MATHEMATICSofMACHINELEARNING

by

March 2020

Contents

1

Introduction

> "No computer has ever been designed that is ever aware of what it's doing; but most of the time, we aren't either."
>
> — Marvin Minsky, 1927-2016

Learning is the process of transforming information and experience into knowledge and understanding. Knowledge and understanding are measured by the ability to perform certain tasks independently. Machine Learning is therefore the study of algorithms and models for computer systems to carry out certain tasks independently, based on the results of a learning process. Learning tasks can range from solving simple classification problems, such as handwritten digit recognition, to more complex tasks, such as medical diagnosis or driving a car.

Machine learning is part of the broader field of Artificial Intelligence, but distinguishes itself from more traditional approaches to problem solving, in which machines follow a strict set of rules they are provided with. As such, it is most useful for tasks such as pattern recognition, that may be simple for humans but where precise rules are hard to come by with, or for tasks that allow for simple rules, but where the complexity of the problem makes any rule-based approach computationally infeasible. An illustrative example of the latter is the recent success of DeepMind's AlphaGo [1], a computer program based on reinforcement learning, at achieving super-human performance at the board game Go (围棋). Even though the rules of the game are simple, the task of beating the best human players seemed impossible only a decade ago due to the daunting complexity that ensues from the number of possible positions.

A General Framework for Learning

Machine learning lies at the intersection of approximation theory, probability theory, statistics, and optimization theory. We illustrate the interplay of these fields with a few basic examples.

In its most basic form, the goal of machine learning is to come up with (learn) a function

$$h: X \to Y,$$

where X is a space of inputs or features, and Y consists of outputs or responses. The input space X is usually modelled as a metric space (such as $\mathbb{R}^d$), and the inputs could represent images, texts, emails, gene sequences, networks, financial time series, or demographic data. The output could consis of quantitative values, such as a temperature or the amount of a certain substance in the body, or of

[1] https://deepmind.com/research/case- studies/alphago- the- story- so- far

qualitative or categorical values, such as {YES,NO} or {0,1,2,3,4,5,6,7,8,9}. The first type of problem is usually called regression, while the latter is called classification. The function h is sometimes called a hypothesis, a predictor, or a classifier. A classifier h that takes only two values (typically 0 and 1,or
−1 and 1) is called a binary classifier. In a machine learning scenario, a function h is chosen from a predetermined set of functions
H, called the hypothesis space.

Machine learning problems can be subdivided into supervised and unsupervised learning problems. In supervised learning, we have at our disposal a collection of input-output data pairs $\{x_i, y_i\}_{i=1}^n \subset X \times Y$, andthegoalistolearnafunctionh: fromthisdata.Thecollectionofpairs $\{(x_i,y_i)\}_{i=1}^n$ iscalled the training set. In unsupervised learning, one does not have access to a training set. The prototypical example of an unsupervised learning task is clustering, where the tasks is to subdivide a given data set into groups based on similarities. This course will deal mainly with supervised learning.

Example 1.1 (Digit recognition). Given a dataset of pixel matrices, each representing a grey-scale image, with associated labels telling us for each image the number it represents, the task is to use this data to train a computer program to recognise new numbers (see Figure 1.1). Such classification tasks are often carried out using deep neural networks, which constitute a powerful class of non-linear functions.

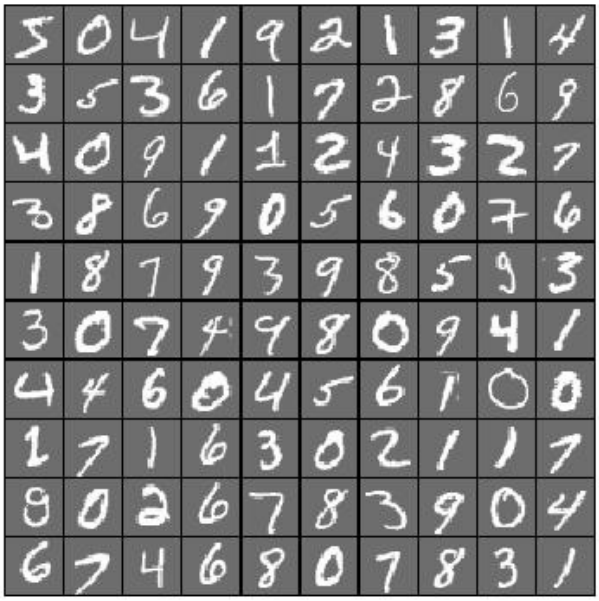


Figure 1.1: The MNIST (Modified National Institute of Standards and Technology, http://yann.lecun.com/exdb/mnist/) dataset is a large collection of images of hand-written digits, and is a frequently used benchmark in machine learning.

Example 1.2. (Clustering) In clustering applications, one observes data $\{x_i\}_{i=1}^n$,andthegoalisto subdivide the data into a number of distinct groups based on similarity, where similarity is measured using a distance function. Figure 1.2 shows an example of an artificial clustering problem and a possible

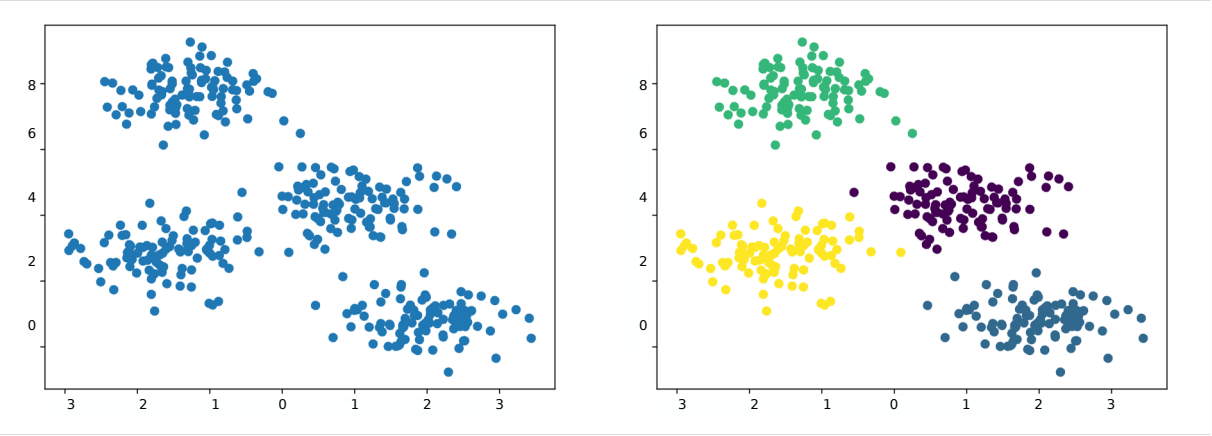


Figure 1.2: A collection of random points on the plane. The image on the right shows the four clusters as determined by the k-means algorithm.

solution. The notion of distance used depends on the application. For example, for binary sequences or DNA sequences one can use the Hamming metric, which simply counts the positions at which two sequences differ. Clustering is used extensively in genetics, biology and medicine. For example, clustering can be used to identify groups of genes (segments of DNA with a function) that encode proteins which are part a common biological pathway. Other uses of clustering are market segmentation and community detection in networks.

Approximation Theory

One often makes the simplified assumption that the observed training data comes from an unknown function f : $X \to Y$. The goal is to approximate the function f with a function h from a hypothesis class H,basedonlyontheknowledgeafinitesetofsamples{xi,yi}n yi≈f(xii=1,whereweassume)for

i $\in [n] := \{1, \ldots, n\}$. Which class of functions is adequate depends on the application at hand, as well as on computational and statistical considerations. In many cases a linear function will do well, while in other situations polynomials or more complex functions, like neural networks, are better suited.

Example 1.3. (Linear regression) Linear Regression is the problem of finding a relationship of the form

$$Y = \beta_0 + \beta 1 X 1 + \ldots + \beta p \, Xp ,$$

where the X1,...,Xp are covariates that describe certain characteristics of a system and Y is the response. Given a set of input-output pairs (xi, yi), arranged in a matrix X and a vector y, we can guess thecorrectβ=(β,β,...,β>01p)bysolvingtheleast-squaresoptimizationproblem

$$\underset{\beta}{\text{minimize}} \quad \| X \beta - y \|_2^2.$$

Figure 1.3 shows an example of linear regression.

Example 1.4. (Text classification) In text classification, the task is to decide to which of a given set of categories a given text belongs. The training data consists of a bag of words: this is a large sparse matrix, whose columns represent words and the rows represent articles, with the (i, j)-th entry containing the number of times word j is contained in text i.

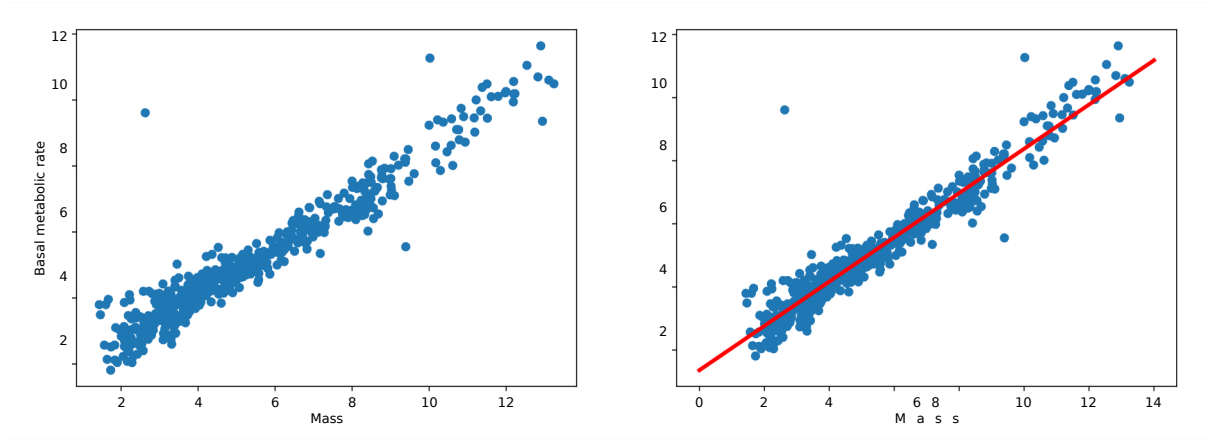
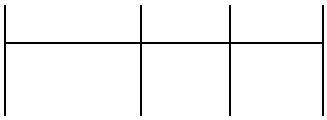


Figure 1.3: The relationship of mass to the logarithm of the basal metabolic rate in mammals. The data consists of 573 samples taken from the PanTHERA database, and the example featured in the episode Size Matters of the BBC series Wonders of Life. The right images shows the regression line determined using linear least squares.

Goal Soup
Article1 5 0
Article2 1 7



For example, in the above set we would classify the first article as "Sports" and the second one as "Food". One such training dataset is the Reuters Corpus Volume I (RCV1) 2, an archive of over 800,000 categorised newswire stories. A typical binary classifier for such a problem would be a linear classifier of the form

h (x > +) = w x b,

with w ∈ R^d and b ∈ R. Given a text, represented as a row of the dataset x, it is classified into one of two classes {+1, −1}, depending on whether h(x) > 0 or h(x) < 0.

Example 1.5: (Deep Neural Networks) Neural networks are functions of the form

f ` ◦ f `_1 ◦ . . . f1,

where each f is the componen-wise composition of an Rdk−1k activation function σ with a linear map → Rdk , x →7Wkx+bk. An activation function could be the sigmoid σ(x)=1/(1+e−x), which takes values between (0, 1), and which can be interpreted as “selecting” certain coordinates of a vector depending on whether they are positive or negative (see Figure 1.4). The coefficients wkij of the matrix W k in each layer are the weights, while the entries of bk are called the bias terms. The weights and bias terms are to be adapted in order to fit observed input-output pairs. A neural network is usually represented as a graph, see Figure 1.5. Neural networks have been extremely successful in pattern recognition, and are widely used in applications ranging from natural language processing to machine translation and medical diagnostics.

One of the earliest theoretical results in approximation theory is a theorem by Weierstrass that shows ~~that we can approximate any continuous f~~unction on an interval to arbitrary precision by polynomials.

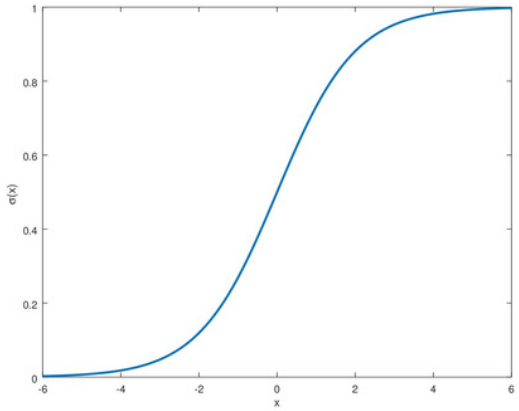


Figure 1.4: The sigmoid function

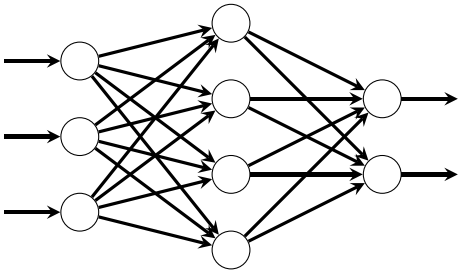


Figure 1.5: A neural network. Each layer correspond to applying a linear map to the outputs of the previous layer, followed by an activation function. Each arrow represents a weight. For example, the transition from the first layer to the second is a map $R3 \to R4$, and the weight associated with the arrow from the second node in layer 1 to the first node in layer 2 is the (1, 2)-entry in the matrix defining the corresponding linear map.

Theorem 1.6 (Weierstrass). Let f be a continuous function on $[a, b]$. Then for any $\varepsilon > 0$ there exists a polynomial $p(x)$ such that

$$\|f - p\|_\infty = \max_{x \in [a,b]} |f(x) - p(x)| < \varepsilon$$

This theorem is remarkable because it shows that we can approximate any continuous function on a compact interval using only a finite number of parameters, the coefficients of a polynomial. The problem with this theorem is that it gives no bound on the size of the polynomial, which can be rather large. It also does not give a procedure of actually computing such an approximation, let alone finding one efficiently. We will see that neural networks have the same approximation properties, i.e., for every continuous function on an interval can be approximated to arbitrary accuracy by a neural network. There are many variations on such results for approximating a class of functions through a simpler class, and we will be interested in cases where such approximations can be efficiently computed. One way of finding good approximating functions is by using optimization methods.

Optimization

The notion of best fit is formalized by using a loss function. A loss function $L : Y \times Y \to R_+$ measures the mismatch between a prediction on a given input $x \in X$ and an element $y \in Y$. The empirical risk of

a function $h: X \to Y$ is the average loss $L(h(x_i), y_i)$ over the training data,

$$\hat{R}_n(h) := \frac{1}{n} \sum_{i=1}^{n} L(h(x_i), y_i)$$

One would then aim to find a function h among a set of candidate function H that minimizes the loss when applying the function to the training data:

$$\text{minimize}_{h \in H} \hat{R}_n(h). \quad (1.1)$$

Problem (1.1) is an optimization problem. Minimizing over a set of functions my look abstract, but functions in H are typically parametrized by few parameters. For example, when the class H consists of linear functions of the form $\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$, as in Example 1.3, then the optimization problem (1.1) amounts to minimizing a function over R^{p+1}. In the case of neural networks, Example 1.5, one optimizes over the weights w_k and biases b_{kj}, i_j.

The form of the loss function depends on the problem at hand and is usually derived from statistical considerations. Two common candidates are the square error for regression problems, which applied to a function $h: R^d \to R$ takes the form

$$L(h(x), y) = (h(x) - y)^2,$$

and the indicator loss function, which takes the general form

$$L(h(x), y) = 1\{h(x) = y\} = \begin{cases} 0 & \text{if } h(x) = y \\ 1 & \text{if } h(x) \neq y \end{cases}$$

As this function is not continuous, in practice one often encounters continuous approximations. A binary classifier is often implemented by a function $h: X \to [0,1]$ that provides a probability of an input belonging to a class. If $h(x) > 1/2$, then x is assigned to class 1, while if $h(x) \leq 1/2$, then x is assigned to class 0. A common loss function for this setting is the log-loss function, or cross-entropy,

$$L(h(x), y) = \begin{cases} -\log(h(x)) & y = 1 \\ -\log(1 - h(x)) & y = 0 \end{cases} \quad (1.2).$$

The function is designed to take on large values if the class predicted by h(x) does not match y, and can be interpreted in the context of maximum-likelihood estimation.

Finding or approximating a minimizer of a function falls into the realm of numerical optimization. While for linear regression we can solve the relevant optimization problem (least-squares minimization) in closed form, for more involved problems such as neural networks we use optimization algorithms such as gradient descent: we start with an initial guess and try to minimize our function by taking steps in direction of steepest descent, that is, along the negative gradient of the function. In the case of composite functions such as neural networks, computing the gradient requires the chain rule, which leads to the famous backpropagation algorithm for training a neural network that will be discussed in detail.

There are many challenges related to optimization models and algorithms. The function to be minimized may have many local minima or saddle points, and algorithms that look for minimizers may find any one of these, instead of a global minimizer. The functions to be minimized may not be differentiable, and methods from the field of non-smooth optimization come into play. The biggest challenge for optimization algorithms in the context of machine learning, however, lies in the particular

form of the objective function: it is given as a sum of many terms, one for each data point. Evaluating such a function and computing its gradient can be time and memory consuming. An old class of algorithms that includes stochastic gradient descent circumvents this issue by not computing the gradient of the whole function at each step, but only of a small random subset of the terms. These algorithms work surprisingly well, considering that they do not make use of all the information available at every step.

Statistics

Suppose we have a binary classification task at hand, with $Y = \{0, 1\}$. We could learn the following function from our data:

$$h(x) = \begin{cases} y_i & \text{if } x = x_i, \\ 1 & \text{otherwise.} \end{cases}$$

This is called learning by memorization, since the function simply memorizes the value y_i for every seen example x_i. The empirical risk with respect to h and the unit loss function for this problem is

$$\hat{R}_n(h) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{h(x_i) \neq y_i\}} = 0.$$

Nevertheless, this is not a good classifier: it will not perform very well outside of the training set. This is an example of overfitting: when the function is adapted too closely to the seen data, it does not generalize to unseen data. The problem of generalization is the problem of finding a classifier that works well on unseen data.

To make the notion of generalization more precise, we assume that the training data points (x_i, y_i) are realizations of a pair of random variables (X, Y), sampled from an (unknown) probability distribution on the product $X \times Y$. The variables X and Y are in general not independent (otherwise there would be nothing to learn), but are related by a relationship of the form $Y = f(X) + \varepsilon$, where ε is a random perturbation with expected value $\mathbb{E}[\varepsilon] = 0$. One could interpret the presence of the random noise ε as indicative of uncertainty or missing information. For example, when trying to predict a certain disease based on genetic markers, the genetic data might simply not carry enough information to always make a correct prediction. The function f is called the regression function. It is the conditional expectation of Y given a value of X,

$$f(x) = \mathbb{E}[Y \mid X = x].$$

Given a classifier $h \in H$ and a loss function L, the *generalization risk* is the expected value

$$R(h) = \mathbb{E}[L(h(X), Y)].$$

If $L(h(x), y) = \mathbb{1}_{\{h(x) = y\}}$ is the unit loss, then this is simply $P\{h(X) = Y\}$, i.e., the probability of misclassifying a randomly chosen input-output pair. The training data can be modelled as sampling form n pairs of random variables

$$(X_1, Y_1), \ldots, (X_n, Y_n)$$

that are identically distributed and independent copies of (X, Y). Given a classifier h, the empirical risk becomes a random variable

$$\hat{R}_n(h) = \frac{1}{n} \sum_{i=1} L(h(X_i), Y_i).$$

The expected value of this random variable is

$$\mathbb{E}[\hat{R}_n(h)] = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}[L(h(X_i), Y_i)] = \mathbb{E}[L(h(X), Y)] = R(h),$$

where we used the linearity of expectation and the fact that the (Xi, Yi) are independent and identically distributed (i.i.d). The empirical risk $\hat{R}\hat{}(h)$ is thus an unbiased estimator of the generalization risk R(h).

Example 1.7. The loss function is often chosen so that the problem of empirical risk minimization becomes a maximum likelihood estimation problem. Consider the example where Y takes values in {0, 1} with probability P{Y = 1 | X = x} = f(x). Conditioned on X = x, Y is a Bernoulli random variable with parameter p = f(x), and the log-loss function (1.2) is precisely the negative log-likelihood function for the problem of estimating the parameter p.

When looking for a good hypothesis h, all we have at our disposal is the empirical risk function constructedfromthetrainingdata.Itturnsoutthatthequalityofanempiricalriskminimizer$\hat{h}\hat{}$froma hypothesis class

H can be measured by the estimation error, which compares the generalization risk of $\hat{h}\hat{}$tothesmallestpossiblegeneralizationriskin

H, and the approximation error, which measures how small the generalization risk can become within

H. There is usually a trade-off between these to errors: if the class of functions

H is large, then it is likely to contain functions with small generalization risk and thus have small approximation error, but the empirical risk minimizer $\hat{h}\hat{}$ is likely to "overfit" the data and not generalize well. On the other hand, if

H is small (in the extreme case, consisting of only one function), then the empirical risk minimizer is likely to be close to the best possible hypothesis in

H, but the approximation error will be large. Figure 1.6 shows an example in which data from a function with noise is approximated by polynomials of different degrees.

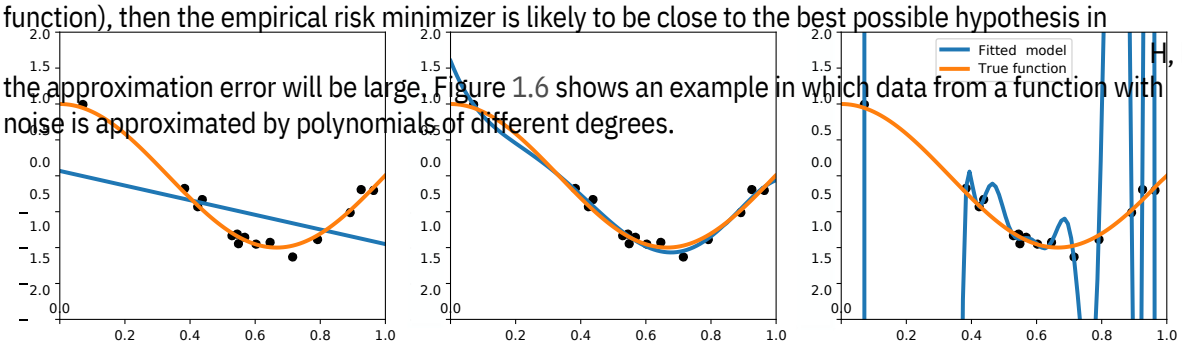


Figure 1.6: The data consists of 15 samples from the graph of a cosine function with added noise. The three displays show an approximation with a linear function, with a polynomial of degree 5, and with a polynomial of degree 15. The linear function has a large error on both the training set and in relation to the true function. The polynomial of degree 15, on the other hand, has zero error on the training data (a polynomial of degree d can fit d + 1 points with distinct x-values exactly), but it will likely perform poorly on new data. This is an example of overfitting: more parameters in the model will not necessarily lead to a better performance.

The field of Statistical Learning Theory aims to understand the relation between the generalization risk, the empirical risk, the capacity of a hypothesis class

H, and the number of samples n. In particular, notions such as the capacity of a hypothesis class are given a precise meaning through concepts such as VC dimension, Rademacher complexity, and covering numbers.

Notes

The ideas from approximation theory, optimization and statistics that underlie modern machine learning are old. Linear regression was known to Legendre and Gauss. The Weierstrass Approximation Theorem

was published by Weierstrass in [31], see [28, Chapter 6] for an account and more history on approximation theory. Neural networks go back to the seminal work by McCulloch and Pitts from 1943 [16], followed by Rosenblatt's Perceptron [22]. The term "Machine Learning" was first coined by Arthur Samuel in 1959 [24]; at the time, "Cybernetics" was still widely used. Gradient descent was known to Cauchy, and the most important algorithm for deep learning today, Stochastic Gradient Descent, was introduced by Robbins and Monro in 1951 [21]. The field of Statistical Learning Theory arose in the 1960s through seminal work by Vapnik and Chervonenkis, see [29] for an overview. For an account of mathematical learning theory, see [5].

Research in machine learning exploded in the 1990s, with striking new results and applications appearing at breathtaking pace. Still, apart from some of the more theoretical developments in learning theory and high-dimensional probability, these breakthroughs rarely relied on mathematics that was not available 50 years ago. So what has changed since the early days of cybernetics? The main reason for the sudden surge in popularity is the availability of vast amounts of data, and equally important, the computational resources to process the data. New applications have in turn led to new mathematical problems, and to new connections between various fields.

2

Overview of Probability

In this lecture we review relevant notions from probability theory, with an emphasis on conditional expectation.

Probability spaces

A probability space is a triple $(\Omega, \mathcal{F}, P)$, consisting of a set Ω, a σ-algebra $\mathcal{F}$ of subsets of Ω, called events, and a probability measure P. That $\mathcal{F}$ is a σ-algebra means that it contains $\emptyset$ and Ω and that it is closed under countable unions and complements. The probability measure P is a non-negative function $P: \mathcal{F} \to [0,1]$ such that $P(\emptyset) = 0$, $P(\Omega) = 1$, and

$$P\left(\bigcup_i A_i\right) = \sum_i P(A_i)$$

for a countable collection $\{A_i\}$ with $A_i \cap A_j = \emptyset$ for $i \neq j$. We interpret $P(A \cup B)$ as the probability of A or B happening, and $P(A \cap B)$ as the probability of A and B happening. Note that $(A \cup B)^c = A^c \cap B^c$, where A^c is the complement of A in Ω. If the A_i are not necessarily disjoint, then we have the important

$$P\left(\bigcup_i A_i\right) \leq \sum_i P(A_i).$$

union bound

This bound is sometimes also referred to as the zero-th moment method. We say that an event A holds almost surely if $P(A) = 1$ (note that this does not mean that the complement of A in Ω is empty).

Random variables

A random variable is a measurable map

$$X : \Omega \to \mathcal{X},$$

where $\mathcal{X}$ is typically $\mathbb{R}$, $\mathbb{R}^d$, $\mathbb{N}$, or a finite set $\{1, \ldots, k\}$. For a measurable set $A \subset \mathcal{X}$, we write

$$P(X \in A) := P(\{\omega \in \Omega : X(\omega) \in A\}).$$

We will usually use upper-case letters X, Y, Z for random variables, lower-case letters x, y, z for the values that these can take, and x, y, z if these are vectors in some $\mathbb{R}^d$.

Example 2.1. A random variable specifies which events “we can see”. For example, let Ω = {1, 2, 3, 4, 5, 6} and define $X : \Omega \to \{0, 1\}$, $X(\omega) = 1_{\{\omega \in \{4,6\}\}}$. Then $P(X = 1) = \frac{1}{3}$, $P(X = 0) = \frac{2}{3}$. If all the information we get about ω from X, then we can only determine whether the result of rolling a die gives an even number greater than 3 or not, but not the individual result.

The map $A \mapsto P(X \in A)$ for subsets of X is called the distribution of the random variable. The distribution completely describes the random variable, and there will often be no need to refer to the domain Ω. If $F : X \to Y$ is another measure map, then $F(X)$ is again a random variable. In particular, if X is a subset of $\mathbb{R}^d$, we can add and multiply random variables to obtain new random variables, and if X and Y are two distinct random variables, then (X, Y) is a random variable in the product space. In the latter case we also write $P(X \in A, Y \in B)$ instead of $P((X, Y) \in A \times B)$. Note that this also has an interpretation in terms of intersections of events: it is the probability that both $X \in A$ and $Y \in B$.

A discrete random variable takes countably many values, for example in a finite set $\{1, \ldots, k\}$ or in $\mathbb{N}$. In such a case it makes sense to talk about the probability of individual outcomes, such as $P(X = k)$ for some $k \in X$. An absolutely continuous random variable takes values in $\mathbb{R}$ or $\mathbb{R}^d$ for $d > 1$, and is defined as having a density $\rho_X = \rho(x)$ such that $P(X \in A) = \int_A \rho(x)\, dx$. In the case where $X = \mathbb{R}$, we consider the cumulative distribution function (cdf) $P(X \leq t)$ for $t \in \mathbb{R}$. The complement, $P(X > t)$ (or $P(X \geq t)$), is referred to as the tail. Many applications are concerned with finding good bounds on the tail of a probability, as the the tail often models the probability of rare events. If X is absolutely continuous, then the probability of taking a particular single value vanishes, $P(X = a) = 0$. For a random variable $Z = (X, Y)$ taking values in $X \times Y$, we can consider the joint density $\rho_Z(x, y)$, but also the individual densities of X and Y, for which we have $\rho_X(x) = \int_Y \rho_Z(x, y)\, dy$.

The ensuing distributions for X and Y are called the marginal distributions.

Example 2.2. Three of the most common distributions are:

- Bernoulli distribution, taking values in $\{0, 1\}$ and defined by $P(X = 1) = p$, $P(X = 0) = 1 - p = q$ for some $p \in [0, 1]$. We can replace the range $\{0, 1\}$ by any other two-element set, for example $\{-1, 1\}$, but then the relation to other distributions may not hold any more.
- Binomial distribution $\mathrm{Bin}(n, p)$, taking values in $\{0, \ldots, n\}$ and defined by $P(X = k) = p^k (1 - p)^{n-k}$ (2.1) for $k \in \{0, 1, \ldots, n\}$ and some $p \in [0, 1]$. We can also write a binomial random variable as a sum of Bernoulli random variables, $X = X_1 + \cdots + X_n$, since $X = k$ if and only if k of the summands have the value 1.

- Normal distribution $N(\mu, \sigma^2)$, also referred to as Gaussian, with mean μ and variance σ^2, defined on $\mathbb{R}$ and with density

$$f(x) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

This is the most important distribution in probability and statistics, as most other distributions can be approximated by it.

Expectation

The expectation (or mean, or expected value) of a discrete random variable is defined as

$$E[X] = \sum_{k \in X} k \cdot P(X = k).$$

For an absolutely continuous random variable with density $\rho(x)$, it is defined as

$$E[X] = \int_X x \rho(x)\, dx.$$

Note that the expectation does not always need to exist since the sum or integral need not converge. When we require it to exist, we often write this as $E[X] < \infty$.

Example 2.3. The expectation of a Bernoulli random variable with parameter p is E[X] = p. The expectation of a Binomial random variable with parameters n and p is E[X] = np. For example, if one were to flip a biased coin that lands on heads with probability p, then this would correspond to the number of heads one would "expect" after n coin flips. The expectation of the normal distribution $N(\mu, \sigma^2)$ is μ. This is the location on which the "bell curve" is centred.

One of the most useful properties is linearity of expectation. If $X_1, \ldots, X_n$ are random variables taking values in a subset of $\mathbb{R}$ and $a_1, \ldots, a_n \in \mathbb{R}$, then

$$E[a_1 X_1 + \ldots + a_n X_n] = a_1 E[X_1] + \ldots + a_n E[X_n].$$

Example 2.4. The expected value of a Bernoulli random variable with parameter p is

$$E[X] = 1 \cdot P(X = 1) + 0 \cdot P(X = 0) = p.$$

The linearity of expectation then immediately gives the expectation of the Binomial distribution with parameters n and p. Since such a random variable can be written as $X = X_1 + \cdots + X_n$, with X_i Bernoulli, we get

$$E[X] = E[X_1 + \ldots + X_n] = E[X_1] + \ldots + E[X_n] = np.$$

This would be (slightly) harder to deduce from the direct definition (2.1), when one would have to use the binomial theorem.

If $F: X \to Y$ is a measurable function, then the expectation of the random variable can be expressed as

$$E[F(X)] = \int_X F(x)\rho(x)dx \quad (2.2)$$

in the case of an absolutely continuous random variable, and similarly in the discrete case. [1]
An important special case is the indicator function

$$F(X) = 1_{\{X \in A\}} = \begin{cases} 1 & X \in A \\ 0 & X \in A \end{cases}.$$

Then

$$E[1_{\{X \in A\}}] = P(X \in A), \quad (2.3)$$

as can be seen by applying (2.2) to the indicator function. The identity (2.3) is useful, as it allows to properties of the expectation, such as linearity, in the study of probabilities of events. The expectation also has the following monotonicity property: if $0 \leq X \leq Y$, where X, Y are real-valued random variables, then $E[X] \leq E[Y]$.

Another important identity for random variables is the following. Assume X is absolutely continuous, takes values in $\mathbb{R}$, and $X \geq 0$. Then

$$E[X] = \int_0^\infty P(X > t)\, dt.$$

Using this identity, one can deduce bounds on the expectation from bounds on the tail of a probability distribution.
The variance of a random variable is the expectation of the square deviation from the mean:

$$\mathrm{Var}(X) = E[(X - E[X])^2].$$

The variance measures the "spread" of a distribution: random variables with a small variance are more likely to stay close to their expectation.

Example 2.5. The variance of the normal distribution is σ^2. The variance of the Bernoulli distribution is $p(1-p)$ (verify this!), while the variance of the Binomial distribution is $np(1-p)$.

The variance scales as $\mathrm{Var}(aX + b) = a^2\mathrm{Var}(X)$. In particular, it is translation invariant. The variance is in general not additive (but it is, if the random variables are independent).

Independence

A set of random variables $\{X_i\}$ taking values in the same range X is called independent if for any subset $\{X_{i_1}, \ldots, X_{i_k}\}$ and any subsets $A_1, \ldots, A_k \subset X$, we have

$$P(X_{i_1} \in A_1, \ldots, X_{i_k} \in A_k) = P(X_{i_1} \in A_1) \cdots P(X_{i_k} \in A_k).$$

In words, the probability of any of the events happening simultaneously is the product of the probabilities of the individual events. A set of random variables $\{X_i\}$ is said to be pairwise independent if every subset of two variables is independent. Note that pairwise independence does not imply independence.

[1] We will not always list the formulas for both the discrete and continuous, when the form of one of these cases can be easily guessed from the form of the other case. In any case, the sum in the discrete setting is also just an integral with respect to the discrete measure.

Example 2.6. Assume you toss a fair coin two times. Let X be the indicator variable for heads on the first toss Y the indicator variable for heads on the second toss, and Z the random variable that is 1 if X = Y , 0 if X and 1/2 6= Y . Taken individually, each of these random variables is a Bernoulli random variable with p . They are also pairwise independent, as is easily verified, but not independent, since

$$P(X = 1, Y = 1, Z = 1) = \frac{1}{4} \neq \frac{1}{8} = P(X = 1)P(Y = 1)P(Z = 1).$$

Intuitively, the information that X = 1 and Y = 1 already implies Z = 1, so adding this constraint does not alter the probability on the left-hand side.

We say that a set of random variables $\{X_i\}$ is i.i.d. if they are independent and identically distributed. This means that each Xi can be seen as a copy of X1 that is independent of it, and in particular all the Xi have the same expectation and variance.

One of the most important results in probability (and, arguably, in nature) is the (strong) law of large numbers. Given random variables $\{X_i\}$, define the sequence of averages as

$$\overline{X_n} = \frac{1}{n}(X_1 + \ldots + X_n).$$

Since each random variable is, by definition, a function on a sample space Ω, we can consider the pointwise limit

$$\lim_{n \to \infty} \overline{X_n},$$

which is the random variable that for each $\omega \in \Omega$ takes the limit $\lim_{n \to \infty} \overline{X_n}(\omega)$ as value.[2]

Theorem 2.7 (Law of Large Numbers). Let $\{X_i\}$ be a sequence of i.i.d. random variables with $E[X_1] = \mu < \infty$. Then the sequence of averages $\overline{X_n}$ converges almost surely to μ:

$$P(\lim_{n \to \infty} \overline{X_n} = \mu) = 1.$$

Example 2.8. Let each Xi be a Bernoulli random variable with parameter p. One could think this as flipping a coin that will show heads with probability p. Then Xn is the average number of heads when flipping the coin n times. The law of large numbers asserts that as n increases, this average approaches p almost surely. Intuitively, when flipping the coin a billion times, the number of heads we get divided by a billion will be indistinguishable from p: if we do not know p we can estimate it in this way.

Some useful inequalities

In applications it is often not possible to get precise expressions for a probability we are interested in, most often because we don't know the exact distribution we are dealing with and only have access to parameters such as the expectation or the variance. There are several useful inequalities that help us bound the tail or deviation probabilities. For the following, we assume $X \subset R$.

• Jensen's Inequality Let $f : R \to R$ be a convex function, that is, $f(\lambda x + (1 - \lambda)y) \leq \lambda f(x) + (1 - \lambda)f(y)$ for $\lambda \in [0, 1]$. Then $f(E[X]) \leq E[f(X)]$.

[2] That this is indeed a random variable in the formal sense follows from measure theory, we will not be concerned with those details.

• Markov'sInequality ("first moment method") For $X \geq 0$ and $\lambda > 0$,

$$P(X \geq \lambda) \leq \frac{E[X]}{\lambda}$$

• Chebyshev'sInequality ("second moment method") For $\lambda > 0$,

$$P(|X - E[X]| \geq \lambda) \leq \frac{Var(X)}{\lambda^2}$$

• ExponentialMomentInequality For any $s, \lambda \geq 0$,

$$P(X \geq \lambda) \leq e^{-s\lambda} E[e^{sX}].$$

Note that both the Chebyshev and the exponential moment inequality follow from the Markov inequality applied to certain transformations of X.

Conditional probability and expectation

Given events $A, B \subset \Omega$ with $P(B) > 0$, the conditional probability of A conditioned on B is defined as

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}.$$

Oneinterpretsthisastheprobabilityof A ifweassume B. Thatis,ifweobservedB,thenwereplacethe whole set Ω by B and consider B to be the new space of events, considering only the part of events A that lie in B. We can rearrange the expression for conditional probability to

$$P(A \cap B) = P(A \mid B) P(B),$$

from which we get the sometimes useful identity

$$P(A) = P(A \mid B) P(B) + P(A \mid B^c) P(B^c), \tag{2.4}$$

whereBcdenotesthecomplementof B in Ω.

Sincebyexchangingtheroleof A and B weget $P(A \cap B) = P(B \mid A) P(A)$, we arrive at the famous Bayes rule for conditional probability:

$$P(A \mid B) = \frac{P(B \mid A) P(A)}{P(B)},$$

defined whenever both A and B have non-zero probability. These concepts clearly extend to random variables, where we can define, for example,

$$P(X \in A \mid Y \in B) = \frac{P(X \in A, Y \in B)}{P(Y \in B)}.$$

Note that if X and Y are independent, then the conditional probability is just the normal probability of X: knowing that $Y \in B$ does not give us any additional information about X! If we fix an event such as $\{Y \in B\}$, then we can define the conditioning of the random variable X to this event as the random variable X′ with distribution $P(X' \in A) = P(X \in A \mid Y \in B)$.
In particular, $P(X \in A \mid Y \in B) + P(X \notin A \mid Y \in B) = 1$.

Example 2.9. Consider the case of testing for doping at a sports event. Let X be the indicator variable for the presence of a certain drug, and Y the indicator variable for whether the person tested has taken the drug. Assume that the test is 99% accurate when the drug is present and 99% accurate when the drug is not present. We would like to know the probability that a person who tested positive actually took the drug, namely $P(Y = 1 \mid X = 1)$. Translated into probabilistic language, we know that

$$P(X = 1 \mid Y = 1) = 0.99, \quad P(X = 1 \mid Y = 1) = 0.01$$
$$P(X = 0 \mid Y = 0) = 0.99, \quad P(X = 0 \mid Y = 0) = 0.01.$$

Assuming that only 1% of the participants have taken the drug, which translates to $P(Y = 1) = 0.01$, we find that the overall probability of a positive test result is, using (2.4),

$$P(X = 1) = P(X = 1 \mid Y = 0)P(Y = 0) + P(X = 1 \mid Y = 1)P(Y = 1)$$
$$= 0.01 \cdot 0.99 + 0.99 \cdot 0.01 = 0.0198$$

Hence, using Bayes' rule, we conclude that

$$P(Y = 1 \mid X = 1) = \frac{P(X = 1 \mid Y = 1)P(Y = 1)}{P(X = 1)} = \frac{0.99 \cdot 0.01}{0.0198} = 0.5.$$

That is, we get the surprising result that even though our test is very unlikely to give false positives and false negatives, the probability that a person tested positive has actually taken the drug is only 50%. The reason is that the event itself is highly unlikely.

We now come to the notion of conditional expectation. Let X,Y be random variables. If X is discrete, then the conditional expectation of X c

∑onditionedonaneventY=yisdefinedasE[X|Y=y]=kP(X=k|Y=y).(2.5)k

This is simply the expectation of the random variable X′ with distribution $P(X' \in A) = P(X \in A \mid Y = y)$. Intuitively, we assume that Y = y is given/has been observed, and consider the expectation of X under this additional knowledge.

Example 2.10. Assume we are rolling dice, let X be the random variable giving the result, and let Y be the indicator variable for the event that the result is at most 4. Then E[X] = 3.5 and $E[X \mid Y = 1] = 2.5$ (verify this!). This is the expected value if we have the additional information that the result is at most 4.

In the absolutely continuous case we can define a conditional density

$$\rho_{X,Y}(x, y) \Big/ \rho_Y(y) = \rho_{X|Y=y}(x),$$

where $\rho_{X,Y}$ is the joint density of (X, Y) and ρ_Y t

[hedensityofY.Theconditionalexpectationisthendefined$E[X|Y=y]=x\rho_{X|Y=y}(x)dx$.(2.7)X

Wereplacethe pofwithanupdateddensity
density $\rho_{X|Y=y}$ has been observed when computing the expectation of $X \mid Y = y$ that takes into account that a value $\in Y$
X When looking at (2.5) and (2.7), we get a different number $E[X \mid Y = y]$ for each $y \in Y$, where we
assume Y to be the space where Y takes values. Hence, we can define a random variable $E[X \mid Y]$ on Y as follows:

$$E[X \mid Y](\omega) = E[X \mid Y = y].$$

If $X = f(Y)$ is completely determined by Y, then clearly

$$E[X \mid Y](y) = E[X \mid Y = y] = E[f(Y) \mid Y = y] = E[f(y) \mid Y = y] = f(y),$$

since the expected value of a constant is just that constant, and hence $E[X \mid Y] = f(Y)$ as a random variable.

Using the definition of the conditional density (2.6), Fubini's Theorem and expression (2.7), we can write the expectation of X as

$$\begin{aligned} E[X] &= \int_X x \rho_X(x)\, dx \\ &= \int_X x \left(\int_Y \rho_{(X,Y)}(x, y)\, dy \right) dx \\ &= \int_Y \left(\int_X x \rho_{(X,Y)}(x,y) dx \right) dy \\ &= \int_Y \left(\int_X x \rho_{X|Y=y}(x) dx \right) \rho_Y(y) dy = \int_Y E[X \mid Y = y] \rho_Y(y) dy. \end{aligned}$$

One can interpret this as saying that we get the expected value of X by integrating the expected values conditioned on $Y = y$ with respect to the density of Y. In the discrete case, the identity has the form

$$E[X] = \sum_y E[X \mid Y = y] P(Y = y).$$

The above identities can be written more compactly as

$$E[E[X \mid Y]] = E[X].$$

In the context of machine learning, we assume that we have a (hidden) map $f : X \to Y$ from an input space to an output space, and that, for a given input $x \in X$, the observed output is $y = f(x) + \varepsilon$, where ε is random noise with $E[\varepsilon] = 0$. If we consider the input as a random variable X, then the output is random variable $Y = f(X) + \varepsilon$. With $E[\varepsilon \mid X] = 0$ ("the expected value of ε, when knowing X, is zero"). We are interested in the value $E[Y \mid X]$ this, we get

$$E[Y \mid X] = E[f(X) \mid X] + E[\varepsilon \mid X] = f(X),$$

since $f(X)$ is completely determined by X.

Concentration of measure

A very important refinement of Chebyshev's inequality are *concentration inequalities*, which state that the probability of exceeding the expectation is *exponentially small*. A prototype of such an inequality is Hoeffding's Bound.

Theorem 2.11 (Hoeffding's Inequality). Let $X_1, \ldots, X_n$ be independent random variables taking values in $[0, 1]$, and let $S_n = \frac{1}{n} \sum_{i=1}^{n} X_i$ be the average. Then for $t \geq 0$

$$P(|S_n - E[S_n]| \geq t) \leq 2e^{-2nt^2}.$$

Note the implication of this: if we have a sequence of random variables $\{X_i\}$ bounded in $[0, 1]$ (for example, the result of repeated, identical experiments or observations) then as n increases, the probability that the average of the random variables deviates from its mean decreases exponentially with n. In particular, if the random variables all thave the same expectation μ, then (by linearity of expectation) we have $E[X_n] = \mu$, and the probability of straying from this value becomes very small very quickly!

Notes

Even though probability theory and statistics are central to machine learning, probabilistic concepts are often not used rigorously. For example, one frequently encounters expressions such as $P(X \mid Y)$ which, taken literally, do not make sense. Depending on context, such an expression may refer to either the conditional expectation $E[X \mid Y]$, the conditional probability $P(X \in A \mid Y \in B)$, or the conditional density $\rho_{X|Y=y}(x)$. It turns out that for most practical purposes it does not really matter, but it is just something that a mathematics student used to rigorous definitions should be aware of.

A good general introduction to probability theory is §1.1 of [27]. Good references for concentration of measure and related topics are [3, 30].

Part I

Statistical Learning Theory

3

Binary Classification

In this lecture we begin the study of statistical learning theory in the case of binary classification. We will characterize the best possible classifier in the binary case, and relate notions of classification error to each other.

Binary Classification

A binary classifier is a function

$$h : X \to Y = \{0, 1\},$$

where X is a space of features. The fact that we use it not very important, and in many cases we will also consider classifiers taking values in $\{0, 1\}$ or $\{-1, 1\}$, whereconvenient.Binaryclassifiersariseina variety of applications. In medical diagnostics, for example, a classifier could take an image of a skin mole and determine if it is benign or if it is melanoma. Typically, can arise from a function $X \to [0,1]$ that assigns to every input x a probability p. If p > 1/2, then x is assigned to class 1, and if p $\leq 1/2$ it is assigned to class 0.

In the context of binary classification we usually use the unit loss function $L(h(x), y) = 1_{\{h(x) \neq y\}} = \begin{cases} 1 & h(x) \neq y \\ 0 & h(x) = y \end{cases}$.

The unit loss function does not distinguish between false positives and false negatives. A false positive is a pair (x, y) with $h(x) = 1$ but $y = 0$, and a false negative is a pair for which $h(x) = 0$ but $y = 1$. We would like to learn a classifier from a set of observations

$$\{(x_i, y_i)\}_{i=1}^{n} \subset X \times Y. \quad (3.1)$$

The classifier should not only match the data, but generalize in order to be able to classify unseen data. For this, we assume that the points in (3.1) are drawn from a probability distribution on $X \times Y$, and replace each data point (x_i, y_i) in (3.1) with a copy (X_i, Y_i) of a random variable (X, Y) on $X \times Y$. We are after a classifier h such that the expected value of the empirical risk

$$R_n(h) = \frac{1}{n} \sum_{i=1}^{n} 1_{\{h(X_i) \neq Y_i\}} \quad (3.2)$$

is small. We can write this expectation as

$$E[\hat{R}_n(h)] \overset{(1)}{=} E\left[\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{h(X_i) \neq Y_i\}\right] \overset{(2)}{=} \frac{1}{n}\sum_{i=1}^{n} P(h(X_i) \neq Y_i) \overset{(3)}{=} P(h(X) \neq Y) =: R(h),$$

where (1) uses the linearity of expectation, expresses the expectation of an indicator function as probability, and (3) uses the fact that all the X_i, Y_i are identically distributed. The function $R(h)$ is the risk: it is the probability that the classifier gets something wrong.

Example 3.1. Assume that the distribution is such that Y is completely determined by X, that is, $Y = f(X)$. Then

$$R(h) = P(h(X) \neq f(X)),$$

and $R(h) = 0$ if $h = f$ almost everywhere. If X is a finite or compact set with the uniform distribution, then $R(h)$ simply measures the proportion of the input space on which h fails to classify inputs correctly.

While for certain tasks such as image classification there may be a unique label to each input, in general this need not be the case. In many applications, the input does not carry enough information to completely determine the output. Consider, for example, the case where X consists of whole genome sequences and the task is to predict hypertension (or any other condition) from it. The genome clearly does not carry enough information to make an accurate prediction, as other factors also play a role. To account for this lack of information, define the regression function $f(X) = E[Y \mid X] = 1 \cdot P(Y = 1 \mid X) + 0 \cdot P(Y = 0 \mid X) = P(Y = 1 \mid X)$.

Note that if we write

$$Y = f(X) + \varepsilon,$$

then $E[\varepsilon \mid X] = 0$, because

$$f(X) = E[Y \mid X] = E[f(X) \mid X] + E[\varepsilon \mid X].$$

The Bayes classifier

While in Example 3.1 we could choose (at least in principle) $h(x) = f(x)$ and get $R(h) = 0$, in the presence of noise this is not possible. However, we could define a classifier h^* by setting

$$h^*(x) = \begin{cases} 1 & f(x) > \frac{1}{2} \\ 0 & f(x) \leq \frac{1}{2}, \end{cases}$$

We call this the Bayes classifier. The following result shows that this is the best possible classifier.

Theorem 3.2. The Bayes classifier h^* satisfies

$$R(h^*) = \inf_h R(h),$$

where the infimum is over all measurable h. Moreover, $R(h^*) \leq 1/2$.

Proof. Let h be any classifier. To compute the risk $R(h)$, we first condition on X and then average over X:

$$R(h) = E[1_{\{h(X) \neq Y\}}] = E[E[1_{\{h(X) \neq Y\}} | X]]. \tag{3.3}$$

For the inner expectation, we have

$$\begin{aligned} E[1_{\{h(X) \neq Y\}} | X] &= E[1_{\{h(X)=1, Y=0\}} + 1_{\{h(X)=0, Y=1\}} | X] \\ &= E[1_{\{h(X)=1\}} 1_{\{Y=0\}} | X] + E[1_{\{h(X)=0\}} 1_{\{Y=1\}} | X] \\ &\overset{(1)}{=} 1_{\{h(X)=1\}} E[1_{\{Y=0\}} | X] + 1_{\{h(X)=0\}} E[1_{\{Y=1\}} | X] \\ &= 1_{\{h(X)=1\}} (1 - f(X)) + 1_{\{h(X)=0\}} f(X). \end{aligned}$$

To see why (1) holds, recall that the random variable $E[1_{\{h(X)=Y\}} | X = x]$ takes values $E[1_{\{h(x)=0\}} | X = x]$ and $E[1_{\{h(x)=1\}} | X = x]$, and will therefore only be non-zero if $h(x) = 0$ or $h(x) = 1$. We can therefore pull the indicator function out of the expectation.

Hence, using (3.3),

$$R(h) = E[\underbrace{1_{\{h(X)=1\}}(1 - f(X))}_{(1)} + \underbrace{1_{\{h(X)=0\}} f(X)}_{(2)}]. \tag{3.4}$$

For (1), we decompose

$$\begin{aligned} 1_{\{h(X)=1\}}(1 - f(X)) &= 1_{\{h(X)=1, f(X)>1/2\}}(1 - f(X)) \\ &\quad + 1_{\{h(X)=1, f(X)\leq 1/2\}}(1 - f(X)) \\ &\geq 1_{\{h(X)=1, f(X)>1/2\}}(1 - f(X)) \\ &\quad + 1_{\{h(X)=1, f(X)\leq 1/2\}} f(X), \end{aligned} \tag{3.5}$$

where the inequality follows since $(1 - f(X)) \geq f(X)$ if $f(X) \leq 1/2$. By the same reasoning, for (2) we get

$$\begin{aligned} 1_{\{h(X)=0\}} f(X) &\geq 1_{\{h(X)=0, f(X)>1/2\}} f(X) \\ &\quad + 1_{\{h(X)=0, f(X)\leq 1/2\}} f(X). \end{aligned} \tag{3.6}$$

Combining the inequalities (3.5) and (3.6) within the bound (3.4) and collecting the terms that are multiplied with $f(X)$ and those that are multiplied with $1 - f(X)$, we arrive at

$$\begin{aligned} R(h) &\geq E[1_{\{f(X)>1/2\}} f(X) + 1_{\{f(X)\leq 1/2\}}(1 - f(X))] \\ &= R(h^*), \end{aligned}$$

where the last equality follows from (3.4) applied to $h = h^*$. The characterization (3.4) also shows that

$$\begin{aligned} R(h^*) &= E[1_{\{f(X)>1/2\}}(1 - f(X)) + 1_{\{f(X)\leq 1/2\}} f(X)] \\ &= E[\min\{f(X), 1 - f(X)\}], \end{aligned}$$

which completes the proof. □

We have seen in Example 3.1 that the Bayes risk is 0 if Y is completely determined by X. At the other extreme, if the response Y does not depend on X at all, then the Bayes risk is 1/2. This means that for every input, the best possible classifier consists of "guessing" without any prior information, which means that we have a 50% chance of being correct!

The error

$$E(h) = R(h) - R(h^*)$$

is called the excess risk or error of h with respect to the best possible classifier.

Approximation and Estimation

We conclude this lecture by relating notions of risk. In what follows, we assume that a class of classifiers H is given, from which we are allowed to choose. We denote by $\hat{h}_n$ the classifier obtained by minimizing the empirical risk $\hat{R}_n(h)$ over H, that is

$$\hat{R}_n(\hat{h}_n) = \inf_{h \in H} \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{h(X_i) \neq Y_i\}},$$

where the (X_i, Y_i) are i.i.d. copies of a random variable (X, Y) on $\mathcal{X} \times \mathcal{Y}$. Note that $\hat{h}_n$ is what we will typically obtain by computation from samples (x_i, y_i). The way it is defined, it depends on n, the class of functions H, and the random variables (X_i, Y_i), and as such is itself a random variable.

We want $\hat{h}_n$ to generalize well, that is, we want

$$R(\hat{h}_n) = P(\hat{h}_n(X) \neq Y)$$

to be small. We know that the smallest possible value this risk can attain is given by $R(h^*)$, where h^* is the Bayes classifier. We can decompose the difference between the risk of $\hat{h}_n$ and that of the Bayes classifier as follows:

$$R(\hat{h}_n) - R(h^*) = \underbrace{R(\hat{h}_n) - \inf_{h \in H} R(h)}_{\text{Estimation error}} + \underbrace{\inf_{h \in H} R(h) - R(h^*)}_{\text{Approximation error}}$$

The first compares the performance of $\hat{h}_n$ against the best possible classifier within the class H, while the second is a statement about the power of the class H. We can reduce the estimation error by making the class H smaller, but then the approximation error increases. Ideally, we would like to find bounds on the estimation error that converge to 0 as the number of samples n increases.

Notes

4

Finite Hypothesis Sets

Given a fixed hypothesis set H, we would like to study the classifier $\hat{h}$ computed from n random samples (X_i, Y_i) by minimizing the empirical risk over the class H,

$$\hat{R}(\hat{h}) = \inf_{h \in H} \hat{R}(h), \quad \text{where} \quad \hat{R}_n(h) = \frac{1}{n} \sum_{i=1}^{n} 1_{\{h(X_i) \neq Y_i\}}.$$

Hence, for any h, $\hat{R}(h)$ is a random variable that depends on the number of samples n and on the underlying probability distribution. The classifier $\hat{h}$ is also a random variable, and depends on n, the class H, and the underlying distribution. This is the object that we can compute from observed data. If h^* denotes the Bayes classifier, then we would like to bound the excess risk

$$E(\hat{h}) = R(\hat{h}) - R(h^*), \quad (A)$$

where we recall that $R(h) = P(h(X) \neq Y)$. As opposed to, $\hat{R}$ is not a random variable: it depends solely on the probability distribution. Moreover, for any fixed h, $E[\hat{R}(h)] = R(h)$. Denote by $\bar{h}$ a best possible classifier in H, that is, the one that generalizes best:

$$R(\bar{h}) = \inf_{h \in H} R(h).$$

The parameter $\bar{h}$ depends only on the class H and the probability distribution. A less ambitious goal is to bound the difference

$$R(\hat{h}) - R(\bar{h}). \quad (B)$$

We emphasize here that $\hat{R}(h)$ and $R(\hat{h})$ are both random variables, and $\hat{R}(\hat{h})$ is a random variable in two ways. Bounds on (A) and (B) are therefore probabilistic. More precisely, for any given tolerance $\delta \in (0, 1)$, we want to find constants $C(n, \delta)$ and $C'(n, \delta)$ such that

$$R(\hat{h}) - R(h^*) \leq C(n,\delta) \text{ and } R(\hat{h}) - R(\bar{h}) \leq C'(n,\delta)$$

holds with probability $1 - \delta$. Ideally, the constants should also depend on properties of the set H, for example the size of H if this set is finite. In addition, we would like the constants to decrease to 0 as $n \to \infty$. In this lecture we will derive bounds on (B) in the case where H is a finite set.

Risk bounds for finite sets of classifiers

In this section we prove the following bound.

Theorem 4.1. Let $H = \{h_1, \ldots, h_K\}$ be a finite dictionary. Then for $\delta \in (0, 1)$,

$$P\left(R(\hat{h}) - \inf_{h \in H} R(h) \leq \sqrt{\frac{2 \log 2K}{n}} \right) \geq 1 - \delta.$$

This important result shows that (with high probability) we can bound the estimation error by a term that is logarithmic in the size of H, and proportional to $n^{-1/2}$, where n is the number of samples. For fixed or moderately growing K, this error goes to zero as n goes to infinity. If we denote by h^* the minimizer of $R(h)$ over H, then we can write the estimation error as

$$R(\hat{h}) - R(h^*) = \hat{R}(\hat{h}) - \hat{R}(h^*) + R(\hat{h}) - \hat{R}(\hat{h}) + \hat{R}(h^*) - R(h^*) \leq 2 \sup_{h \in H} |R(h) - \hat{R}(h)| \tag{4.1}$$

As a first step towards bounding the supremum, we need to bound the difference $|R(h) - \hat{R}(h)|$ of an individual, fixed h. The key ingredient for such a bound is a concentration of measure inequality known as Hoeffding's bound.

Theorem 4.2 (Hoeffding's Inequality). Let $Z_1, \ldots, Z_n$ be independent random variables taking values in $[0, 1]$, and let $\bar{Z}_n = (1/n) \sum_{i=1}^{n} Z_i$ be the average. Then for $t > 0$,

$$P\left(|\bar{Z}_n - E[\bar{Z}_n]| > t \right) \leq 2e^{-2nt^2}.$$

Using Hoeffding's Inequality we obtain the following bound on the difference between the empirical risk and the risk of a classifier.

Lemma 4.3. For any classifier h and $\delta \in (0, 1)$,

$$|\hat{R}(h) - R(h)| \leq \sqrt{\frac{\log(2/\delta)}{2n}}$$

holds with probability at least $1 - \delta$.

Proof. Set $Z_i = 1_{\{h(X_i) \neq Y_i\}}$. Then

$$\bar{Z}_n = \frac{1}{n} \sum_{i=1}^{n} Z_i = \hat{R}(h),$$

$$E[\bar{Z}_n] = E[\hat{R}(h)] = R(h),$$

and the Z_i satisfy the conditions of Hoeffding's inequality. Set $\delta = 2e^{-2nt^2}$ and resolve for t, which gives

$$t = \sqrt{\frac{\log(2/\delta)}{2n}}.$$

Hence, by Hoeffding's inequality,

$$P\left(|\hat{R}(h) - R(h)| > t \right) \leq \delta$$

and therefore, by taking the complement,

$$P(|\hat{R}(h) - R(h)| \leq t) = 1 - P(|\hat{R}(h) - R(h)| > t) \geq 1 - \delta,$$

which was claimed. □

Proof of Theorem 4.1. The goal is to bound the supremum

$$\sup_{h \in H} |\hat{R}(h) - R(h)|. \tag{4.2}$$

For this, we use the union bound. Indeed, for each i we can apply Lemma 4.3 with δ/K to show that

$$P\left(|\hat{R}(h_i) - R(h_i)| > \frac{t}{2}\right) \leq \frac{\delta}{K},$$

where $t = \sqrt{\frac{2\log(2K/\delta)}{n}}$. The probability of (4.2) being bounded by t can be expressed equivalently as

$$\begin{aligned} &P(\sup_{h \in H} |\hat{R}(h) - R(h)| \leq t/2) \\ &= P(|\hat{R}(h_1) - R(h_1)| \leq t/2, \ldots, |\hat{R}(h_K) - R(h_K)| \leq t/2). \end{aligned}$$

Since the right-hand side is an intersection of events, the complement of this event is the union of the events $|\hat{R}(h_i) - R(h_i)| > t$, and we can apply the union bound:

$$\begin{aligned} P(\sup_{h \in H} |\hat{R}(h) - R(h)| > t/2) &\leq \sum_{i=1}^{K} P(|\hat{R}(h_i) - R(h_i)| > t/2) \\ &\leq \delta K \cdot \frac{1}{K} = \delta. \end{aligned}$$

Therefore, with probability at least $1 - \delta$ we have

$$\sup_{h \in H} |\hat{R}(h) - R(h)| \leq \sqrt{\frac{2\log(2K/\delta)}{n}},$$

and using (4.1) the claim follows. □

One drawback of the bound in Theorem 4.1 is that it does not take into account properties of the underlying distribution. In particular, it is the same in the case where Y is completely determined by X as it is in the case in which Y is completely independent on X. Intuitively, in the first situation we would hope to get better rates of convergence than in the second. We will see that using concentration inequalities such as the Bernstein inequality, that take into account the variance of the random variables, we can get better rates of convergence in situation in which the "variance" is not too big.

Notes

5

Probably Approximately Correct

As before, we consider a fixed dictionary H and select one classifier $\hat{h}$ that optimizes the empirical risk $\hat{R}(h)$. Recall:

- The empirical risk $\hat{R}$ and the classifier $\hat{h}$ depend on the data (X_i, Y_i), $1 \le i \le n$, and are random variables. In particular, they depend on n;
- The risk $R(h) = E[\hat{R}(h)]$ depends on the underlying distribution on $X \times Y$, but not on n.

We have seen that if $H = \{h_1, \ldots, h_K\}$ is finite, then with probability $1 - \delta$, we have

$$R(\hat{h}) - \inf_{h \in H} R(h) \le \sqrt{\frac{2\log(K) + 2\log(2/\delta)}{n}}. \quad (5.1)$$

Note that $\log(K)$ is proportional to the bit size of K: this is the amount of bits needed to represent numbers up to K, and can be seen as a measure of complexity for the set H (the "space" necessary to represent K elements). Bounds such as (5.1) are called generalization bounds.

Probably Approximately Correct Learning

An alternative point of view to generalization bounds would be to ask, for given accuracy $\varepsilon > 0$ and confidence $\delta \in (0, 1)$, how many samples n are needed to get an accuracy of ε with confidence $1 - \delta$:

$$P(R(\hat{h}) - \inf_{h \in H} R(h) \le \varepsilon) \ge 1 - \delta.$$

Assuming $h^* = \arg\min_{h \in H} R(h)$, we have $f(X) = R(h^*|X) \in H$ and $Y = f(X)$, and h^* would be the correct classifier. The classifier $\hat{h}$ is then probably (with probability $1 - \delta$) approximately (up to an misclassification probability of at most ε) correct. This leads us to the notion of Probably Approximately Correct (PAC) learning.

In what follows, we denote by $\text{size}(H)$ the complexity of representing an element of H. This is not a precise definition, but depends on the case at hand. For example, if $H = \{h_1, \ldots, h_K\}$ is a finite set, then we can index this set using K numbers. On a computer, numbers up to K can be represented as binary numbers using $\lceil \log_2(K) \rceil$ bits, and hence (up to a constant factor) $\text{size}(H) = \log(K)$ would be adequate here. Similarly, we denote by $\text{size}(X)$ the complexity of representing an element of the input space. For example, if $X \subset \mathbb{R}^d$, then we would use d as size parameter (possibly multiplied by a constant factor to account for the size of representing a real number in floating point arithmetic on a computer). Note that $\text{size}(X)$ or $\text{size}(H)$ is not the same as the cardinality of these sets!

Definition 5.1. (PAC Learning [1]) A hypothesis class H is called PAC-learnable if there exists a classifier $\hat{h} \in H$ depending on n random samples (X_i, Y_i), $i \in \{1, \ldots, n\}$, and a polynomial function $p(x, y, z, w)$, such that for any $\varepsilon > 0$ and $\delta \in (0,1)$, for all distributions on (X, Y),

$$P(R(\hat{h}) \leq \inf_{h \in H} R(h) + \varepsilon) \geq 1 - \delta$$

holds whenever $n \geq p(1/\varepsilon, 1/\delta, \text{size}(X), \text{size}(H))$. We also say that H is efficiently PAC-learnable, if the algorithm that produces $\hat{h}$ from the data runs in time polynomial in $1/\varepsilon$, $1/\delta$, size(X) and size(H).

Remark 5.2. In our context, to say that an algorithm "runs in time p(n)" means that the number of steps, with respect to some suitable model of computation, is bounded by p(n). Note that in this definition we disregard specific constants in the lower bound on n, but only require that it is polynomial. In computer science, polynomial time or space is considered efficient, while problems that require exponential time and/or space to solve are considered inefficient. For example, sorting n numbers can be performed in $O(n \log(n))$ operations and is efficient, while it is not known if finding the shortest route through n cities (the Traveling Salesman Problem) can be solved in a number of computational steps that is polynomial in n. This is the subject of the famous P vs NP conjecture.

In the case of a finite hypothesis space H with K elements, we have seen that H is PAC-learnable, since $n \geq \frac{1}{2\varepsilon^2}(\log(K) + \log(\frac{1}{\delta}))$ which is polynomial in all the relevant parameters.

Generalization bounds and noise

We conclude by commenting briefly on an improvement of the generalization bound (5.1) when incorporating assumptions on the distribution. While the bound (5.1) incorporates the number of samples and the size of H, it does not take into account properties of the distribution, for example, the uncertainty $\varepsilon = Y - f(X)$, where $f(X) = E[Y|X]$ is the regression function. Let $\gamma \in (0, 1/2]$ and assume that

$$|f(X) - 1/2| \geq \gamma$$

almost surely. This condition is known as Massart's noise condition. If $\gamma = 1/2$, then $f(X)$ is either 1 or 0 and we are in the deterministic case, where Y is completely determined by X. If, on the other hand, $\gamma \approx 0$, then we are barely placing any restrictions on $f(X)$, and we are allowing for the case where $f(X)$ is close to 0, and hence where Y is almost independent of X.

Theorem 5.3. Let $H = \{h_1, \ldots, h_K\}$ be a finite dictionary and assume that $h^* \in H$, where h^* is the Bayes classifier. Then for $\delta \in (0,1)$,

$$P(R(\hat{h}) - R(h^*) \leq \frac{1}{\gamma n}(\log \frac{K}{\delta})) \geq 1 - \delta.$$

[1] In some references, such as the book "Foundations of Machine Learning" by Mohri, Rostamizadeh and Talwalkar, this version of PAC learning is called Agnostic PAC Learning.

In the PAC learning context, we see that

$$n \geq \frac{1}{\gamma\varepsilon}(\log(K) + \log(1/\delta))$$

samples are necessary to approximate the Bayes classifier up to a factor of ε with confidence $1 - \delta$. We also see here that the number of samples needed increases as $\gamma \to 0$, reflecting the fact that in the presence of high uncertainty, more observations are needed than if we have low uncertainty. The proof of this result relies on a concentration of measure result similar to Hoeffding's inequality, called Bernstein's inequality.

Theorem 5.4 (Bernstein's Inequality). Let $\{Z_i\}_{i=1}^n$ be centred random variables (that is, $\mathbb{E}[Z_i] = 0$) with $|Z_i| \leq c$ and set $\sigma^2 = \frac{1}{n}\sum_{i=1}^n \mathrm{Var}(Z_i)$. Then for $t > 0$,

$$\mathbb{P}\left(\sum_{i=1}^n Z_i > t\right) \leq e^{-\frac{nt^2}{2(\sigma^2 + ct/3)}}.$$

We outline the idea of the proof. The proof proceeds by defining the random variables

$$Z_i(h) = 1\{h(X_i) = Y_i\} - 1\{h^*(X_i) = Y_i\}$$

for each $h \in \mathcal{H}$. The average and expectation of these random variables is then

$$\hat{R}_n(h) - \hat{R}_n(h^*) = \frac{1}{n}\sum_{i=1}^n Z_i(h),$$
$$R(h) - R(h^*) = \mathbb{E}[Z_i(h)].$$

Based on this, one gets a bound

$$R(\hat{h}_n) - R(h^*) \leq \sup_{h \in \mathcal{H}} \left| \frac{1}{n}\sum_{i=1}^n Z_i(h) - \mathbb{E}[Z_i(h)] \right| \tag{5.2}$$

The random variables $\overline{Z}_i(h)$ are centred, satisfy $|\overline{Z}_i(h)| \leq 2 =: c$ and we can bound the variance by

$$\mathrm{Var}(\overline{Z}_i(h)) = \mathrm{Var}(Z_i(h)) \leq \mathbb{P}(h(X) \neq h^*(X)) =: \sigma^2(h).$$

We can now apply Bernstein's inequality to the probability that the sum (5.2) exceeds a certain value for each individual h, and use a union bound to get a corresponding bound for the maximum that involves the variance $\sigma^2(\hat{h}_n)$. Using the property that $h^* \in \mathcal{H}$, one can also derive a lower bound on the excess risk in terms of the variance, and hence combine both bounds to get the desired result.

Notes

6

Learning Shapes

So far we have considered learning with finite dictionaries of classifiers H. We now illustrate an example where H is not finite, and show how PAC-learnability and generalization bounds can be derived in this setting. We then move on to the more general framework of Rademacher complexity.

Learning Rectangles

Assume that our data describes a rectangle: the input space X is a subset of R2, and the function f: 0,Xistheindicatorfunctionofaclosedrectangle B,sothatforanypointx,f(x)=1ifxis in the rectangle, and 0 if not. Suppose that all we have access to is a random sample of labelled points, (xi,yi){i 1,...,n } (see Figure 6.1, left panel).

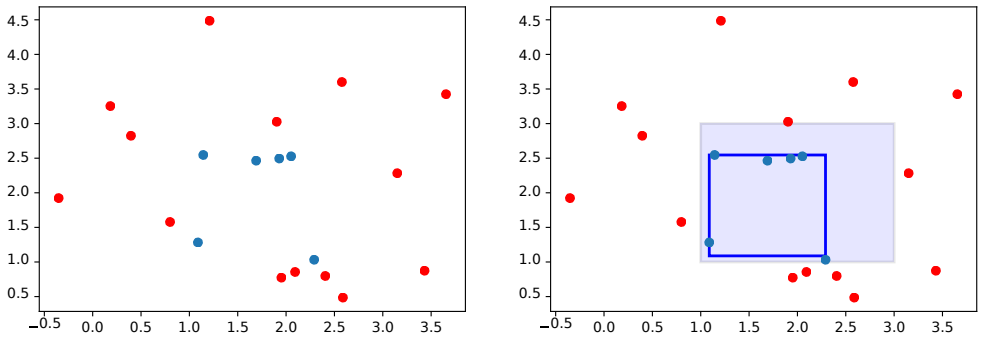


Figure 6.1: The blue points are in the true rectangle, while the red points are outside of it. The smaller blue rectangle represents the computed classifier ĥ ˆ, while the shaded are corresponds to the true rectangle that generated the original labelling.

We compute a candidate ĥ ˆ : R2 → {0, 1} as the indicator function of the smallest enclosing rectangle of the point with label 1 (i.e., the blue points in Figure 6.1). It is clear that if we have lots of sampled points, then we should get a good approximation to the "true" rectangle that generated the data, while with few points this is not possible. How can we quantify this fact? Let ε and δ ∈ (0, 1) be given, and let X be a random point with associated label Y = f(X). Then P(ĥ ˆ(X) = Y) is the probability measure of the true rectangle, while

is the risk of $\hat{h}^*$. We would like to find out the number of samples that would ensure

$$P(R(\hat{h}) \leq \varepsilon) \geq 1 - \delta.$$

First, note that since the rectangle defined by $\hat{h}$ is always contained in the true rectangle that we would like to discover, we can only get false negatives from $\hat{h}^*$ (that is, if $h(x)$ then x is in the true rectangle, but there may be points in the true rectangle for which $\hat{h}(x) = 0$). Hence,

$$R(\hat{h}) = P(\hat{h}^*(X) = 0, f(X) = 1)$$
$$= P(\hat{h}^*(X) = 0 \neq f(X)) = P(\hat{h}^*(X) = 0 = \hat{h}(X), f(X) = 1).$$

If we denote by $\hat{B}^* = \{x \in \mathbb{R}^2 : \hat{h}^*(x) = 1\}$ the smallest enclosing rectangle, then the risk $R(\hat{h})$ can be described more geometrically as

$$R(\hat{h}) = P(X \in B \setminus \hat{B}^*),$$

namely the probability of an input being in the true rectangle but not in the computed one.
Now let $\varepsilon > 0$ and $\delta \in (0,1)$ be given. If $P(X \in B) \leq \varepsilon$, then clearly also $R(\hat{h}^*) \leq \varepsilon$. Assume therefore that $P(X \in B) > \varepsilon$. Denote by R_i, $i \in \{1,2,3,4\}$, the smallest sub-rectangles or B with $P(X \in R_i) \geq \varepsilon/4$ that bound each of the four sides of B, respectively (see Figure 6.2). We could, for example, start with the whole rectangle and move one of its sides towards the opposite side for as long as the measure is not less than $\varepsilon/4$.

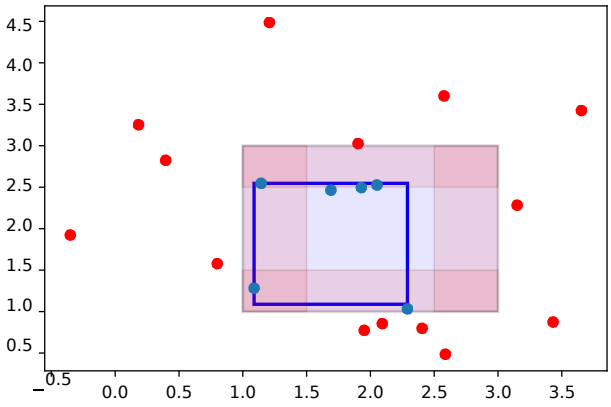


Figure 6.2: Four boundary regions with probability mass $\varepsilon/4$ each.

Denote by R_i° the rectangles with their inward-facing sides removed. Then clearly the probability measure of the union of these sets is $P(X \in \bigcup_i R_i^\circ) \leq \varepsilon$, since the measure of each of the R_i° is at most $\varepsilon/4$. If the computed rectangle $\hat{B}^*$ intersects all the R_i, then

$$P(X \in B \setminus \hat{B}^*) = P(X \in \bigcup_i R_i^\circ) \leq \varepsilon.$$

We now need to show that the probability that $\hat{B}$ does not intersect all the rectangles is small:

$$P(\exists i : \hat{B} \cap R_i = \emptyset) = P(\{\hat{B} \cap R_i = \emptyset\}) \leq \sum_{i=1}^{4} P(\hat{B} \cap R_i = \emptyset),$$

where we used the union bound. The probability that $\hat{B}$ does not intersect one of the rectangles Ri, each of which has probability mass ε/4, is equal to the probability that the n randomly sampled points that gave rise to $\hat{B}$ do not fall in Ri. For each of these points, the probability of not falling into Ri is at most $1 - \varepsilon/4$, so the probability that none of the points falls into is ε/4). Hence, $(1 - \varepsilon/4)^n$

$$P(\exists i : \hat{B} \cap R_i = \emptyset) \leq 4(1 - \varepsilon/4)^n \leq 4e^{-n\varepsilon/4},$$

where we used the inequality $1 - x \leq e^{-x}$. Setting the righ-hand side to δ, we conclude that if

$$n \geq \frac{4}{\varepsilon} \log \frac{4}{\delta}$$

then $R(\hat{h}) \leq \varepsilon$ holds with probability at least $1 - \delta$.

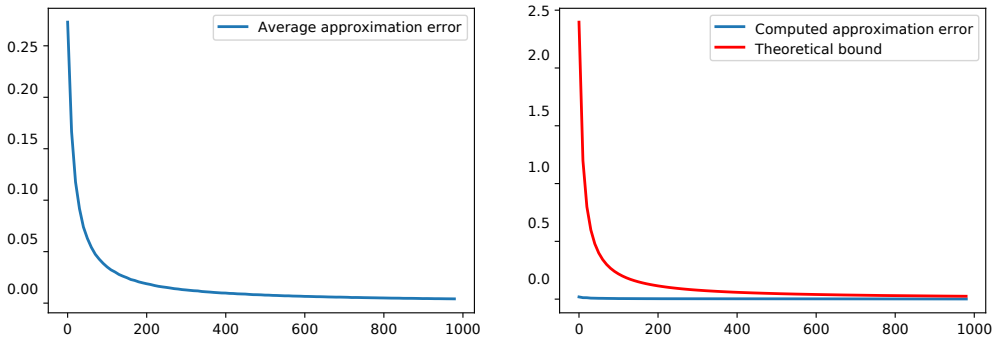


Figure 6.3: The left graphs shows the average risk as n increases. The right right graph shows the risk bound ε when given a confidence $1 - \delta = 0.99$ (blue curve), and the theoretical generalization bound as derived in the example.

Remark 6.1. Note that we did not make any assumptions on the probability distribution when deriving the bound on n in the rectangle-learning example. If the distribution is absolutely continuous with respect to the Lebesgue measure on R2, then we could have required the probability measures of the rectangles to be exactly ε/4, but the way the proof is written it also applies to distributions that are not supported on all of R2, such as the uniform distribution on a compact subset of R2 that may or may not cover the area of B, or a discrete distribution supported on countably many points. The requirement $P(X \in B) > \varepsilon$ still ensures that enough probability mass is contained within the confines of B for the argument to work. We may, however, end up looking at degenerate cases where, for example, all the probability mass is on an edge, one of the rectangles Ri is an edge or the whole of B, etc. Note that in such cases the intuitive view of the generalization risk as the "area" of the complement $B \setminus \hat{B}$ is no longer accurate! In practice we will only consider distributions that are natural to the situation under consideration.

Notes

7

Rademacher Complexity

The key to deriving generalization bounds for sets of classifiers H was to bound the maximum possible difference between the empirical risk and its expected value,

$$\sup_{h \in H} \left| \hat{R}(h) - R(h) \right| = \sup_{h \in H} \left| \frac{1}{n} \sum_{i=1}^{n} \{h(X_i) = Y_i\} - E\left[\{h(X_i) = Y_i\}\right] \right|. \tag{A}$$

For finite H, the probability that this quantity exceeds a given t can be bounded using the union bound, but this approach does not work for infinite sets. We therefore derive a different method that is based on intrinsic complexity measures of H. The first such measure that we will encounter is the Rademacher complexity.

Rademacher Complexity

In the following, we write $Z_i = (X_i, Y_i) \in Z = X \times Y$ for the set of (random) pairs of samples and labels and $z = (x, y) \in Z$ denote by $i \in \{1, \ldots, n\}$, for the set of (random) pairs of samples and labels. To every $h \in H$ we associate a function $g: Z \to \{0, 1\}$ by setting

$$g(z) = \{h(x) \neq y\},$$

and we denote the class of these functions by G.[1] Using this notation, we get

$$\sup_{h \in H} \left| \hat{R}(h) - R(h) \right| = \sup_{g \in G} \left| \frac{1}{n} \sum_{i=1}^{n} g(Z_i) - E\left[g(Z_i)\right] \right|.$$

We will bound this expression in terms of a property of the set G, the Rademacher complexity. For what follows, we say that a random variable has the Rademacher distribution if it takes the values 1 or −1 with probability 1/2 each. When an expression potentially depends on different random quantities, for example random variables X and Y , we write E_X to denote the expectation with respect to only X.

Definition 7.1. (Rademacher Complexity) Let G be a class of real-valued functions on Z. Let $\sigma = (\sigma_1, \ldots, \sigma_n)$ be a random vector, such that the σ_i are independent Rademacher random variables, and let $z = (z_1, \ldots, z_n)$. The empirical Rademacher complexity of the family of functions G with respect to z is defined as

$$\hat{R}_z(G) = E_\sigma\left[\sup_{g \in G} \frac{1}{n} \sum_{i=1}^{n} \sigma_i \cdot g(z_i)\right]. \tag{7.1}$$

[1] We could, and will eventually, use any other loss function.

The Rademacher complexity is the expectation:

$$R() = EZ[nG\hat{R}\hat{}Z(G)]. \quad (7.2)$$

Remark 7.2. Note that the expected value in (7.1) is only over the random signs, and that the point z is fixed. It is only in (7.2) that we replace the point by a random vector Z = (Z1,...,Zn) and take the expectation. As the distribution of σ is discrete, we could also rewrite the empirical Rademacher complexity as average:

$$\hat{R}\hat{}z(\hat{G}i = \frac{1}{2n}\sum(\sum)n1\sup\sigma i\cdot g(zi).n\sigma\in\{-ng\in G1,1i=1$$

}

Note also that the definition works for any set of real-valued functions g, not only those that arise from the loss function applied to a classifier. In the literature there are various variations on the notion of Rademacher complexity. It can be defined simply for sets: Given a set S $\subset$ Rn, the Rademacher complexity of S is defined as

$$R(S) = E\sigma\sup\sigma ixi. \quad (7.3)$$

Ournotionisaspecialcase:whenS 1)): =(g(z),...,g(zng z()= $\in$ G}, then we have Using this notion of complexity, we can derive the following bound.

Theorem 7.3. Let $\delta \in (0, 1)$ be given. Then with probability at least $1-\delta$ over $z \in Z n$, the following holds for all $g \in G$:

$$E[g(Z)] \le \frac{1}{n}\sum_{i=1}^{n} g(zi) + 2Rn(G) + \sqrt{\frac{\ln(1/\delta)}{n}}. \quad (7.4)$$

Before going into the proof, we list some examples.

Example 7.4. Assume H consists of only one function. Then for any point $(z1,...,zn) \in Z n$, the values $g(zi) = 1\{h(xi) = yi\}$ form a fixed 0-1 vector . The empirical Rademacher complexity is

$$\hat{R}(G) = \frac{1}{2}E\sigma\left[\left|\sum_{i=1}^{n}\sigma ig(zi)\right|\right] = \frac{1}{2}E\sigma\left[\left|\sum_{i=1}^{n}\sigma ig(zi)\right|\right]$$

since for each i, g(zi) and $-g(zi)$ appear an equal number of times when averaging over all possible sign vectors.

Example 7.5. Let H be the set of all binary classifiers. It follows that for any $(z1,...,zn) \in Z n$, the set of vectors $(g(z1), \ldots, g(zn))$ for $g \in G$ runs through all binary 0-1 vectors. The empirical Rademacher complexity of the set of functions $G = \{h(\cdot) = 1\{h(\cdot) = yi\} : h \in H\}$ is thus the same as the Rademacher complexity of the hypercube $S = [0,1]n$ as a set. Note that for each sign vector σ we can always pick out a function $g \in G$ such that $g(zi) = 1$ if $\sigma i = 1$ and $g(zi) = 0$ if $\sigma i =$, and this function maximizes the sum $\sum \sigma ig(zi)$. From this observation it is not hard to conclude that

$$\hat{R}(z, G) \le \frac{r\sqrt{2\log(2n)}}{n},$$

Example 7.6. We will see that for a finite set $\{z1,...,zn\}$ and $G = \{g1,...,gK\}$, we get the bound where $r = \max_{g\in G}\sqrt{\sum_{i=1}^{n} g(zi)^2}$. This bound is known as Massart's Lemma.

Example 7.7. The Rademacher complexity of a set S equals the Rademacher complexity of the convex hull of S. For example, the Rademacher complexity of the hypercube [0, 1]n equals the Ra

√demachercomplexity√ofthesetofitsven√rtices.Sincethereare2vertices,Example7.6givesthebound2log(2)

(weusedthatr=nhere).WesawinExample7.5thattheexactvalueis1/2.

The proof of Theorem 7.3 depends on yet another concentration of measure inequality for averages of random variables, namely McDiarmid's inequality.

Theorem 7.8 (McDiarmid's Inequality). Let $Z_1, \ldots, Z_n$ be a set of independent random variables defined on a space Z, let $f : Z \to \mathbb{R}$ be a function such that for all $i \in \{1, \ldots, n\}$ and for all $z_1, \ldots, z_n, z_i'$,

$$|f(z_1, \ldots, z_i, \ldots, z_n) - f(z_1, \ldots, z_i', \ldots, z_n)| \leq c_i.$$

Then for all $t > 0$, the following inequalities hold:

$$P(f(Z_1, \ldots, Z_n) - E[f(Z_1, \ldots, Z_n)] + t) \leq e^{-\frac{2t^2}{\sum_{i=1}^n c_i^2}}$$
$$P(E[f(Z_1, \ldots, Z_n)] - f(Z_1, \ldots, Z_n) + t) \leq e^{-\frac{2t^2}{\sum_{i=1}^n c_i^2}} \tag{7.5}$$

Using the union bound we can combine the two inequalities (7.5) to one inequality for the absolute value $|f(Z_1, \ldots, Z_n) - E[f(Z_1, \ldots, Z_n)]|$, with an additional factor of 2 in front of the exponential bound. Note that McDiarmid's inequality contains Hoeffding's inequality as a special case when f is the average.

Proof of Theorem 7.3. Define the function

$$\Phi(z_1, \ldots, z_n) = \sup_{g \in G} \left(E[g(Z_i)] - \frac{1}{n} \sum_{i=1}^n g(z_i) \right).$$

Then for $i \in \{1, \ldots, n\}$ and z_i', and using the fact that the difference of sup remains not bigger than the supremum of a difference, we get

$$\Phi(z_1, \ldots, z_i, \ldots, z_n) - \Phi(z_1, \ldots, z_i', \ldots, z_n) \leq \sup_{g \in G} \frac{1}{n} (g(z_i') - g(z_i)) \leq \frac{1}{n},$$

from which we conclude that

$$|\Phi(z_1, \ldots, z_i, \ldots, z_n) - \Phi(z_1, \ldots, z_i', \ldots, z_n)| \leq \frac{1}{n}.$$

The function Φ thus satisfies the conditions of McDiarmid's inequality with $c_i = 1/n$, and from this inequality we get

$$P(\Phi(Z_1, \ldots, Z_n) - E[\Phi(Z_1, \ldots, Z_n)] > t) \leq -2n^2 t^2 / n) = e^{-2nt^2}.$$

Setting the right-hand bound to δ and resolving for t, we conclude that with probability at least $1 - \delta$,

$$\Phi(Z_1, \ldots, Z_n) \leq E[\Phi(Z_1, \ldots, Z_n)] + \sqrt{\frac{\log(1/\delta)}{2n}}.$$

The last, and crucial, step is to bound the expected value on the right-hand side with the Rademacher complexity. The idea is to introduce identical but independent copies Z_i' of the random variables Z_i. Denote by Z and Z' the vectors of random variables Z and Z_i', respectively. We will use repeatedly the

fact that if f(Z′) only depends on the random variables in Z′, then f(Z′) = E ′Z[f(Z)], that is we can pull an expression "into the expectation" if the terms involved are independent of the variables over which the expectation is taken. We will also use the linearity of expectation repeatedly without explicitly saying so. We can then set

$$
\begin{aligned}
E_Z[\Phi(Z_1,\ldots,Z_n)] &= E[\ ]\\
&= E\\
&= E\\
&\le E\\
&= [Z\sup(g(Z_i)-g(Z_i))\,g\in G\,i=1\,n\,1\,E'',Z\sup(g(Z_i)-g(Z_i)),g\in G\,n\,i=1
\end{aligned}
$$

where for the inequality we used the fact that the sup of an expectation is not more than the expectation of the sup. We next use an idea known as symmetrization. The key observation is that each summand $g(Z'_i) - g(Z_i)$ is just as likely to be positive as it is to be negative. In other words, if we replace $g(Z'_i) - g(Z_i)$ with its negative $g(Z_i) - g(Z'_i)$ then the above expectation does not change. More generally, we can pick any sign vector $\sigma = (\sigma_1, \ldots, \sigma_n)$ with $\sigma_i \in \{-1, 1\}$, and will then have

$$E_{Z,Z'}[\ ] = E_{Z,Z'}[\ ] \cdot (g(Z'_i) - g(Z_i)).$$

Now we use a "sheep counting trick"[2], and sum these terms over all possible sign patterns, dividing by the total number of sign patterns, 2^n:

$$E_{Z,Z'}[\ ] = \frac{1}{2^n}\sum_{\sigma}[\ ] = E_{\sigma,Z,Z'}[\ ].$$

where for the last equality we simply rewrote the average as an expectation over a vector of Rademacher random variables. We can now bound the supremum of the differences by the difference of suprema in order to get

$$E_{\sigma,Z,Z'}[\ ] = 2R_n(G).$$

[2] When a shepherd wants to count her sheep, she counts the legs and divides the result by four.

The last equality shows why we averaged over all sign vectors: by symmetry, averaging over $-\sigma$ is the same as averaging over σ. □

The empirical Rademacher complexity $\hat{R}_z$ is a function of $(z_1,\ldots,z_n)$, and by the same argument as for the Φ function in the proof of Theorem 7.3 one can show that if z' arises from z by changing z to z'_i, then $|\hat{R}_z(G) - \hat{R}_{z'}(G)| \leq \frac{1}{n}$.

We can therefore apply McDiarmid's inequality to the random variable $\hat{R}_z(G)$ to conclude that

$$R(G) \leq \hat{R}_z(G) + \sqrt{\frac{\log(1/\delta)}{2n}} \tag{7.6}$$

with probability at least $1 - \delta$. One can combine (7.6) with (7.4) using the union bound to get a generalization bound analogous to (7.4) but in terms of the empirical Rademacher complexity (with slightly different parameters).

To conclude, note that the inequalities (7.4) and (7.6) are one-sided inequalities: they bound a difference but not the absolute value of this difference. These can easily be adapted to give a bound on the supremum of the absolute difference $|\hat{R}_z(h) - R(h)|$. As a consequence of Example 7.5 we see that when considering the set of all binary classifiers, we do not get a generalization bound that converges to 0 as $n \to \infty$ using the Rademacher complexity bound.

Notes

8

VC Theory

As usual we operate on a pair of input-output spaces $Z = X \times Y$ with $Y = \{-1, 1\}$. Let H be a set of classifiers with associated set

$$G = \{g : Z \to \{0, 1\} : g(x, y) = 1 \{h(x) \neq y\}, h \in H\}.$$

We saw that we could bound the maximum difference between the generalization risk $R(h)$ and the empirical risk $\hat{R}_n(h)$ of a classifier using the Rademacher complexity $R_n(G)$:

$$\sup_{h \in H} (R(h) - \hat{R}_n(h)) \leq 2R_n(G) + \sqrt{\frac{\log(1/\delta)}{2n}}. \tag{8.1}$$

Instead of considering the set G, we can also consider the Rademacher complexity of H itself. For what follows, assume that the classifiers in H take the values $\{-1, 1\}$.

Lemma 8.1. The Rademacher complexities of H and G satisfy $R_n(G) = \frac{1}{2} R_n(H)$.

The proof depends on writing $1\{h(x) \neq y\} = (1 - yh(x))/2$, and is left as an exercise. The definition of Rademacher complexity involved taking the expectation with respect to a distribution on the space Z that we do not know. We saw, however, that in some examples we could bound this expectation in a way that does not depend on the distribution. We next develop this idea in a more principled way, deriving generalization bounds in terms of parameters of the set of classifiers H that do not make reference to the underlying distribution. The theory is named after Vladimir Vapnik and Alexey Chervonenkis, who developed it in the 1960s.

Vapnik-Chervonenkis Theory

In binary classification, a classifier h can take at most two possible values on a fixed input. Denote by $x = (x_1, \ldots, x_n) \in X^n$ an n-tuple of inputs and set

$$h(x) := (h(x_1), \ldots, h(x_n)) \in \{-1, 1\}^n$$

for the corresponding vector of values of the classifier. For any fixed $x \in X^n$, we could get up to 2^n different values $h(x)$ as h runs through H. Note that his is a finite bound even if the set H is infinite! A possible classification $h(x)$ is called a dichotomy and one way to measure the expressiveness or richness of a set of classifiers H would be to count the possible dichotomies.

Example 8.2. Let $X = R$ and let H be the set of indicator functions of closed half-lines: for each $a \in R$,

$$h^+_a(x) = \begin{cases} 1 & x \geq a, \\ -1 & x < a \end{cases} \qquad h^-_a(x) = \begin{cases} 1 & x \leq a. \\ -1 & x > a \end{cases}$$

Given two distinct samples, $\{x, x, \text{there are 4 possible dichotomies: for each pattern} \in \{-1,1\}^2$ we can find $h \in H$ such that the tuple $(h(x_1), h(x_2)) = \rho$. For three distinct points $\{x_1, x_2, x_3\} \subset R$ this is no longer possible. If we assume, for example, that $x_1 < x_2 < x_3$, then any classifier h with $h(x_1) = -1$ and $h(x_2) = 1$ will automatically also satisfy $h(x_3) = 1$.

Clearly, if H consists of only one element, then for every $x \in X^n$ there is only one possible dichotomy, while if H consists of all classifiers then there are 2^n possible dichotomies if the entries of x are distinct. Somewhere in between these two extreme cases, the number of dichotomies may depend on x in more intricate ways.

Definition 8.3. Let H be a set of classifiers h. The growth function Π_H is defined as

$$\Pi_H(n) = \max_{x \in X^n} |\{(h(x_1), \ldots, h(x_n)) : h \in H\}|.$$

Note that this function depends only on the set H. We can use it to bound the Rademacher complexity of a set of functions G. The bound depends on a result known as Massart's Lemma, which is derived in the Exercises. Recall that the Rademacher complexity of a set $S \subset R^n$ is defined as

$$R_n(S) = \frac{1}{n} \mathbb{E}_\sigma \sup_{x \in S} \sum_{i=1}^n \sigma_i x_i$$

where $\sigma = (\sigma_1, \ldots, \sigma_n)$ is a vector of i.i.d. Rademacher random variables ($P(\sigma_i = +1) = P(\sigma_i = -1) = 1/2$) and $x = (x_1, \ldots, x_n)$.

Lemma 8.4. (Massart's Lemma) If $S = \{x_1, \ldots, x_K\} \subset R^n$ consists of K elements, then

$$R_n(S) \leq r \sqrt{\frac{2\log(K)}{n}},$$

where $r = \max_{x \in S} \|x\|$ and $\|x\|^2 = \sum_{i=1}^n x_i^2$.

Theorem 8.5. The Rademacher complexity of H such that

$$R_n(H) \leq \sqrt{\frac{2\log(\Pi_H(n))}{n}}$$

Proof. Consider a fixed tuple $x = (x_1, \ldots, x_n)$

$$\Pi_H(x) = |\{(h(x_1), \ldots, h(x_n)) : h \in H\}|$$

The vectors $(h(x_1), \ldots, h(x_n))$ all have norm $\sqrt{n}$, and the claim follows from Massart's Lemma. □

Combining this with (8.1) we immediately arrive at the following bound.

Corollary 8.6. For all $h \in H$,

$$R(h) \leq \hat{R}(h) + \sqrt{\frac{2\log(\Pi_H(n))}{n}} + \sqrt{\frac{\log(1/\delta)}{2n}} \tag{8.2}$$

Let H be a set of classifiers and $S = \{x_i\}_{i=1}^n \subset X$ a set of inputs. Then S is shattered by H, if

$$|\{(h(x_1), \dots, h(x_n)) : h \in H\}| = 2^n,$$

that is, if all dichotomies are possible.

Example 8.7. If H is the set of all binary classifiers and all the samples in S are distinct, then S is shattered by H.

Example 8.8. If $H = \{h_1, h_2\}$ consists of only two classifiers, one of which is constant 1 and the other is constant 0, then a subset $S \subset X$ of samples is shattered by H if and only if $|S| = 1$, that is, we only look at one sample.

If the number of samples n increases, it can become harder to find a subset that can be shattered. While in the case where H consists of all binary classifiers we always have $\Pi_H(n) = 2^n$, in the case where H consists of indicator functions of half-lines we saw that three distinct points on the line cannot be shattered.

There exists a subset $S = \{x_i\}_{i=1}^n$ that can be shattered by H if and only if $\Pi_H(n) = 2^n$.

The maximum possible n such that a subset of n points can be shattered is the VC dimension.

Definition 8.9. The Vapnik-Chernovenkis (VC) dimension of a set of classifiers H is

$$VC(H) = \max\{n \in \mathbb{N} : \Pi_H(n) = 2^n\}.$$

If $VC(H) = d$, then there exists a set of $n = d$ samples that can be shattered by H: all possible dichotomies occur when applying classifiers in H to these samples. Note that this does not mean, however, that all sets of n samples can be shattered by H.

We will see that if H has VC dimension d, then we can bound $\log(\Pi_H(n)) \le d \log\left(\frac{en}{d}\right)$, which allows to rephrase the bound (8.2) in terms of the VC dimension. We will then study plenty of examples (including practical ones) where we can compute or bound the VC dimension.

Notes

9

The VC Inequality

The VC dimension of a hypothesis set H is the maximal cardinality of a subset of the input space X that is shattered by H:

$$VC(H) = \max\{n \in \mathbb{N} : \Pi_H(n) = 2^n\},$$

where the growth function $\Pi_H(n)$ counts the number of possible dichotomies,

$$\Pi_H(n) = \max_{(x_1,\ldots,x_n) \in X^n} |\{(h(x_1),\ldots,h(x_n)) : h \in H\}|.$$

The VC dimension is a combinatorial quantity that depends only on H. It acts as a surrogate for cardinality when dealing with infinite sets.

VC dimension of families of sets

The notion of VC dimension was defined for classes of functions H. Equivalently, we can identify each h with the indicator function of a set $A \subset X$ and consider the set of sets

$$\mathcal{A} = \{A \subset X : h(x) = 1 \Leftrightarrow x \in A\}.$$

Given a subset $S \subset X$, we say that $\mathcal{A}$ shatters S if every subset of S (including the empty set) can be obtained by intersecting S with elements of $\mathcal{A}$:

$$\{A \cap S : A \in \mathcal{A}\} = \mathcal{P}(S) := \{S' : S' \subset S\}.$$

In other words, we can use the collection $\mathcal{A}$ to select any subset of S. It should be clear that if S is shattered by $\mathcal{A}$, then so is any subset of S. In this context, the growth function is defined analogously,

$$\Pi_{\mathcal{A}}(n) = \max\{|\{A \cap S : A \in \mathcal{A}\}| : S \subset X, |S| = n\},$$

as the maximal number of subsets of a set of n that can be selected using $\mathcal{A}$. The VC dimension of $\mathcal{A}$ is the cardinality of the set system $\mathcal{A}$:

$$VC(\mathcal{A}) = \max\{n \in \mathbb{N} : \Pi_{\mathcal{A}}(n) = 2^n\}.$$

Note that the VC dimension is monotone in the sense that it does not decrease when $\mathcal{A}$ is enlarged. One of the most important results in VC theory is a bound on $\Pi_{\mathcal{A}}(n)$ in terms of the VC dimension. This result is usually attributed to Sauer (who credits Perles) and Shelah, but was discovered independently by Vapnik & Chervonenkis.

Lemma 9.1. If $VC(A) = d$, then for $n \geq d$,

$$\Pi_{A(n)} \leq \sum_{i=0}^{d} \binom{n}{i}.$$

Proof. The proof of the first inequality is by induction on $n + d$. If $d = 0$ and $n = 0$, then $\Pi_{A(0)} = 1$ (only the empty set can be considered) and the bound is valid. The statement is also easily verified for $n = 1$ and $d \geq 1$. Assume now that $n > 0$ and $d > 0$, and that the statement holds for the pairs $(d, n-1)$ and $(d-1, n-1)$. Let $S \subset X$ be a subset with $|S| = n$, such that $|\{A \cap S : A \in \mathcal{A}\}| = \Pi_A(n)$, and select an arbitrary element $s \in S$. Consider the sets

$$\mathcal{A}' = \{A \setminus \{s\} : A \in \mathcal{A}\}, \quad \mathcal{A}'' = \{A \in \mathcal{A} : s \notin A, A \cup \{s\} \in \mathcal{A}\}.$$

Note that

$$VC(\mathcal{A}') \leq VC(\mathcal{A}) \leq d, \text{ and } VC(\mathcal{A}'') \leq VC(\mathcal{A}) - 1 = d - 1.$$

The first inequality follows from the fact that if $\mathcal{A}'$ shatters a set T, then so does $\mathcal{A}$. The second inequality follows from the fact that if a set T is shattered by $\mathcal{A}''$, then the set $T \cup \{s\}$ is shattered by $\mathcal{A}$.

Consider the map

$$\{A \cap S : A \in \mathcal{A}\} \to \{A \cap S : A \in \mathcal{A}'\} = \{A \cap S \setminus \{s\} : A \in \mathcal{A}\},$$
$$A \cap S \to A \cap S \setminus \{s\}.$$

This map is one-to-one, except in the case where $A \in \mathcal{A}$ and $\tilde{A} = A \cup \{s\} \in \mathcal{A}$. For such A, both hold, and therefore $A \cap S = \tilde{A} \cap S$, but

$$A \cap S \setminus \{s\} = \tilde{A} \cap S \setminus \{s\}.$$

It follows that

$$|\{A \cap S : A \in \mathcal{A}\}| = |\{A \cap S \setminus \{s\} : A \in \mathcal{A}'\}| + |\{A \cap (S \setminus \{s\}) : A \in \mathcal{A}''\}|.$$

By the induction hypothesis,

$$|\{A \cap (S \setminus \{s\}) : A \in \mathcal{A}'\}| \leq \sum_{i=0}^{d-1} \binom{n-1}{i} \leq \binom{n-1}{i},$$

$$|\{A \cap (S \setminus \{s\}) : A \in \mathcal{A}''\}| \leq \sum_{i=0}^{d-1} \binom{n-1}{i} \leq \binom{n-1}{i},$$

so that combining the terms we get

$$|\{A \cap S : A \in \mathcal{A}\}| \leq \sum_{i=0}^{d} \binom{n-1}{i} + \sum_{i=0}^{d-1} \binom{n-1}{i} = \sum_{i=0}^{d} \binom{n}{i}.$$

For the second claimed inequality, we extend the sum to n and multiply each summand by $(n/d)^{d-i}$, to obtain

$$\sum_{i=0}^{d} \binom{n}{i} \leq \sum_{i=0}^{n} \binom{n}{i} \frac{n^{d-i}}{d^{d-i}} = \frac{n^d}{d^d} \sum_{i=0}^{n} \binom{n}{i} \frac{d^i}{n^i} = \frac{n^d}{d^d} \left(1 + \frac{d}{n}\right)^n \leq \frac{n^d}{d^d} e^d,$$

where we used the inequality $1 + x \leq e^x$ at the end. □

The VC Inequality

A central result in statistical learning is an inequality that relates the VC dimension of a set of classifiers to the difference between the empirical and the generalization risk.

Theorem 9.2. (VC Inequality) Let H be a set of classifiers $h : X \to \{-1, 1\}$ with VC dimension $VC(H) = d$, and let $\delta \in (0, 1)$. Then with probability at least $1 - \delta$,

$$\sup_{h \in H} R(h) - \hat{R}(h) \leq \sqrt{\frac{2d}{n} \log \frac{en}{d}} + \sqrt{\frac{\log(1/\delta)}{2n}}.$$

Proof. In Corollary 8.6, Lecture 8, we saw that

$$\sup_{h \in H} R(h) - \hat{R}(h) \leq \sqrt{\frac{2 \log(\Pi_H(n))}{n}} + \sqrt{\frac{\log(1/\delta)}{2n}}.$$

The claim now follows directly from Lemma 9.1. □

We remark that the bounds here were all stated as one-sided bounds: that is, they are stated as bounds on the difference $R(h) - \hat{R}(h)$ and not the absolute value of the difference. We can get two-sided bounds by making adjustments in the derivation of bounds using the Rademacher complexity (using the second case of McDiarmid's inequality) and arrive at the following bound, which holds with probability at least $1 - \delta$:

$$\sup_{h \in H} \left| R(h) - \hat{R}(h) \right| \leq \sqrt{\frac{2d}{n} \log \frac{en}{d}} + \sqrt{\frac{\log(2/\delta)}{2n}}$$

Note that the only difference is the factor of $2/\delta$ which is a consequence of combining two one-sided inequalities to one two-sided inequality using the union bound.

Rectangle learning revisited

Let H be the set of functions that take the value 1 on rectangles in the plane $X = R^2$ and -1 otherwise[1]. The question is:

[1] Depending on context, we will consider H as consisting of functions into $\{0, 1\}$ or into $\{-1, 1\}$, this does not alter any of the results involving the VC dimension.

Given n, can we find a configuration of n points in the plane such that for any labelling we can find a rectangle containing those points labelled with 1?

This is clearly possible when n = 2 (just choose two distinct points) and n = 3 (choose three points that form a triangle such as 4). For n = 4 there are 16 possible labellings, and if we arrange the points in diamond form , then all labellings can be captured by rectangles (try this!). For n = 5 this is no longer possible: take the smallest enclosing rectangle of five points {x 5i}i=1.Thisrectanglewillcontain(atleast) one of the xi on each boundary (if not, we could make it smaller). If each xi, i ∈{1,...,4},liesona different boundary, we can assign 1 to these points and −1 to x5. This dichotomy cannot be realized by a rectangle: any rectangle containing {x 4i}i=1mustalsocontaintheirsmallestenclosingrectangle,hence also x5.

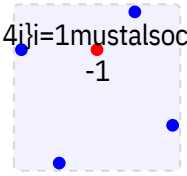


Figure 9.1: A dichotomy that is not captured by rectangles.

Notes

10

General Loss Functions

So far we looked at the problem of binary classification. We considered a set H of classifiers $h: X \to Y$, where Y was a set with two elements ($\{-1, 1\}$ or $\{0, 1\}$, for example). Given a distribution on $X \times Y$ we considered the generalization risk,

$$R(h) = E[1_{\{h(X) \neq Y\}}] = P((h(X) \neq Y), \tag{10.1}$$

and the empirical risk,

$$\hat{R}_n(h) = \frac{1}{n} \sum_{i=1}^{n} 1_{\{h(X_i) \neq Y_i\}}, \tag{10.2}$$

which is based on a sequence of random observations $(X_i, Y_i) \in \{\ldots, n\}$. For each set of realizations $\{(x_i, y_i)\}$ one can (in principle) construct a classifier $\hat{h} \in H$ that minimizes the empirical risk (10.2). We are ultimately interested in the generalization risk and not the empirical risk $\hat{R}(\hat{h})$ (this would just tell us something about how well our classifier works on the training data). Specifically, we are interested in how close the risk $R(\hat{h})$ is to the optimal generalization risk $R(h^*) = \inf_{h \in H} R(h)$. To analyse this, we split the difference into two parts:

$$\begin{aligned} R(\hat{h}) - \inf_{h \in H} R(h) &\leq R(\hat{h}) - \hat{R}(\hat{h}) + \hat{R}(h^*) - \inf_{h \in H} R(h) \\ &\leq 2 \sup_{h} \left| \hat{R}(h) - R(h) \right| \end{aligned}$$

The term within the sup to be bounded is just the difference between the average of i.i.d. random variables and their expectation. To bound this difference we used:

- Hoeffding's or Bernstein's inequality and the union bound for finite H, to obtain a bound in terms of $\log(|H|)$;
- McDiarmid's inequality applied directly to the supremum and symmetrization to obtain a bound in terms of the Rademacher complexity.

The Rademacher complexity could then be bounded, via Massart's inequality, in terms of the growth function of H, which in turn could be bounded again, via the Sauer-Shelah Lemma, in terms of the VC dimension of H. It is natural to ask how much of this depends on the fact that we used binary classification and the unit loss $1_{\{h(X) = Y\}}$, and to what extent these bounds generalize to other types of classifiers and loss functions.

General Loss Functions

Consider a set of function $H = \{h: X \to Y\}$ with $Y = [-1, 1]$ (or any other subset of R), and a loss function L:

$Y \times Y \to R \geq 0$. Besides the unit-loss that we studied so far, examples include:

- $L(x,y) = (x - y)2$ (quadratic loss);
- $L(x,y) = |x - y|$ (`1-loss);
- $L(x,y) = |x - y|p$ (`p-loss);
- $L(x, y) = \log(1 + e-xy)$ (log-loss).

$\in -11] \quad y \in [-1\ 1]$

Sometimes the loss function is scaled to ensure that it is in a certain range. For example, for the quadratic loss, $(h(x) \; R)) = E \; [\; (L(\; n \;) \; X \; , \; Y$ $-y)2/2 \in [0,1]$ if $h(x)[$ and . The generalization riskis defined as while the empirical risk is

$$\hat{R}\hat{}\,1(h) = \frac{}{n} \sum nL(h(Xi),Yi).=1$$

Just as in the binary case with unit loss, we denote by $\bar{h}$ a classifier with optimal generalization risk in H, and by $\hat{h}\hat{}$ a minimizer of $\hat{R}\hat{}(h)$. Also as in the binary case, we can bound

$$R(\hat{h}) - R(\bar{h}) \leq 2 \sup_{h \in H} | \hat{R}\hat{}(h) - R(h) |.$$

Consider the set of functions $[0, 1]: g(z) = L(h(x), y)\}$.

$$L \circ H := \{ g: Z \to z = (z1,\ldots,zn)$$

This plays the role of the set G defined in Lecture 6. For , with $zi = (xi, yi)$, we can define the empirical Rademacher complexity as

$$\hat{R}\hat{}z(L \circ H) = E \left[\sum \right] n1\sigma \sup_{g \in L \circ H} \sigma ig(zi), ni=1$$

and the Rademacher complexity as

$$[]) = E\hat{R}\hat{}ZZ(L \circ H.$$

$$R\,n(L \circ H$$

Assuming that $L(x, y) \in [0, 1]$ (which we can by scaling the loss function accordingly) one can use the McDiarmid's bounded difference inequality and the same symmetrisation argument as in the binary case to derive the bound

$$\sup_{h \in H} | \hat{R}\hat{}(h) - R(h) | \leq 2R(\circ H \sqrt{\log(2/\delta nL)+,2n} \quad H =$$

which holds with probability at least . For finite sets of classifiers same $\{h1,\ldots,hk\}$ we get the cardinality bounds as in the binary case.

Theorem 10.1. Let $H = \{h_1, \ldots, h_K\}$ be a set of classifiers $h : X \to [-1, 1]$, and let L be a loss function taking values in $[0, 1]$. Then

$$R_n(L \circ H) \leq \sqrt{\frac{2\log(K)}{n}}.$$

Proof. Fix z = (z1,...,zn). Then by Massart's Lemma (Lemma 8.4 in Lecture 8), the empirical Rademacher complexity is bounded by

$$\hat{R}_z(L \circ H) \leq r\sqrt{\frac{2\log(K)}{n}},$$

where

$$r = \sup\{\|x\| : x \in S\}, \quad S = \{(g(z_1), \ldots, g(z_n)) : g \in L \circ H\}.$$

Since by assumption $g(z) = L(h(x), y) \in [0, 1]$, we have $r \leq \sqrt{n}$, and the result follows by taking the expectation over Z. □

For infinite H we cannot repeat the arguments used in the case of binary classifiers with unit loss. The bounds based on growth function and VC dimension depend on the fact that the number of possible dichotomies is finite and this is no longer the case when considering functions h with infinite range. One way of dealing with such a situation is to approximate an infinite set by a finite set in such a way, that every element of the infinite set is close to a point in the finite subset.

Notes

11

Covering Numbers

In this lecture we consider hypothesis $H = \{h : X \to Y = [-1, 1]\}$ and a loss function $L : Y \times Y \to$ set $[0, 1]$. Define the set of functions

$$L \circ H = \{g : Z \to [0, 1] \; g(z) = (L(h)x, y)\}.$$

This set plays the role of the set G defined in Lecture 6. For $z = (z_1, \ldots, z_n)$, with $z_i = (x_i, y_i)$, we can define the empirical Rademacher complexity as

$$\hat{R}_z(L \circ H) = E_\sigma \left[\sup_{g \in L \circ H} \frac{1}{n} \sum_{i=1}^{n} \sigma_i g(z_i) \right],$$

where the expectation is over all sign vector $\sigma = (\sigma_1, \ldots, \sigma_n)$ with independent σ_i that satisfy $P\{\sigma_i = +1\} = P\{\sigma_i = 1 = 1/2\}$. The Rademacher complexity is defined as as

$$R_n(L \circ H) = E_Z [\hat{R}_Z(L \circ H)],$$

where the expectation is over all n-tuples $Z_1, \ldots, Z_n)$, where $Z_i = (X_i, Y_i)$. As in the case of binary classification with unit loss, one can use McDiarmid's bounded difference inequality and a symmetrisation argument to derive the bound

$$\sup_{h \in H} \left| \hat{R}(h) - R(h) \right| \leq 2 R_n(L \circ H) + \sqrt{\frac{\log(2/\delta)}{2n}},$$

which holds with probability at least $1 - \delta$. For a finite set of binary classifiers $H = \{h_1, \ldots, h_K\}$, the arguments used in the case of binary classification carry over seamlessly.

Theorem 11.1. Let $H = \{h_1, \ldots, h_K\}$ be a set of functions $h : X \to [-1, 1]$, and let L be a loss function taking values in $[0, 1]$. Then

$$R_n(L \circ H) \leq \sqrt{\frac{2 \log(K)}{n}}.$$

Proof. Fix $z = (z_1, \ldots, z_n)$. Then by Massart's Lemma (Lemma 8.4 in Lecture 8), the empirical Rademacher complexity is bounded by

$$\hat{R}_z(L \circ H) \leq r \sqrt{\frac{2 \log(K)}{n}},$$

where

$$r = \sup\{\|x\| : x \in S\} \qquad S = \{(g(z_1), \ldots, g(z_n)) : g \in L \circ H\}.$$

Since by assumption $g(z) = L(h(x) - y) \in [0, 1]$, and the result follows by taking the expectation over Z. □

For infinite H, we cannot repeat the arguments that lead to the VC inequality in the binary classification case, since these arguments were based on the fact that even for infinite H, the number of possible dichotomies is finite. This limitation can be circumvented by approximating $L \circ H$ by a finite subset that is sufficiently dense. This leads to the concept of covering numbers.

Covering Numbers

Recall that a metric on a set S is a function $d: S \times S \to \mathbb{R}_{\geq 0}$ such that $d(x,y) = 0$ if and only if $x = y$, $d(x, y) = d(y, x)$, and $d(x, y) + d(y, z) \leq d(x, z)$. A pseudo-metric is defined like a metric, but replacing the first condition with the looser requirement that $d(x, x) = 0$ (that is, $d(x, y) = 0$ may also be possible if $x = y$).

Definition 11.2. Given a pseudo-metric space (S, d), an ε-net is a subset $T \subset S$ such that for every $x \in S$ there exists $y \in T$ with $d(x, y) \leq \varepsilon$. The covering number corresponding to S and ε is the smallest cardinality of an ε-net:

$$N(S, d, \varepsilon) = \inf\{|T| : T \text{ is an } \varepsilon\text{-net}\}.$$

Example 11.3. Consider the Euclidean unit ball $B_d = B_{d2} = \{x \in \mathbb{R}^d : \|x\|_2 \leq 1\}$. We construct an ε-net T as follows. Start with an arbitrary point $T = \{x_1\}$ and add points to T as follows: if $T = \{x_1, \ldots, x_k\}$, then choose a point x_{k+1} such that $\|x_{k+1} - x_j\| > \varepsilon$ for all j if possible, and add it to T, and if this is not possible, then stop. This process terminates since B_{d2} is bounded. The resulting set $T = \{x_1, \ldots, x_N\}$ is an ε-net by construction, and the distance between any two points in this set it larger than ε. Hence, the balls $B_d(x_i, \varepsilon/2)$ of radius $\varepsilon/2$ around the points in T are disjoint. Since the union of these balls is contained in the larger ball $(1 + \varepsilon/2)B_d$ (the scaling of the unit ball by a factor of $(1 + \varepsilon/2)$), we get the volume inequality

$$N \cdot \mathrm{vol}(B_d(x_1, \varepsilon/2)) = \mathrm{vol}\Big(\bigcup_{i=1}^{N} B_d(x_i, \varepsilon/2)\Big) \leq \mathrm{vol}((1 + \varepsilon/2)B_d). \tag{11.1}$$

The volume of a ball of radius r is $r^d \mathrm{vol}(B_d)$, hence

$$\mathrm{vol}(B_d(x_1, \varepsilon/2)) = (\varepsilon/2)^d \mathrm{vol}(B_d) \quad \text{and} \quad \mathrm{vol}((1 + \varepsilon/2)B_d) = (1 + \varepsilon/2)^d \mathrm{vol}(B_d),$$

and we get from (11.1) that

$$N \leq \frac{(1 + \varepsilon/2)^d}{(\varepsilon/2)^d} = \left(\frac{2}{\varepsilon} + 1\right)^d \leq \left(\frac{3}{\varepsilon}\right)^d$$

if $\varepsilon \leq 1$. Of course, if $\varepsilon \geq 1$ then we have an ε-net of size 1.

We next consider the set $L \circ H$ with the empirical 1-distance $d_{z1}(g_1, g_2) = \frac{1}{n}\sum_{i=1}^{n} |g_1(z_i) - g_2(z_i)|$.

We get the following bound for the empirical Rademacher complexity at z.

Theorem 11.4. Let H be a hypothesis set of functions taking values in $[-1, 1]$ and let L be a loss function taking values in $[0, 1]$. Then for any $z \in Z^n$ we have

$$\hat{R}_z(L \circ H) \leq \inf_{\varepsilon > 0} \left\{ \varepsilon \sqrt{\frac{2 \log(N(L \circ H, \varepsilon, d_z))}{n}} + \varepsilon \right\}.$$

The covering number takes over the role of VC dimension. Note that the covering number increases as ε decreases, and we get a trade-off between small ε and small covering number.

Proof of Theorem 11.4. Fix $z \in Z^n$ and let T be a minimal ε-net for $(L \circ H, d_z)$. It follows that for any $g \in L \circ H$ we can find a $g' \in T$ such that $d_z(g, g') \leq \varepsilon$. Hence,

$$\begin{aligned}
\hat{R}_z(L \circ H) &= \mathbb{E}_\sigma \left[\sup_{g \in L \circ H} \frac{1}{n} \sum_{i=1}^n \sigma_i g(z_i) \right] \\
&= \mathbb{E}_\sigma \left[\sup_{g \in L \circ H} \frac{1}{n} \sum_{i=1}^n \sigma_i (g(z_i) - g'(z_i) + g'(z_i)) \right] \\
&\leq \mathbb{E}_\sigma \left[\sup_{g \in L \circ H} \frac{1}{n} \sum_{i=1}^n \sigma_i (g(z_i) - g'(z_i)) \right] + \mathbb{E}_\sigma \left[\sup_{g \in L \circ H} \frac{1}{n} \sum_{i=1}^n \sigma_i g'(z_i) \right] \\
&\leq \mathbb{E}_\sigma \left[\sup_{g \in L \circ H} \frac{1}{n} |g(z_n) - g'(z_n)| \right] + \mathbb{E}_\sigma \left[\sup_{g' \in T} \frac{1}{n} \sum_{i=1}^n \sigma_i g'(z_i) \right] \\
&\leq \sup_{g \in L \circ H} d_{z1}(g, g') + \mathbb{E}_\sigma \max_{g' \in T} \frac{1}{n} \sum_{i=1}^n \sigma_i g'(z_i) \\
&\leq \varepsilon + \sqrt{\frac{2 \log(N(L \circ H, d_z, \varepsilon))}{n}},
\end{aligned}$$

where we used the bound for finite sets, Theorem 11.1. Since $\varepsilon > 0$ was arbitrary, this bounds holds for the infimum among $\varepsilon > 0$. □

The bound derived is for the set $L \circ H$. Under a boundedness condition on the loss function, we can bound the Rademacher complexity of this set in terms of that of H.

Theorem 11.5. If $L : Y \times Y \to [0, 1]$ then

$$\hat{R}_z(L \circ H) \leq 2 \hat{R}_x(H),$$

where $x = (x_1, \ldots, x_n)$ and $z = (z_1, \ldots, z_n)$ with $z_i = (x_i, y_i)$.

In light of this result, we conclude this section with a bound on $\hat{R}_z(H)$ for a specific class of functions. In what follows we denote by $B_p^d \subset \mathbb{R}^d$ the unit ball with respect to the p-norm. In particular,

$$\begin{aligned}
B_1^d &= \{x \in \mathbb{R}^d : \|x\|_1 = \sum_{i=1}^d |x_i| \leq 1\}, \\
B_\infty^d &= \{x \in \mathbb{R}^d : \|x\|_\infty = \max_i |x_i| \leq 1\}.
\end{aligned}$$

Notice that the set B_∞^d is a hypercube. We now consider $X = \{x \in \mathbb{R}^d : \|x\|_1 \leq 1\} = B_1^d$ and the class of functions

$$H = \{h : X \to Y : h(x) = \langle a, x \rangle, a \in B_\infty^d\}.$$

By the Hölder inequality, the functions h satisfy

$$|h(x)| = |\langle a, x \rangle| \leq \|a\|_\infty \|x\|_1 \leq 1.$$

Two functions $h, g \in H$ are thus represented by two vectors $\in \mathbb{R}^d$, and for their distance we have

$$d_{x1}(h, g) = \frac{1}{n} \sum_{i=1}^{n} |h(x_i) - g(x_i)| = \frac{1}{n} \sum_{i=1}^{n} |\langle a - b, x_i \rangle| .$$

From the Hölder inequality we get $|\langle a - b, x \rangle| \leq \|a - b\|_\infty \|x\|_1$, and since by assumption each $x_i \in B_1^d$, we have $\|x_i\|_1 \leq 1$ and

$$d_x(h, g) \leq \|a - b\|_\infty .$$

An ε-net therefore corresponds to a set $T = \{b_1, \ldots, b_N\} \subset B_\infty^d$ such that for every $a \in B_\infty$ there exists b_j such that $\|a - b_j\|_\infty \leq \varepsilon$. It follows from Exercise 4.1 that the covering number is bounded by

$$N(H, d_1, \varepsilon) \leq (3/\varepsilon)^d .$$

We thus get the bound

$$\hat{R}_x(H) \leq \inf_{\varepsilon > 0} \{\varepsilon + \sqrt{2d \log(3/\varepsilon)/n}\} .$$

If n is sufficiently large, then setting $\varepsilon = 3\sqrt{d \log(n)/n}$, we get the bound

$$\hat{R}_x(H) \leq 4\sqrt{d \log(n)/n} .$$

Note that the resulting bound does not depend on X. It is possible to remove the logarithmic factor $\log(n)$ using a more sophisticated technique called chaining.

Notes

12

Model Selection

Given an input space X, an output space Y, a class H of functions $h : X \to Y$, a loss function $L : Y \times Y \to \mathbb{R}_{\geq 0}$, and data $S = \{(x_i, y_i)\}_{i=1}^n$, we would like to solve the problem

$$\text{minimize}_{h} \hat{R}(h) = \frac{1}{n} \sum_{i=1}^n L(h(x_i), y_i) \quad \text{(A)}$$

subject to $h \in H$.

This is an example of a constrained optimization problem. The function to be minimized is the objective function and a solution $\hat{h}_S$ of (A) is called a minimizer. Several problems can arise when trying to solve this minimization problem.

1. The problem (A) may be hard to solve. This can be the case if the class H is large, the number of samples n is large, or when the objective function is not differentiable or not even continuous.

2. Do we even want to solve (A)? If the class H is large we may find a minimizer $\hat{h}_S$ that fits the data well but does not generalize.

Since a certain generalization error is unavoidable, we can often replace (A) with a surrogate that is computationally easier to handle and provides a solution that is close enough to the one we are looking for. The choice of such an approximation is also informed by the choice of H. We therefore first study the problem of finding a suitable class H, also known as model section.

Model Selection

Consider now the set of inputs again as a set of pairs of random variables $S = \{(X_i, Y_i)\}_{i=1}^n \subset X \times Y$, so that $\hat{h}_S$ is a random variable. Ideally we would like the generalization risk $R(\hat{h}_S)$ to be close to the Bayes risk R^*, which is the best possible generalization risk. Recall the decomposition

$$R(\hat{h}_S) - R^* = \underbrace{R(\hat{h}_S) - \inf_{h \in H} R(h)}_{\text{Estimation error}} + \underbrace{\inf_{h \in H} R(h) - R^*}_{\text{Approximation error}} \quad \text{(12.1)}$$

of the excess risk. Denote by $\bar{h}$ the minimizer of $R(h)$ in H. In previous lectures we saw that

$$R(\hat{h}_S) - R(\bar{h}) \leq 2 \sup_{h \in H} \left| R(h) - \hat{R}_S(h) \right|, \quad \text{(12.2)}$$

and therefore any bound that holds for the right-hand side with high probability (such as those based on VC dimension or covering numbers) also holds for the left-hand side. If $\mathcal{H}$ is large, then the bound may not be good enough, and in addition the minimizer $\hat{h}_S$ may be hard to compute. If, on the other hand, $\mathcal{H}$ is too small, then the approximation error can be large. One way to address this issue is to consider a nested family of sets $\mathcal{H}_k \subset \mathcal{H}_{k+1}$ of increasing complexity and to choose a k with optimal overall performance. We illustrate this using an example.

Example 12.1. Let $T = \{x \in \mathbb{R}^2 : \|x\| \le 1\}$ be a disk and let (X, Y) be a pair of random variables on $\mathbb{R}^2 \times \{0, 1\}$ such that $Y = 1_{\{X \in T\}}$. Consider the unit loss function, so that $R(h) = \mathbb{P}\{h(X) \neq Y\}$ for some function $h : \mathbb{R}^2 \to \{0, 1\}$. The function f_T is the Bayes classifier, as it satisfies $R(f_T) = R^* = 0$. For any hypothesis set $\mathcal{H}$ we can combine (12.1) and the bound (12.2) to get

$$R(\hat{h}_S) \le \underbrace{2 \sup_{h \in \mathcal{H}} |R(h) - \hat{R}(h)|}_{\text{Bound on estimation error}} + \underbrace{\inf_{h \in \mathcal{H}} R(h)}_{\text{Approximation error}}.$$

Let $\mathcal{H}_k$ denote the set of indicator functions of regions bounded by s convex polygons with at most k sides (see Figure 25.1). To bound the estimation error, we use the fact that the VC dimension of the class of convex k-gons is $2k + 1$ (see Problem Set 4), and get the bound

$$\sup_{h \in \mathcal{H}_k} |\hat{R}(h) - R(h)| \le \sqrt{\frac{(4k + 2) \log(n)}{n}} + \sqrt{\frac{\log(2/\delta)}{2n}} \tag{12.3}$$

with probability $1 - \delta$ (we simplified the logarithmic term in the first part of the bound). For the approximation error, we look at the well-known problem of approximating the circle with a regular polygon. The area enclosed by a regular k-gon inscribed in a circle of radius r is $r^2(k/2)\sin(2\pi/k)$, so the area of the complement in the disk is

$$\pi r^2 - r^2(k/2)\sin(2\pi/k) = O(k^{-2}), \tag{12.4}$$

where the equality follows from the Taylor expansion of the sine function. If the underlying probability distribution on $\mathbb{R}^2$ is the uniform distribution on a larger set or can be approximated as such, then (12.4) gives an upper bound for the approximation error (it can be less), and combined with (12.3) illustrates the estimation-approximation trade-off. The larger the number of sides of the polygons, the smaller the approximation error becomes, but the estimation error can become large due to overfitting. Thus even if the unknown shape we want to learn is a circle, if the number of samples is small we may be better off restricting to simpler models! This also has the additional advantage of saving computational cost.

There are general strategies for optimizing for k. One theoretical method is known as structural risk minimization (SRM). In this model, the parameter k enters into the optimization problem to be solved. An alternative, practical approach is cross-validation. In this approach, the training set S is subdivided into a smaller training set and a validation set. In a nutshell, the ERM problem is solved for different parameters k on the training set, and the one that performs best on the validation set is chosen.

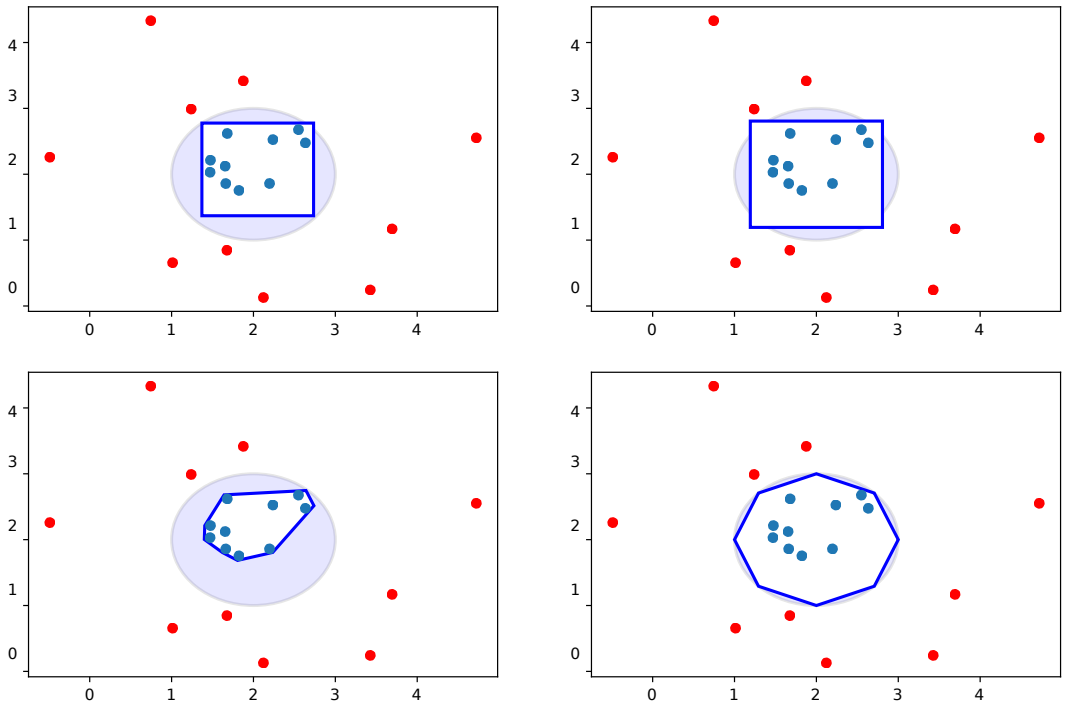


Figure 12.1: Learning a circle with polygons. The left panel shows the estimation error when trying to learn the shape using polygons with at most 4 and with at most 8 sides from the data. This is the error typically incurred by empirical risk minimization. The right panel illustrates the approximation error. This error measures how good we can approximate the ground truth with our function class.

Part II

Optimization

13

Optimization

> "[N]othing at all takes place in the universe in which some rule of maximum or minimum does not appear."
>
> — Leonhard Euler

Mathematical optimization, traditionally also known as mathematical programming, is the theory of optimal decision making. Other than in machine learning, optimization problems arise in a large variety of contexts, including scheduling and logistics problems, finance, optimal control and signal processing. The underlying mathematical problem always amounts to finding parameters that minimize (cost) or maximize (utility) an objective function in the presence or absence of a set of constraints. In the context of machine learning, the objective function is usually related to the empirical risk, but we first take a step back and consider optimization problems in greater generality.

What is an optimization problem?

A general mathematical optimization problem is a problem of the form

$$\begin{aligned} &\text{minimize} && f(x) \\ &\text{subject to} && x \in \Omega \end{aligned} \tag{13.1}$$

where $f : \mathbb{R}^d \to \mathbb{R}$ is a real-valued objective function and $\Omega \subseteq \mathbb{R}^d$ is a set defining the constraints. If $\Omega = \mathbb{R}^d$, then the problem is an unconstrained optimization problem. Among all $x \in \Omega$, we seek one with smallest f-value. Typically, the constraint set Ω will consist of such $x \in \mathbb{R}^d$ that satisfy certain equations and inequalities, $f_1(x) \leq 0, \ldots, f_m(x) \leq 0, g_1(x) = 0, \ldots, g_p(x) = 0$.

A vector x^* satisfying the constraints is called an optimum, a solution, or a minimizer of the problem (13.1), if $f(x^*) \leq f(x)$ for all other x that satisfy the constraints. Note that replacing f by $-f$, we could equivalently state the problem as a maximization problem.

Optimality conditions

In what follows we will study the unconstrained problem

$$\text{minimize } f(x), \tag{13.2}$$

where $x \in \mathbb{R}^d$.

Optimality conditions aim to identify properties that potential minimizers need to satisfy in relation to f(x). We will review the well known local optimality conditions for differentiable functions from calculus. We first identify different types of minimizers.

Definition 13.1. A point $x^* \in \mathbb{R}^d$ is a

- global minimizer of (13.2) if for all $x \in \mathbb{R}^d$, $f(x^*) \le f(x)$;
- a local minimizer, if there is an open neighbourhood U of x^* such that $f(x^*) \le f(x)$ for all $x \in U$;
- a strict local minimizer, if there is an open neighbourhood U of x^* such that $f(x^*) < f(x)$ for all $x \in U$;
- an isolated minimizer if there is an open neighbourhood U of x^* such that x^* is the only local minimizer in U.

Without any further assumptions on f, finding a minimizer is a hopeless task: we simply can not examine the function at all points in $\mathbb{R}^d$. The situation becomes more tractable if we assume some smoothness conditions. Recall that $C^k(U)$ denotes the set of functions that are k times continuously differentiable on some set U. The following first-order necessary condition for optimality is well known. We write $\nabla f(x)$ for the gradient of f at x, i.e., the vector

$$\nabla f(x) = \left(\frac{\partial f}{\partial x_1}(x), \dots, \frac{\partial f}{\partial x_d}(x)\right)^\top$$

Theorem 13.2. Let x^* be a local minimizer of f and assume that $f \in C^1(U)$ for a neighbourhood U of x^*. Then $\nabla f(x^*) = 0$.

There are simple examples that show that this is not a sufficient condition: maxima and saddle points will also have a vanishing gradient. If we have access to second-order information, in form of the second derivative, or Hessian, of f, then we can say more. Recall that the Hessian of f at x, $\nabla^2 f(x)$, is the $d \times d$ symmetric matrix given by the second derivatives,

$$\nabla^2 f(x) = \left(\frac{\partial^2 f}{\partial x_i \partial x_j}\right)_{1 \le i,j \le d}.$$

In the one-variable case we have learned that if x^* is a local minimizer of $f \in C^2([a,b])$, then $f'(x^*) = 0$ and $f''(x^*) \ge 0$. Moreover, the conditions $f'(x^*) = 0$ and $f''(x^*) > 0$ guarantee that we have a local minimizer. These conditions generalise to higher dimension, but first we need to know what $f''(x) > 0$ when we have more than one variable.

Recall also that a matrix A is positive semidefinite, written $A \succeq 0$, if for every $x \in \mathbb{R}^d$, $x^\top A x \ge 0$, and positive definite, written $A \succ 0$, if $x^\top A x > 0$. The property that the Hessian matrix is positive semidefinite is a multivariate generalization of the property that the second derivative is nonnegative. The known conditions for a minimizer involving the second derivative generalize accordingly.

Theorem 13.3. Let $f \in C^2(U)$ for some open set U and $x^* \in U$. If x^* is a local minimizer, then $\nabla f(x^*) = 0$ and $\nabla^2 f(x^*)$ is positive semidefinite. Conversely, if $\nabla f(x^*) = 0$ and $\nabla^2 f(x^*)$ is positive definite, then x^* is a strict local minimizer.

Unfortunately, the above criteria are not able to identify global minimizers, as differentiability is a local property. For convex functions, however, local optimality implies global optimality.

Examples

We present two examples of optimization problems that can be interpretred as machine learning problems, but have mainly been studied outside of the context of machine learning. The examples below come with associated Python code and it is not expected that you understand them in detail; they are merely intended to illustrate some of the problems that optimization deals with, and how they can be solved.

Example 13.4. Suppose we want to understand the relationship of a quantity y (for example, sales data) to a series of predictors x1, . . . , xp (for example, advertising budget in different media). We can often assume the relationship to be approximately linear, with distribution

$$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p + \varepsilon,$$

with noise ε that satisfies $\mathbb{E}[\varepsilon] = 0$. As function class we take

$$H = \{h : h(X) = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p, \beta = (\beta_0, \ldots, \beta_p) \in \mathbb{R}^{p+1}\}. \quad (13.3)$$

The goal is to determine the model parameters $\beta_0, \ldots, \beta_p$ from data.

To determine these, we can collect $n \geq p$ sample realizations (from observations or experiments),

$$\{(y_i, x_{i1}, \ldots, x_{ip}), 1 \leq i \leq n\}$$

and minimize the empirical risk with respect to the (normalized) L^2 loss function:

$$\hat{R}(h) = \frac{1}{2n} \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}))^2.$$

Collecting the data in matrices and vectors,

$$y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \quad X = \begin{pmatrix} 1 & x_{11} & \ldots & x_{1p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \ldots & x_{np} \end{pmatrix}, \quad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix},$$

we can write the empirical risk concisely as

$$\hat{R}(h) = \frac{1}{2n} \|y - X\beta\|_2^2.$$

Minimizing over $h \in H$ means minimizing over vectors $\beta \in \mathbb{R}^{p+1}$, and the best β is then the vector that solves the unconstrained optimization problem

$$\text{minimize} \frac{1}{2n} \|X\beta - y\|_2^2.$$

This is an example of an optimization problem with variables β, no constraints (all β are valid candidates and the constraint set is $\mathbb{R}^{p+1}$), and a quadratic objective function

$$f(\beta) = \frac{1}{2n} \|X\beta - y\|_2^2 = \frac{1}{2n} (X\beta - y)^\top (X\beta - y) = \frac{1}{2n} \beta^\top X^\top X \beta - 2y^\top X\beta + y^\top y, \quad (13.4)$$

where X> is the matrix transpose. Quadratic functions of the form (13.4) are convex, so this is a convex optimization problem. If the columns of X are linearly independent (which, by the way, requires there to be more data than the number of parameters p), this simple optimization problem has a unique closed form solution,

$$\beta* = (X>X)^{-1}X>y. \quad (13.5)$$

In practice one would not compute $\beta*$ by evaluating (13.5). There are more efficient methods available, such as gradient descent, the conjugate gradient method, and several variations of these. It is important to note that even in this simple example, solving the optimization problem can be problematic if the number of samples is large.

To illustrate the least squares setting using a concrete example, assume that we have data relating the basal metabolic rate (energy expenditure per time unit) in mammals to their mass.[1] The model we

use is $Y = \beta 0 + \beta 1X$, with Y the basal metabolic rate and X the mass. Using data for 573 mammals from the PanTHERIA database[2], we can assemble the vector y and the matrix X $\in Rn\times(p+1)$ in order to compute the $\beta=(\beta 0,\beta >1)$ .Here,p=1andn=573.Weillustratehowtosolvethisprobleminpython.

As usual, we first have to import some relevant libraries: numpy for numerical computation, pandas for loading and transforming datasets, cvxpy for convex optimization, and matplotlib for plotting.

In [1]:
```
# Import some important Python
modules import numpy as np
import pandas as pd
from cvxpy import *
import matplotlib.pyplot as plt
```

We next have to load the data. The data is saved in a table with 573 rows and 2 columns, where the first column list the mass and the second the basal metabolic rate.

In [2]:
```
# Load data into numpy array
bmr = pd.read_csv('../../data/bmr.csv',header=None).as_matrix() #
We can find out the dimension of the data
bmr.shape
```

Out [2]: (573, 2)

To see the first three and the last three rows of the dataset, we can use the "print" command.

[1] This example is from the episode “Size Matters” of the BBC series Wonders of Life.

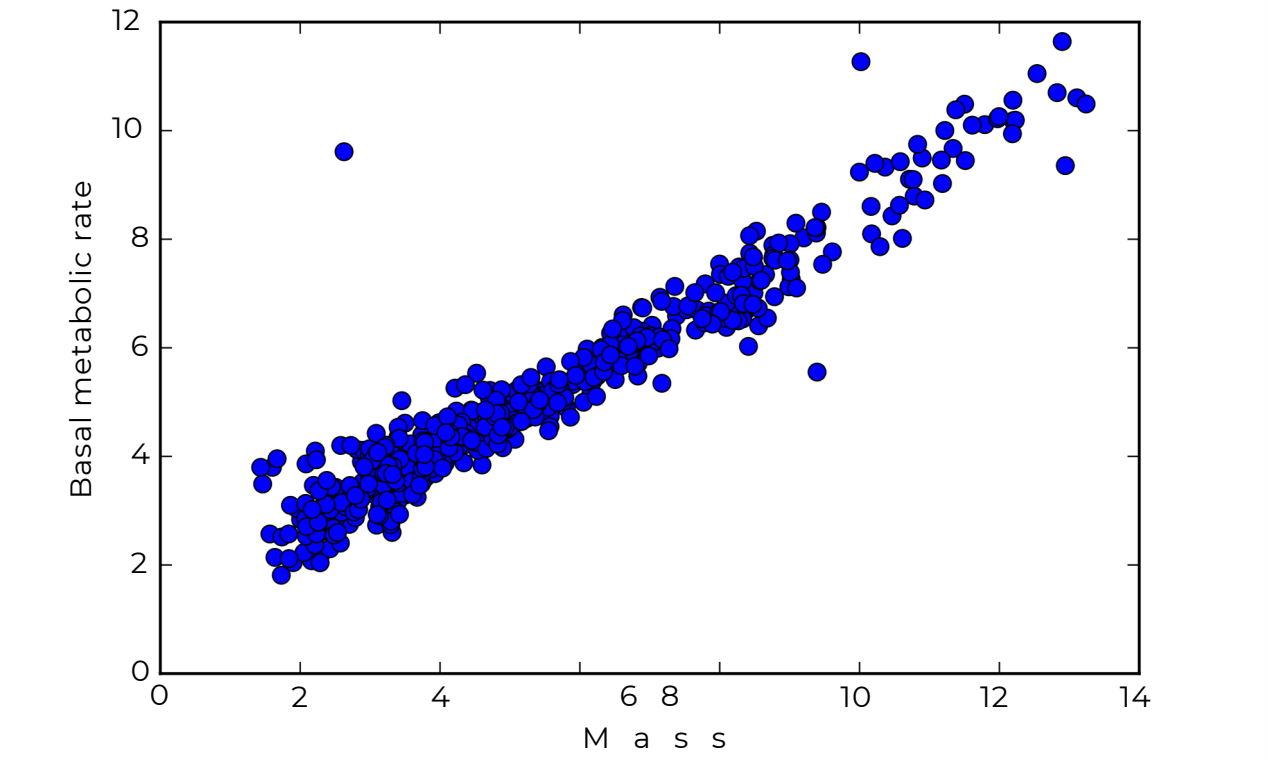
[2] http://esapubs.org/archive/ecol/E090/184/#data

In [3]:
```
print (bmr[0:3,:])
```

```
[[ 13.108      10.604 ]
 [ 9.3918      8.2158]
 [ 10.366      9.3285]]
```

To visualise the whole dataset, we can make a scatterplot by interpreting each row as a coordinate on the plane, and marking it with a dot.

In [4]:
```
# Display scatterplot of data (plot all the rows as points) bmr1
= plt.plot(bmr[:,0],bmr[:,1],'o')
plt.xlabel("Mass")
plt.ylabel("Basal metabolic rate")
plt.show()
```



The plot above suggests that the relation of the basal metabolic rate to the mass is linear, i.e., of the form

$$Y = \beta_0 + \beta 1X,$$

where X is the mass and Y the BMR. We can find $\beta\beta$ and 1bysolvinganoptimizationproblemas described above. We first have to assemble the matrix X and the vector y.

In [5]:
```
n = bmr.shape[0]
p = 1
X = np.concatenate((np.ones((n,1)),bmr[:,0:p]),axis=1) y =
bmr[:,-1]
```

In [6]:

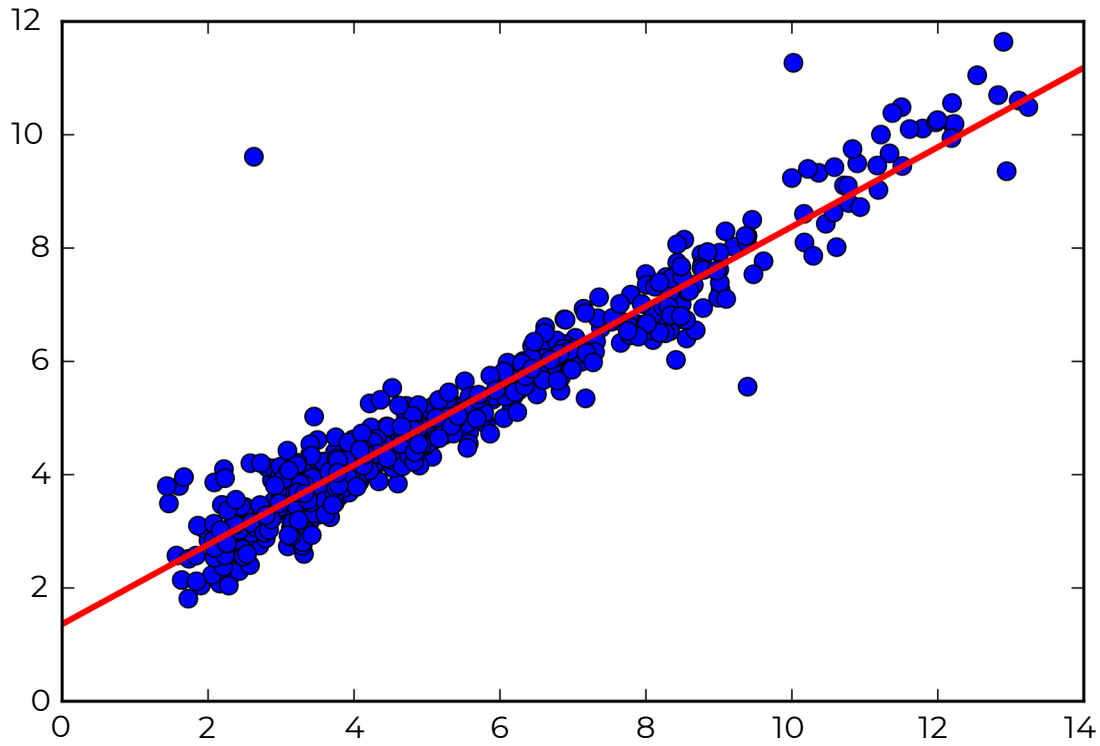
```
# Create a (p+1) vector of variables
Beta = Variable(p+1)

# Create sum-of-squares objective function
objective = Minimize(sum_entries(square(X*Beta - y)))
# Create problem and solve it
prob = Problem(objective)
prob.solve()

print("status: ", prob.status)
print("optimal value: ", prob.value)
print("optimal variables: ", Beta[0].value, Beta[1].value)
```

```
status: optimal
optimal value: 152.736200529558
optimal variables: 1.3620698558275837 0.7016170245505547
```

Now that we solved the problem and have the values $\beta 0 = 1.362$ and $\beta 1 = 0.702$ we can plot the line and see how it fits the data.

In [6]:

```
plt.plot(bmr[:,0],bmr[:,1],'o')

xx = np.linspace(0,14,100)
bmr = plt.plot(xx, Beta[0].value+Beta[1].value*xx, color='red',\
linewidth=2)
plt.show()
```



Even though for illustration purposes we used the CVXPY package, this particular problem can be solved directly using the least squares solver in numpy.

In [7]:
```
import numpy.linalg as la
beta = la.lstsq(X,y)
print(beta[0])
```

```
[ 1.36206997  0.70161692]
```

Example 13.5. (Image inpainting) Even problems in image processing that do not appear to be machine learning problems can be cast as such. An image can be viewed as an m × n matrix U , with each entry uij corresponding to a light intensity (for greyscale images), or a colour vector, represented by a triple of red, green and blue intensities (usually with values between 0 and 255 each). For simplicity the following discussion assumes a greyscale image. For computational purposes, the matrix of an image is often viewed as an mn-dimensional vector u, with the columns of the matrix stacked on top of each other.

In the image inpainting problem, one aims to learn the true value of missing or corrupted entries of an image. There are different approaches to this problem. A conceptually simple approach is to replace the image with the closest image among a set of images satisfying typical properties. But what are typical properties of a typical image? Some properties that come to mind are:

• Images tend to have large homogeneous areas in which the colour doesn't change much;

• Images have approximately low rank, when interpreted as matrices.

Total variation image analysis takes advantage of the first property. The total variation or TV-norm is the sum of the norm of the horizontal and vertical differences,

$$\|U\|_{TV} = \sum_{i=1}^{m}\sum_{j=1}^{n} \sqrt{(u_{i+1,j} - u_{i,j})^2 + (u_{i,j+1} - u_{i,j})^2},$$

where we set entries with out-of-bounds indices to 0. The TV-norm naturally increases with increased variation or sharp edges in an image. Consider for example the two following matrices (imagine that they represent a 3 × 3 pixel block taken from an image).

$$U_1 = \begin{bmatrix} 0 & 1 & 7 & 3 \\ 7 & 3 & 2 & 0 \\ 2 & 9 & 2 & 7 \end{bmatrix}, \quad U_2 = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix}$$

The left matrix has TV-norm $\|U_1\|_{TV} = 200$, while the right one has TV-norm $\|U_2\|_{TV} = 14.721$ (verify this!) Intuitively, we would expect a natural image with artifacts added to it to have a higher TV norm.

Now let U be an image with entries uij, and let $\Omega \subset [m] \times [n] = \{(i,j) \mid 1 \le i \le m, 1 \le j \le n\}$ be the set of indices where the original image and the corrupted image coincide (all the other entries are missing). One could attempt to find the image with the smallest TV-norm that coincides with the known pixels uij for $(i, j) \in \Omega$. This is an optimization problem of the form

$$\text{minimize} \quad \|X\|_{TV} \quad \text{subject to} \quad x_{ij} = u_{ij} \text{ for } (i,j) \in \Omega.$$

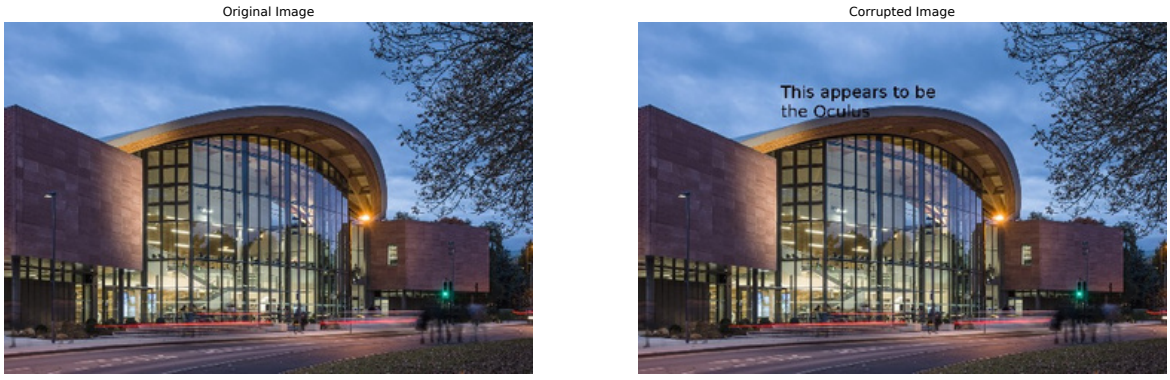
The TV-norm is an example of a convex function and the constraints are linear conditions which define a convex set. This is again an example of a convex optimization problem and can be solved efficiently by a range of algorithms. For the time being we will not go into the algorithms but solve it using CVXPY.

The example below is based on an example from the CVXPY Tutorial[3], and it is recommended to look at this tutorial for other interesting examples!

Warning: the example below uses some more advanced Python programming, it is not necessary to understand.

In our first piece of code below, we load the image and a version of the image with text written on it, and display the images. The Python Image Library (PIL) is used for this purpose.

In [9]:
```
from PIL import Image

# Load the images and convert to numpy arrays for processing.
U = np.array(Image.open("../images/oculus.png"))
Ucorr = np.array(Image.open("../images/oculus-corr.png"))

# Display the images
fig, ax = plt.subplots(1, 2,figsize=(10, 5))

ax[0].imshow(U);
ax[0].set_title("Original Image")
ax[0].axis('off')

ax[1].imshow(Ucorr);
ax[1].set_title("Corrupted Image")
ax[1].axis('off');
```



After having the images at our disposal, we determine which entries of the corrupted image are known. We store these in a mask M, with entries mijk = 1 if the colour k of the (i, j)-th pixel is known, and 0 otherwise.

In [10]:
```
# Each image is now an m x n x 3 array, with each pixel
# represented by three numbers between 0 and 255,
# corresponding to red, green and blue
rows, cols, colours = U.shape

# Create a mask: this is a matrix with a 1 if the corresponding #
pixel is known, and zero else
M = np.zeros((rows, cols, colours))
for i in range(rows):
for j in range(cols):
for k in range(colours):
if U[i, j, k] == Ucorr[i, j, k]:
M[i, j, k] = 1
```

[3] http://www.cvxpy.org/en/latest/tutorial/index.html

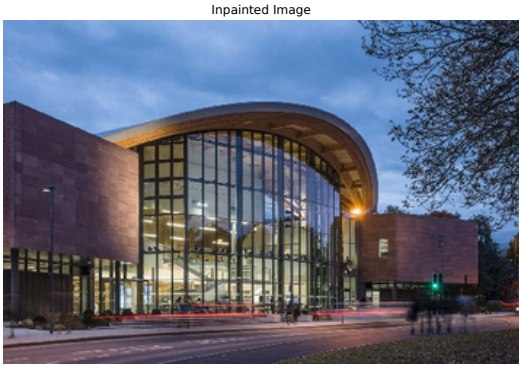
We are now ready to solve the optimization problem using CVXPY. As the problem is rather big (more than a million variables), it is important to choose a good solver that will solve the problem to sufficient accuracy in an acceptable amount of time. For the example at hand, we choose the SCS solver, which can be specified when calling the solve function.

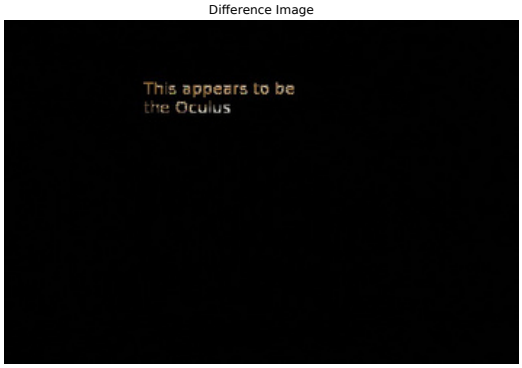
In [11]:
```
# Determine the variables and constraints
variables = []
constraints = []
for k in range(colours):
X = Variable(rows, cols)
# Add variables
variables.append(X)
        # Add constraints by multiplying the relevant variable matrix
# elementwise with the mask
constraints.append(mul_elemwise(M[:, :, k], X) ==
\ (M[:, :, k], Ucorr[:, :, k]))

# Create a problem instance with
objective = Minimize(tv(variables[0],variables[1],variables[2]))

# Create a problem instance and solve it using the SCS solver
prob = Problem(objective, constraints)
prob.solve(verbose=True, solver=SCS)
```

Out [11]: 8263910.812250629

Now that we solved the optimization problem, we have a solution stored in 'variables'. We have to transform this back into an image and display the result.

In [12]:
```
# Load variable values into a single array.
Urec = np.zeros((rows, cols, colours), dtype=np.uint8)
for i in range(colours):
Urec[:, :, i] = variables[i].value

fig, ax = plt.subplots(1, 2,figsize=(10, 5))

# Display the inpainted image.
ax[0].imshow(Urec);
ax[0].set_title("Inpainted Image")
ax[0].axis('off')

ax[1].imshow(np.abs(Ucorr[:,:,0:3] - Urec));
ax[1].set_title("Difference Image")
ax[1].axis('off');
```





Another typical structure of images is that the singular values of the image, considered as matrix, decay quickly. The singular value decomposition (SVD) of a matrix $A \in \mathbb{R}^{m \times n}$ is the matrix product

$$A = U \Sigma V^T,$$

where $U \in \mathbb{R}^{m \times m}$ and $V \in \mathbb{R}^{n \times n}$ are orthogonal matrices, and $\Sigma \in \mathbb{R}^{m \times n}$ is a diagonal matrix with entries $\sigma_1, \ldots, \sigma_{\min\{m,n\}}$ on the diagonal. Instead of minimizing the TV-norm of an image X, one may instead try to minimize the Schatten 1-norm, defined as the sum of the singular values, $\|U\|_{S1} = \sigma_1 + \cdots + \sigma_{\min\{m,n\}}$. The problem is then

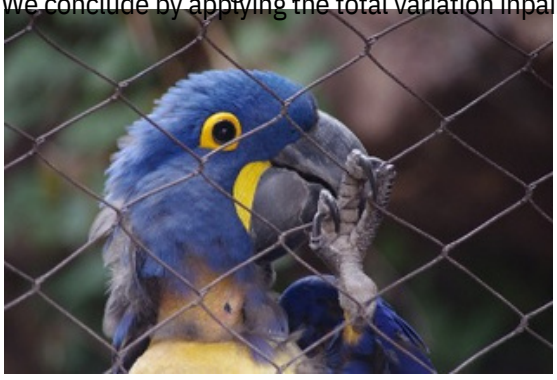
$$\text{minimize} \quad \|X\|_{S1} \quad \text{subject to} \quad x_{ij} = u_{ij} \quad \text{for} \quad (i,j) \in \Omega.$$

This is an instance of a type of convex optimization problem known as semidefinite programming.

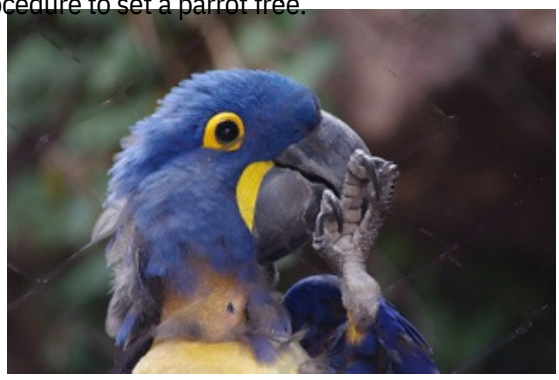
Alternatively, one may also use the 1-norm of the image applied to a discrete cosine transform (DCT) or a discrete wavelet transform (DWT). As this examples (and many more to come) shows: there is no unique choice of loss function, and hence of the objective function, for a particular problem. These choices depend on model assumptions and require some knowledge of the problem one is trying to solve.

We conclude by applying the total variation inpainting procedure to set a parrot free.

Caged parrot



Free parrot



Notes

14

Convexity

Convexity is a central theme in optimization. The reason is that convex optimization problems have many favourable properties, such as lower computational complexity and the property that local minima are also global minima. Despite being a seemingly special property, convex optimization problems arise surprisingly often.

Convex functions

Definition 14.1. A convex body

A set $C \subseteq \mathbb{R}^d$ is convex if for all $x, y \in C$ and $\lambda \in [0, 1]$, the line $\lambda x + (1 - \lambda) y \in C$. is a convex set that is closed and bounded.

Definition 14.2

Let $S \subseteq \mathbb{R}^d$. A function $f : S \to \mathbb{R}$ is convex if S is convex and for all $x, y \in S$ and $\lambda \in [0, 1]$,

$$f(\lambda x + (1 - \lambda) y) \leq \lambda f(x) + (1 - \lambda) f(y).$$

The function f is called strictly convex if

$$f(\lambda x + (1 - \lambda) y) < \lambda f(x) + (1 - \lambda) f(y).$$

A function f is called concave, if
−f is convex.

Figure 14.1 illustrates what a convex function of one variable looks like. The graph of the function lies below any line connecting two points on it. A function f has a domain dom(f), which is where the function is defined. For example, if f(x) = log(x), then dom(f) = R+, the positive integers. The definition of a convex function thus states that the domain of f is a convex set S. We can also restrict a function on a smaller domain, even though the function could be defined more generally. For example, f(x) = x3 is a convex function if restricted to the domain R
≥0, but is not convex on R.

A convex optimization problem is an optimization problem in which the set of constraints Ω and the function f are convex. While most general optimization problems are practically intractable, convex optimization problems can be solved efficiently, and still cover a surprisingly large range of applications!

Convex function have pleasant properties, while at the same time covering many of the functions that arise in applications. Perhaps the most important property is that local minima are global minima.

Theorem 14.3. Let $f : \mathbb{R}^d \to \mathbb{R}$ be a convex function. Then any local minimizer x^* is a global minimizer.

Proof. Let x^* be a local minimizer and assume that it is not a global minimizer. Then there exists a vector $y \in \mathbb{R}^d$ such that $f(y) < f(x^*)$. Since f is convex, for any $\lambda \in [0, 1]$ we have $x = \lambda y + (1 - \lambda) x^*$

$$f(x) \leq \lambda f(y) + (1 - \lambda) f(x^*) < \lambda f(x^*) + (1 - \lambda) f(x^*) = f(x^*).$$

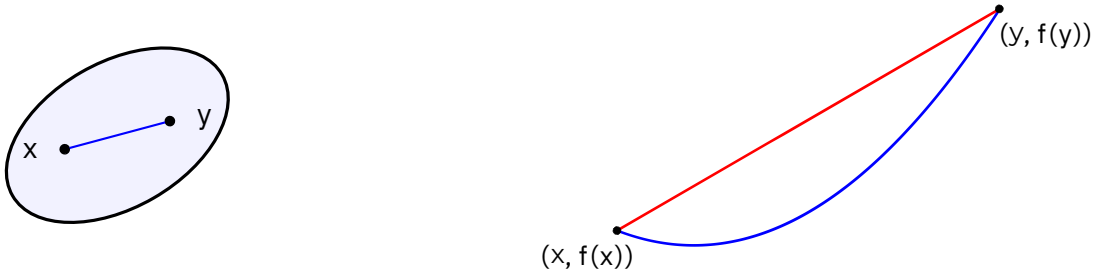


Figure 14.1: A convex set and a convex function

This holds for all x on the line segment connecting y and $x*$. Since every open neighbourhood U of $x*$ contains a bit of this line segment, this means that every open neighbourhood U of $x*$ contains an $x \neq x*$ such that f(x) $\leq$ f($x*$), in contradiction to the assumption that $x*$ is a local minimizer. It follows that $x*$ has to be a global minimizer. □

Remark 14.4. Note that in the above theorem we made no assumptions about the differentiability of the function f! In fact, while a convex function is always continuous, it need not be differentiable. The function f(x) = |x| is a typical example: it is convex, but not differentiable at x = 0.

Example 14.5. Affine functions f(x) = $\langle x, a \rangle$ + b and the exponential function ex are examples of convex functions.

Example 14.6. In optimization we will often work with functions of matrices, where an m × n matrix is considered as a vector in $R^{m \times n} = R^{mn}$. If the matrix is symmetric, that is, if $A^{\top} = A$, then we only care about the upper diagonal entries, and we consider the space S^n of symmetric matrices as a vector space of dimension d = n(n + 1)/2 (the number of entries on and above the main diagonal). Important functions on symmetric matrices that are convex are the operator norm $\|A\|_2 = \max_{x \neq 0} \frac{\|Ax\|_2}{\|x\|_2}$, $\|A\|_2$, defined as $\|A\|_2 = \max_{\|x\|=1} \|Ax\|_2$, or the function log det(X), defined on the set of positive semidefinite symmetric matrices S_+^d.

There are useful ways of characterising convexity using differentiability.

Theorem 14.7. 1. Let $f \in C^1(R^d)$. Then f is convex if and only if for all $x, y \in R^d$,

$$f(y) \geq f(x) + \nabla f(x)^{\top}(y - x).$$

2. Let $f \in C^2(R^d)$. Then f is convex if and only if $\nabla^2 f(x)$ is positive semidefinite for all x. If $\nabla^2 f(x)$ is positive definite for all x, then f is strictly convex.

Example 14.8. Consider a quadratic function of the form

$$f(x) = \frac{1}{2} x^{\top} A x + b^{\top} x + c,$$

where $A \in R^{n \times n}$ is symmetric. Writing out the product, we get

)□□□a11 ⋯ a1..□.□.n..n

xTAx (= x 1 ⋯ x n

□□□□□□ x 1 . . □ . □ x n

annn

) □ a 1 ⋯ □□ a 1 1 x 1 + ⋯ . . . n 1 x 1 + ⋯

□+a1nxn

(= x 1 ⋯ x n

□□+annxn

a

∑n∑n=ijjai=xjx=1

∑i+21≤i<j≤n

aij = aji

Because A is symmetric, we have , and the above product simplifies to

f

xTAx ∑n= a x2iii=1 ∂2f aij xi xj .

bi,

∂xi∂xj

xi

This is a quadratic function, because it involves products of the . The gradient and the Hessian of f (x) are found by computing the partial derivatives of :

∑n=aijxj+ij=1

∂f

∂ ∇f(x) = Ax + , ——— = aij.

x f

In summary, we have

b ∇2f (x)= A.

Using the previous theorem, we see that is convex if and only if A is positive semidefinite. A typical example for such a function is

– – – f (x)= ||Ax − b ||2 = (Ax b)T(Ax b)= xT AT Ax 2bT Ax+ bTb.

ATA Thematrix isalwayssymmetricandpositivesemidefinite(why?)sothatthefunctionfis convex.

Aconvenientwaytovisualiseafunction isthroughcontourplots.Alevelsetofthe

f R<→ R

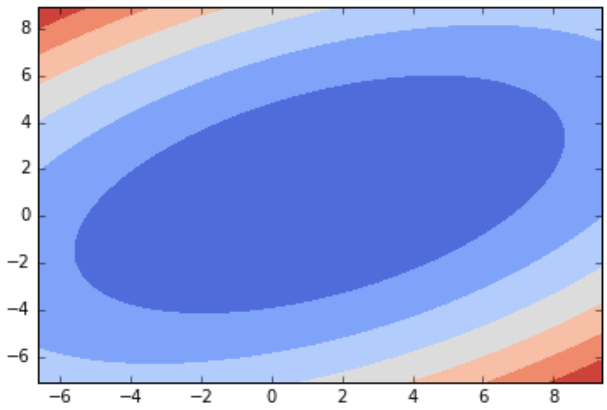
function f is a set of the form

x |f (x)=c

{ },

where c is the level. Each such level set is a curve in $\mathbb{R}^2$, and a contour plot is a plot of a collection of such curves for various c. If one colours the areas between adjacent curves, one gets a plot as in the following figure. A convex function is has the property that there is only one sink in the contour plot.

Notes


8
6
4
2
0
−2
−4
−6
−6
−4
−2
0
2
4
6
8

15

Lagrangian Duality

In this lecture we study optimality conditions for convex problems of the form

$$\begin{aligned} \underset{x}{\text{minimize}} \quad & f(x) \\ \text{subject to} \quad & f_i(x) \leq 0 \\ & h_j(x) = 0, \end{aligned} \tag{1}$$

where $x \in \mathbb{R}^n$, $f = (f_1, \ldots, f_m)$, $h = (h_1, \ldots, h_p)$, and the inequalities are componentwise. We assume that f and the f_i are convex, and the h_j are linear. It is also customary to write the conditions $h(x) = 0$ as $Ax = b$, with $h_j(x) = a_j^T x - b_j$, a_j being the j-th row of A. The feasible set of (1) is the set

$$F = \{x \mid f_i(x) \leq 0, Ax = b\}$$

It is easy to see that F is convex if the f_i are convex. If the objective f and the f_i are also linear, then (1) is called a linear programming problem, and F is a polyhedron. Such problems have been studied extensively and can be solved with efficient algorithms such as the simplex method.

A first-order optimality condition

We first generalize the standard first-order optimality conditions for differentiable functions to the setting of constrained convex optimization.

Theorem 15.1. Let $f \in C^1(\mathbb{R}^n)$ be a convex, differentiable function, and consider a convex optimization problem of the form (1). Then x^* is an optimal point of the optimization problem

$$\text{minimize} \quad f(x) \quad \text{subject to} \quad x \in F$$

if and only if for all $y \in F$,

$$\langle \nabla_x f(x^*), y - x^* \rangle \geq 0, \tag{15.1}$$

where F is the feasible set of the problem.

Proof. Suppose x^* is such that (1) holds. Then, since f is a convex function, for all $y \in F$ we have,

$$f(y) \geq f(x^*) + \langle \nabla_x f(x^*), y - x^* \rangle \geq f(x^*),$$

which shows that x^* is a minimizer in F. To show the opposite direction, assume that but x^* is a minimizer that (15.1) does not hold. This means that there exists a $y \in F$ such that $\langle \nabla f(x^*), y - x^* \rangle < 0$. Since both x^* and y are in F and is convex, any point $x^* + \lambda y$ with $z(\lambda) = (1 - \lambda)x^* + \lambda y$, $\lambda \in [0, 1]$ is also in F. At $\lambda = 0$ we have $\frac{df}{d\lambda} f(z(\lambda))|_{\lambda=0} = \langle \nabla f(x^*), y - x^* \rangle < 0$. Since the derivative at $\lambda = 0$ is negative, the function is decreasing at $\lambda = 0$, and therefore, for small $\lambda > 0$, $f(z(\lambda)) < f(x^*)$, in contradiction with the assumption that x^* is a minimizer. □

Example 15.2. In the absence of constraints, $F = R^n$, and the statement says that

$$\forall y \in R^n : \langle \nabla f(x^*), y - x^* \rangle \geq 0.$$

Given y such that $\langle \nabla f(x^*), y - x^* \rangle \geq 0$, then replacing y by $2x^* - y$ we also have the converse inequality, and therefore the optimality condition is equivalent to saying that $\nabla f(x^*) = 0$. We therefore recover the well-known first order optimality condition.

Geometrically, the first order optimality condition means that the set

$$\{x \mid \langle \nabla f(x^*), x \rangle \leq \langle \nabla f(x^*), x^* \rangle\}$$

defines a supporting hyperplane to the set F.

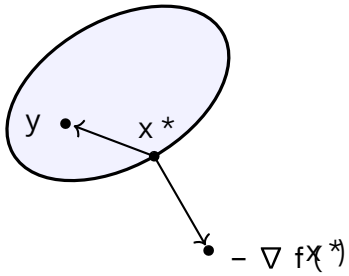


Figure 15.1: Optimality condition

Lagrangian duality

Recall the method of Lagrange multipliers. Given two functions $f(x, y)$ and $h(x, y)$, if the problem

$$\text{minimize } f(x,y) \text{ subject to } h(x, y) = 0$$

has a solution (x^*, y^*), then there exists a parameter λ, the Lagrange multiplier, such that

$$\nabla f(x^*, y^*) = \lambda \nabla h(x^*, y^*). \tag{15.2}$$

In other words, if we define the Lagrangian as

$$L(x, y, \lambda) = f(x, y) - \lambda h(x, y),$$

then (15.2) says that $\nabla L(x^*, y^*, \lambda) = 0$ for some λ. The intuition is as follows. The set

$$M = \{(x, y) \in R2 \mid h(x, y) = 0\}$$

is a curve in R2, and the gradient $\nabla h(x, y)$ is perpendicular to M at every point $(x, y) \in M$. For someone living inside M, a vector that is perpendicular to M is not visible, it is zero. Therefore the gradient $\nabla f(x, y)$ is zero as viewed from within M if it is perpendicular to M, or equivalently, a multiple of $\nabla h(x, y)$.

Alternatively, we can view the graph of f(x, y) in three dimensions. A maximum or minimum of f(x, y) along the curve defined by h(x, y) = 0 will be a point at which the direction of steepest ascent $\nabla f(x, y)$ is perpendicular to the curve h(x, y) = 0.

Example tion 2withtheconstraint 22

√15.3.Consˉderthefuncf(x,y)=xyh(x,y)=x+y−3(acircleofradius3).TheLagrangianisthefunction

$$L(x, y, \lambda) = x^2 y - \lambda(x^2 + y^2 - 3).$$

Computing the partial derivatives gives the three equations

$$\frac{\partial}{\partial} L = 2xy - 2\lambda x = 0$$

$$\frac{x}{\partial} L = x^2 - 2\lambda y = 0$$

$$\frac{\partial}{\partial} L = x + y^2 - 3 = 0.$$

$$y$$

From the second equation we get $\lambda = \frac{x^2}{2y}$ and the first and third equations become

$$\lambda \quad 2xy \quad \frac{x^3}{y} = 0$$

$$x^2 + y^2 \quad 3 = 0 .$$

√Solvingthissystem,wegetsixcriticalpoint, $(\sqrt{2}, -1)$ $(0, \pm\sqrt{3})$.Tofindoutwhichoneoftheseisthe minimizers, we just evaluate the function f on each of these.

We now turn to convex problems of the more general form form

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & h(x) \leq 0, \end{array} \tag{15.3}$$

Denote by D the domain of all the functions f, f_i, h_j, i.e.,

$$D = \text{dom}(f) \cap \text{dom}(f_1) \cap \ldots \cap \text{dom}(f_m) \cap \text{dom}(h_1) \cap \ldots \cap \text{dom}(h_p)$$

Assume that D is not empty and let p^* be the optimal value of (15.3).

The Lagrangian of the system is defined as

$$L(x, \lambda, \mu) = f(x) + \lambda^T f(x) + \mu^T h(x) = f(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{i=1}^{p} \mu_i h_i(x).$$

The vectors λand μ are called the dual variables or Lagrange multipliers of the system. The domain of L $D \times R^m \times R^p$ is .

Definition 15.4. The Lagrange dual of the problem (15.3) is the function

$$g(\lambda, \mu) = \inf_{x \in D} L(x, \lambda, \mu).$$

If the Lagrangian L is unbounded from below, then the value is $-\infty$.

The Lagrangian L is linear in the λ_i and μ_j variables. The infimum of a family of linear functions is concave, so that the Lagrange dual is a concave function. Therefore the negative $-g(\lambda, \mu)$ is a convex function.

Lemma 15.5. For any $\mu \in \mathbb{R}^p$ and $\lambda \geq 0$ we have

$$g(\lambda, \mu) \leq p^*.$$

Proof. Let x^* be a feasible point for (15.3), that is,

$$f_i(x^*) \leq 0, \quad h_j(x^*) = 0, \quad 1 \leq i \leq m, \quad 1 \leq j \leq p.$$

Then for $\lambda \geq 0$ and any μ, since each $h_j(x^*) = 0$ and $\lambda_i f_i(x^*) \leq 0$,

$$L(x^*, \lambda, \mu) = f(x^*) + \sum_{i=1}^{m} \lambda_i f_i(x^*) + \sum_{j=1}^{p} \mu_j h_j(x^*) \leq f(x^*).$$

In particular,

$$g(\lambda, \mu) = \inf_{x} L(x, \lambda, \mu) \leq L(x^*, \lambda, \mu) \leq f(x^*).$$

Since this holds for all feasible x^*, it holds in particular for the x^* that minimizes (15.3), for which $f(x^*) = p^*$. □

A point (λ, μ) with $\lambda \geq 0$ and $(\lambda, \mu) \in \operatorname{dom}(g)$ is called a feasible point of the dual problem. The Lagrange dual problem of the optimization problem (15.3) is the problem

$$\text{maximize} \quad g(\lambda, \mu) \quad \text{subject to} \quad \lambda \geq 0. \tag{15.4}$$

If q^* is the optimal value of (15.4), then $q^* \leq p^*$. In the special case of linear programming we actually have $q^* = p^*$. We will see that under certain conditions, we have $q^* = p^*$ for more general problems, but this is not always the case.

Notes

16

KKT Conditions

For convex problems of the form

$$\begin{aligned} &\text{minimize} && f(x) \\ &\text{subject to} && f_i(x) \leq 0 \\ &&& Ax = b, \end{aligned} \tag{1}$$

we introduced the Lagrangian $L(x, \lambda, \mu)$ and defined the Lagrange dual as

$$g(\lambda, \mu) = \inf_{x \in D} L(x, \lambda, \mu).$$

We saw that $g(\lambda, \mu)$ is a lower bound on the optimal value of (1). Note that here we wrote the conditions $h_j(x) = 0$ as system of linear equations $Ax = b$, since for the problem to be convex, we require that the h_j be linear functions. We will derive conditions under which the lower bound provided by the Lagrange dual matches the upper bound, and derive a system of equations and inequalities that certify optimality, the
Karush-Kuhn-Tucker (KKT) conditions. These conditions can be seen as generalizations of the first-order optimality conditions to the setting when equality and inequality constraints are present.

Constraint qualification

Consider a linear programming problem of the form

$$\begin{aligned} &\text{minimize} && \langle c, x \rangle \\ &\text{subject to} && Ax = b \\ &&& x \geq 0. \end{aligned}$$

The inequality constraints are $-x_i \leq 0$, while the equality constraints are $a_i^\top x - b_i$. The Lagrangian has the form

$$\begin{aligned} L(x, \lambda, \mu) &= \langle c, x \rangle - \sum_{i=1}^{n} \lambda_i x_i + \sum_{j=1}^{m} \mu_j (a_j^\top x - b_j) \\ &= (c - \lambda + A^\top \mu)^\top x - b^\top \mu. \end{aligned}$$

The infimum over x of this function is $-\infty$ unless $c - \lambda + A^\top \mu = 0$. The Lagrange dual is therefore

$$g(\lambda, \mu) = \begin{cases} -b^\top \mu & \text{if } c - \lambda + A^\top \mu = 0 \\ -\infty & \text{else.} \end{cases}$$

From Lemma 15.5 we conclude that

$$\max\{-b^\top \mu \mid c + A^\top \mu = 0, \lambda \geq 0\} \leq \min\{c^\top x \mid Ax = 0, x \geq 0\}.$$

Notethatifwewrite $= \lambda$ and $= \mu$, thenwegetthedualversionofthelinearprogrammingproblem we started out with, and in this case we know that

$$\max_{\lambda \geq 0, \mu} g(\lambda, \mu) = p^*.$$

In the example of linear programming, we have seen that the optimal value of the dual problem is equal to the optimal value of the primal problem. In general, we have

$$d^* = \sup_{\lambda \geq 0, \mu} g(\lambda, \mu) \leq \inf_{x \in D} \{f(x) \mid f_i(x) \leq 0, Ax = b\} = p^*.$$

Once certain conditions, called constraint qualifications, hold, we can ensure that strong duality holds, which means $d^* = p^*$. One particular such constraint qualification is Slater's Theorem.

Theorem 16.1. (Slater conditions) Assume that the interior of the domain D is non-empty, that the problem (1) isconvex,andthatthereexistsapoint $x \in D$ such that

$$f_i(x) < 0, \quad 1 \leq i \leq m, \quad Ax = b, \quad 1 \leq j \leq p.$$

Then $d^* = p^*$, the primal optimal value coincides with the dual optimal value.

Example 16.2. The problem minimize e^{-x} subject to $x^2/y \leq 0$, $y \geq 0$ is an example of a convex problem that does not satisfy strong duality.

Example 16.3. Consider the problem

$$\text{minimize}_x \; x^\top x \text{ subject to } Ax = b.$$

The Lagrangian is $L(x, \mu) = x^\top x + \mu^\top (Ax - b)$. For any μ, we can find the infimum by setting the derivative of the Lagrangian to zero:

$$g(\mu) = \inf_x L(x, \mu)$$

$$\nabla_x L(x, \mu) = 2x + A^\top \mu = 0,$$

which gives the solution

$$x = -\frac{1}{2} A^\top \mu.$$

The dual function is therefore

$$g(\mu) = -\frac{1}{4} \mu^\top A^\top A \mu - b^\top \mu.$$

As the negative of a positive semidefinite quadratic function, it is concave. Moreover, we get the lower bound

$$-\frac{1}{4} \mu^\top A^\top A \mu - b^\top \mu \leq \inf\{x^\top x \mid Ax = b\}.$$

The problem we started out with is convex, and if we assume that there exists a feasible primal point, then the above inequality is in fact an equality by Slater's conditions.

Karush-Kuhn-Tucker optimality conditions

Consider now a not necessarily convex problem of the form

$$\begin{aligned} &\text{minimize} && f(x) \\ &\text{subject to} && f_i(x) \le 0, \\ & && h_j(x) = 0, \end{aligned} \tag{16.1}$$

If p^* is the optimal solution of (16.1) and (λ, μ) dual variables, then we have seen that (this holds even in the non-convex case)

$$p^* \ge g(\lambda, \mu).$$

From this is follows that for any primal feasible point x,

$$f(x) - p^* \le f(x) - g(\lambda, \mu).$$

The difference $f(x) - g(\lambda, \mu)$ between the primal objective function at a primal feasible point and the dual objective function at a dual feasible point is called the duality gap at x and (λ, μ). For any such pairs of points we know that

$$p^* \in [g(\lambda, \mu), f(x)]$$

and if the gap is small we have a good approximation of the primal and dual optimal values. The duality gap can be used in iterative algorithms to define stopping criteria: if the algorithm generates a sequence of primal-dual variables (x_k, λ_k, μ_k), then we can stop if the duality gap is less than, say, a predefined tolerance ε.

Now suppose that we have points $x^*, (\lambda^*, \mu^*)$ such that the duality gap is zero. Then

$$\begin{aligned} f(x^*) &= g(\lambda^*, \mu^*) \\ &= \inf_x \left\{ f(x) + \sum_{i=1}^m \lambda_i^* f_i(x) + \sum_{j=1}^p \mu_j^* h_j(x) \right\} \\ &\le f(x^*) + \sum_{i=1}^m \lambda_i^* f_i(x^*) + \sum_{j=1}^p \mu_j^* h_j(x^*) \\ &\le f(x^*), \end{aligned}$$

where the last inequality follows from the fact that $h_j(x^*) = 0$ and $\lambda_i f_i(x) \le 0$ for $1 \le j \le p$ and $1 \le i \le m$. It follows that the inequalities are in fact equalities. From the identity

$$f(x^*) + \sum_{i=1}^m \lambda_i^* f_i(x^*) = f(x^*)$$

and $\lambda_i^* \ge 0$ and $f_i(x^*) \le 0$ we also conclude that at such optimal points,

$$\lambda_i^* f_i(x^*) = 0, \quad 1 \le i \le m.$$

This condition is known as complementary slackness. From the above we also see that x^* minimizes the Lagrangian $L(x, \lambda^*, \mu^*)$, so that the gradient of that function is zero:

$$\nabla_x L(x^*, \lambda^*, \mu^*) = 0.$$

Collecting these conditions (primal and dual feasibility, complementary slackness, vanishing gradient), we arrive at a set of optimality conditions known as the Karush-Kuhn-Tucker (KKT) conditions.

Theorem 16.4. (KKT conditions) Let $x*$ and $(\lambda*, \mu*)$ be primal and dual optimal solutions of (16.1) with zero duality gap. The the following conditions are satisfied:

$$\begin{aligned} f(x*) &\leq 0 \\ h(x*) &= 0 \\ \lambda* &\geq 0 \\ \lambda* f_i(x*) &= 0, 1 \leq i \leq m \end{aligned}$$

$$\nabla_x f(x*) + \sum_{i=1}^{m} \lambda_i* \nabla_x f_i(x*) + \sum_{j=1}^{p} \mu_j* \nabla* h_j(x*) = 0.$$

Moreover, if the problem is convex and the Slater Conditions (Theorem 16.1) are satisfied, then any points satisfying the KKT conditions have zero duality gap.

Notes

17

Support Vector Machines I

In this lecture we return to the task of classification. As seen earlier, examples include spam filters, letter recognition, or text classification. In this lecture we introduce a popular method for classification, Support Vector Machines (SVMs), from the point of view of convex optimization.

Linear Support Vector Machines

In the simplest case there is a set of labels $Y = \{-1, 1\}$ and the set of training points $\{x_1, \ldots, x_n\} \subset \mathbb{R}^d$ is linearly separable: this means that there exists an affine hyperplane $h(x) = w^\top x + b$ such that $h(x_i) > 0$ if $y_i = 1$ and $h(x_j) < 0$ if $y_j = -1$. We call the points for which $y_i = 1$ positive, and the ones for which $y_j = -1$ negative. The problem of finding such a hyperplane can be posed as a linear programming feasibility problem as follows: we look for a vector of weights w and a bias term b (together a $(p+1)$-dimensional vector) such that

$$w^\top x_i + b \geq 1, \quad \text{for } y_i = 1, \qquad w^\top x_j + b \leq -1, \quad \text{for } y_j = -1.$$

Note that we can replace the $+1$ and -1 with any other positve or negative quantity by rescaling the w and b, so this is just convention. We can also describe the two inequalities concisely as

$$y_i(w^\top x_i + b) - 1 \geq 0. \tag{17.1}$$

A hyperplane separating the two point sets will in general not be unique. As we want to use the linear classifier on new, yet unknown data, we want to find a separating hyperplane with best possible margin. Let δ^+ and δ^- denote the distance of a separating hyperplane to the closest positive and closest negative point, respectively. The quantity $\delta = \delta^+ + \delta^-$ is then called the margin of the classifier, and we want to find a hyperplane with largest possible margin.

We next show that the margin for a separating hyperplane that satisfies (17.1) is $\delta = 2/\|w\|_2$. Given a hyperplane H described in (17.1) and a point x such that we have the equality $w^\top x + b = 1$ (the point is as close as possible to the hyperplane, also called a support vector), the distance of that point to the hyperplane can be computed by first taking the difference of x with a point p on H (an anchor), and then computing the dot product of $x - p$ with the unit vector $w/\|w\|$ normal to H.

As anchor point p we can just choose a multiple cw that is on the plane, i.e., that satisfies

$$\langle w, cw \rangle + b = 0$$

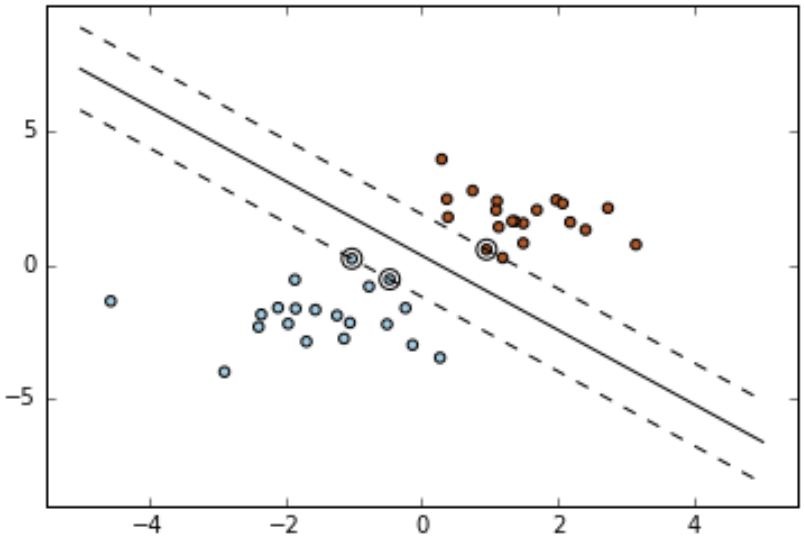


Figure 17.1: A hyperplane separating two sets of points with margin and support vectors.

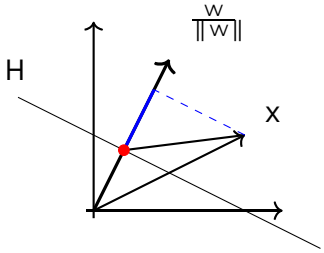


Figure 17.2: Computing the distance to the hyperplane

0. This implies that $c = b/\|w\|^2$, and consequently $p = -(b/\|w\|^2)w$. The distance is then

$$\delta = \left\langle x, \frac{w}{\|w\|} \right\rangle + \frac{b}{\|w\|^2}\left\langle w, \frac{w}{\|w\|} \right\rangle$$

$$= \frac{1}{\|w\|}\left(\langle w, x\rangle + b\right) = \frac{1}{\|w\|}(1-1) = 0.$$

Similarly, we get $\delta = -1/\|w\|$. The margin of this particular separating hyperplane is thus $\delta = 2/\|w\|$. If we want to find a hyperplane with largest margin, we thus have to solve the quadratic optimization problem

$$\underset{w,b}{\text{minimize}} \frac{1}{2}\|w\|^2$$

$$\text{subject to } y_i(\langle w, x_i\rangle + b) \geq 1.$$

Note that b is also an unknown variable in this problem! The factor $1/2$ in the objective function is just to make the gradient look nicer. The Lagrangian of this problem is

$$L(w,b,\lambda) = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{m} \lambda_i y_i (w^\top x_i - b) + \sum_{i=1}^{m} \lambda_i$$
$$= \frac{1}{2} w^\top w - w^\top X^\top \lambda + b\, y^\top \lambda + \sum_{i=1}^{m} \lambda_i,$$

where we denote by X the matrix with the $y_i x_i$ as rows. We can then write the conditions on the gradient with respect to w and b of the Lagrangian as

$$\begin{aligned} \nabla_w L(w,b,\lambda) &= w - X^\top \lambda = 0 \\ \frac{\partial L}{\partial b}(w,b,\lambda) &= y^\top \lambda = 0. \end{aligned} \tag{17.2}$$

If $y^\top \lambda \neq 0$, then the conditions (17.2) cannot be satisfied and the problem is unbounded from below. If $y^\top \lambda = 0$, then the first condition in (17.2) is necessary and sufficient for a minimizer. Replacing w by $X^\top \lambda$ and $\lambda^\top y$ by 0 in the Lagrangian gives the expression for the Lagrange dual $g(\lambda)$, $-\frac{1}{2}\lambda^\top X X^\top \lambda + m\lambda^\top y$ if $\lambda = 0$ $g(\lambda) = -\infty$ else.

Finally, maximizing this function and changing the sign, so that the maximum becomes a minimum, we can formulate the Lagrange dual optimization problem as

$$\text{minimize}_{\lambda} \quad \frac{1}{2}\lambda^\top X X^\top \lambda - e^\top \lambda \quad \text{subject to } \lambda \geq 0, \tag{17.3}$$

where e is the vector of all ones.

Notes

18

Support Vector Machines II

The linear separation problem was introduced as the problem of finding a hyperplane

$$H = \{x \in \mathbb{R}^d : w^T x + b = 0\}$$

that would separate data points x_i with label y_i from data points x_j with label $y_j = -1$. For convenience, we collect our data in matrices and vectors, where $X \in \mathbb{R}^{n \times d}$ is the matrix with rows $y_i x_i^T$ and y the vectors of labels y_i.

Assuming that the data is linearly separable, the problem of finding a separating hyperplane of maximal margin can be formulated as

$$\text{minimize} \quad \frac{1}{2}\|w\|^2 \quad \text{subject to} \quad e - by - X^T x \leq 0, \tag{P}$$

where e is the vector of all ones. The Lagrange dual optimization problem is

$$\text{minimize} \quad \frac{1}{2}\lambda^T X X^T \lambda - \lambda^T e \quad \text{subject to} \quad \lambda \geq 0. \tag{D}$$

Note that there is one dual variable λ_i per data point i. We can find the optimal value by solving the dual problem (D), but that does not give us automatically the weights w and the bias b. We can find the weights by $w = X^T\lambda$. As for b, this is best determined from the KKT conditions of the problem. These can be written by combining the constraints of the primal problem with the conditions on the gradient of the Lagrangian, the condition $\lambda \geq 0$, and complementary slackness as

$$\begin{aligned} Xw + yb - e &\geq 0 \\ \lambda &\geq 0 \\ \lambda_i(1 - y_i(w^T x_i + b)) &= 0 \quad \text{for } 1 \leq i \leq n \\ w - X^T \lambda &= 0 \\ y^T \lambda &= 0 \end{aligned}$$

To get b, we can choose one of the equations in which $\lambda_i > 0$, and then find b by setting $b = y_i(1 - y_i w^T x_i)$. With the KKT conditions written down, we can go about solving the problem of finding a maximum margin linear classifier using methods such as the barrier method.

Extensions

So far we looked at the particularly simple case where (a) the data falls into two classes, (b) the points can actually be well separated, and (c) they can be separated by an affine hyperplane. In reality, these three assumptions may not hold. We briefly discuss extensions of the basic model to account for the three situations just mentioned.

Non-exact separation

What happens when the data can not be separated by a hyperplane? In this case the constraints can not be satisfied: there is no feasible solution to the problem. We can still modify the problem to allow for misclassification: we want to find a hyperplane that separates the two point sets as good as possible, but we allow for some mistakes.

One approach is to add an additional set of n slack variables s1, . . . , sn, and modify the constraints to

w > x + b

≥1−s,fory=1,w>iiixj+b≤−1+sj,foryj=−1,si≥0.

T ∑hei-

thdatapointcanlandonthewrongsideofthehyperplaneifsi>1,andconsequentlythesumni=1siisanupperboun

donthenumberoferrorspossible.Ifwewanttominimizethenumberofmisclassifiedpoints,wemaywanttomini

mizethisupperbound,soasensiblechoiceforobjective

function would be to add a multiple of this sum. The new problem thus becomes

$$\begin{aligned} &\text{minimize} && \tfrac{1}{2}\| w \|^2 + \mu \sum_{j=1}^{n} s_j \\ &\text{subject to} && y_i(\langle w, x_i \rangle + b) \geq 1 - s_i, \\ & && s_i \geq 0, \quad 1 \leq i \leq n \end{aligned}$$

for some parameter μ. The Lagrangian of this problem and the KKT conditions can be derived in a similar way as in the separable case and are left as an exercise.

Non-linear separation and kernels

The key to extending SVMs from linear to non-linear separation is the observation that the dual form of the optimization problem (D) depends only on the dot products ⟨xi,xj ofthedatapoints.Infact,the

(i, j)-th entry of the matrix XX> is precisely ⟨xi, xj ⟩ ! H

If we map our data into a higher (possibly infinite) dimensional space ,

$$\Phi : \mathbb{R}^p \to H,$$

and consider the data points ϕ(xi), 1 n ≤i≤ ,thenapplyingthesupportvectormachinetothesehigher dimensional vectors will only depend on the dot products

$$K(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle .$$

The function K is called a kernel function. A typical example, often used in practice, is the Gaussian radial basis function (RBF),

$$K(x,y) = e^{-\|x-y\|^2/2\sigma^2}.$$

Note that we don't need to know how the function ϕ looks like! In the equation for the hyperplane we simply replace w>x with K(w, x). The only difference now is that the function ceases to be linear in x: we get a non-linear decision boundary.

Multiple classes

One is often interested in classifying data into more than two classes. There are two simple ways in which support vector machines can be extended for such problems: one-vs-one and one-vs-rest. In the one-vs-one case, given k classes, we train one classifier for each pair of classes in the training data, obtaining a total of k(k – 1)/2 classifiers. When it comes to prediction, we apply each of the classifiers to our test data and choose the class that was chosen the most among all the classifiers. In the one-vs-rest approach, each train k binary classifiers: in each one, one class corresponds to a chosen class, and the second class corresponds to the rest. By associating confidence scores to each classifier, we choose the one with the highest confidence score.

Example 18.1. An example that uses all three extensions mentioned is handwritten digit recognition. Suppose we have a series of pixels, each representing a number, and associated labels $\{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\}$. We would like to train a support vector machine to recognize new digits. Given the knowledge we have, we can implement this task using standard optimization software such as CVXPY. Luckily, there are packages that have this functionality already implemented, such as the SCIKIT-LEARN package for Python.
We illustrate its functioning below. The code also illustrates some standard procedures when tackling a machine learning problem:

- Separate the data set randomly into training data and test data;
- Create a support vector classifier with optional parameters;
- Train (using `FIT`) the classifier with the training data;
- Predict the response using the test data and compare with the true response;
- Report the results.

An important aspect to keep in mind is that when testing the performance using the test data, we should compare the classification accuracy to a naive baseline: if, for example, 80% of the test data is classified as +1, then making a prediction of +1 for all the data will give us an accuracy of 80%; in this case, we would want our classifier to perform considerably better than getting the right answer 80% of the time!

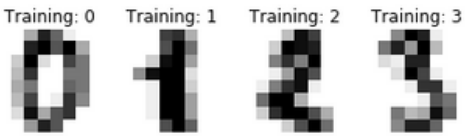
In [15]:
```
import numpy as np
import matplotlib.pyplot as plt
% matplotlib inline
from sklearn import svm, datasets, metrics
from sklearn.model_selection import train_test_split
```

In [16]:

```
digits = datasets.load_digits()

# Display images and labels
images_and_labels = list(zip(digits.images, digits.target))
for index, (image, label) in enumerate(images_and_labels[:4]):
plt.subplot(2, 4, index + 1)
plt.axis('off')
        plt.imshow(image, cmap=plt.cm.gray_r, interpolation='nearest')
plt.title('Training: %i' % label)

# Turn images into 1-D arrays
n_samples = len(digits.images)
data = digits.images.reshape((n_samples, -1))

# Create classifier
svc = svm.SVC(gamma=0.001)

# Randomly split data into train and test set
X_train, X_test, y_train, y_test = train_test_split(data,
digits.target, test_size = 0.4, random_state=0)
svc.fit(X_train, y_train)
```

Out [2]:

```
SVC(C=1.0, cache_size=200, class_weight=None, coef0=0.0,
decision_function_shape=None, degree=3, gamma=0.001,
        kernel='rbf', max_iter=-1, probability=False, random_state=None,
shrinking=True, tol=0.001, verbose=False)
```



Now apply prediction to test set and report performance.

In [3]:

```
predicted = svc.predict(X_test)
print("Classification report for classifier %s:\n%s\n"
                % (svc, metrics.classification_report(y_test, predicted)))
```

Out [3]:

```
Classification report for classifier SVC(C=1.0, cache_size=200,
class_weight=None, coef0=0.0,vdecision_function_shape=None,
      degree=3, gamma=0.001, kernel='rbf', max_iter=-1, probability=False,
random_state=None, shrinking=True, tol=0.001, verbose=False):
precision recall f1-score support

0 1.00 1.00 1.00 60
1 0.97 1.00 0.99 73
2 1.00 0.97 0.99 71
3 1.00 1.00 1.00 70
4 1.00 1.00 1.00 63
5 1.00 0.98 0.99 89
6 0.99 1.00 0.99 76
7 0.98 1.00 0.99 65
8 1.00 0.99 0.99 78
9 0.99 1.00 0.99 74

avg/total 0.99 0.99 0.99 719
```

In [4]:

```
import skimage
from skimage import data
from skimage.transform import resize
from skimage import io
import os
```

Now try this out on some original data!

In [5]:

```
mydigit1 = io.imread('images/digit9.png')
mydigit2 = io.imread('images/digit4.png')
plt.figure(figsize=(8, 4))
plt.subplot(1,2,1)
plt.imshow(mydigit1,  cmap=plt.cm.gray_r,  interpolation='nearest')
plt.axis('off')
plt.subplot(1,2,2)
plt.imshow(mydigit2,  cmap=plt.cm.gray_r,  interpolation='nearest')
plt.axis('off')
plt.show()
```

In [6]:

```
smalldigit1 = resize(mydigit1, (8,8))
smalldigit2 = resize(mydigit2, (8,8))
mydigits = np.concatenate((np.round(15*(np.ones((8,8))-
                                        smalldigit1[:,:,0])).reshape((64,1)).T,
np.round(15*(np.ones((8,8))-
smalldigit2[:,:,0])).reshape((64,1)).T),axis=0)
# After some preprocessing, make prediction
guess = svc.predict(mydigits)
print guess
```

[94]

Notes

19

Iterative Algorithms

Most modern optimization methods are iterative: they generate a sequence of points x0, x1, . . . in Rd in the hope that this sequences will converge to a local or global minimizer x^* of a function f(x). A typical rule for generating such a sequence would be to start with a vector x0, chosen by an educated guess, and then for k ≥ 0, move from step k to k + 1 by

$$x_{k+1} = x_k + \alpha_k p_k,$$

in a way that ensures that $f(x_{k+1}) \leq f(x_k)$. The parameter α_k is called the steplength, while p_k is the search direction. The are many ways in which the direction pk and the step length αk can be chosen. If we take

$$p_k = -\nabla f(x_k), \quad (19.1)$$

then we take a step in the direction of steepest descent and the resulting method is (unsurprisingly) called gradient descent. If there is second-order information available, then we can take steps of the form

$$p_k = -\nabla^2 f(x_k)^{-1} \nabla f(x_k). \quad (19.2)$$

The resulting method is called Newton's Method. If applicable, Newton's method tends to converge faster to a solution, but the computation at each step is more expensive.

Gradient descent

In the method of gradient descent, the search direction is chosen as

$$p_k = -\nabla f(x_k).$$

To see why this makes sense, let p be a direction and consider the Taylor expansion

$$f(x_k + \alpha p) = f(x_k) + \alpha \langle p, \nabla f(x_k) \rangle + O(\alpha^2).$$

Considering this as a function of α, the rate of change in direction p at x_k is the derivative of this function at α=0,

$$\frac{d f(x_k + \alpha p)}{d \alpha} \Big|_{\alpha=0} = \langle p, \nabla f(x_k) \rangle,$$

also known as the directional derivative of f at x_k in the direction p. This formula indicates that the rate of change is negative, and we have a descent direction, if $\langle p, \nabla f(x_k) \rangle < 0$.

The Cauchy-Schwarz inequality gives the bounds

$$-\|p\|_2 \|\nabla f(x_k)\|_2 \leq \langle p, \nabla f(x_k) \rangle \leq \|p\|_2 \|\nabla f(x_k)\|_2$$

We see that the rate of change is the smallest when the first inequality is an equality, which happens if

$$p = -\alpha \nabla f(x_k)$$

for some $\alpha > 0$.

For a visual interpretation of what it means to be a descent direction, note that the angle θ between a vector p and the gradient $\nabla f(x)$ at a point is given by (see Preliminaries, Page 9)

$$\langle p, \nabla f(x) \rangle = \|p\|_2 \|\nabla f(x)\| \cos(\theta)$$

This is negative if the vector p forms and angle greater than with the gradient. Recall that the gradient points in the direction of steepest ascent, and is orthogonal to the level sets. If you are standing on the slope of a mountain, walking along the level set lines will not change your elevation, the gradient points to the steepest upward direction, and the negative gradient to the steepest descent.

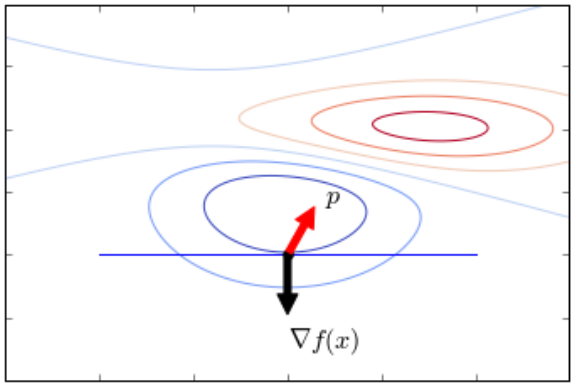


Figure 19.1: A descent direction

Any multiple $\alpha \nabla f(x_k)$ points in the direction of steepest descent, but we have to choose a sensible parameter α to ensure that we make sufficient progress, but at the same time don't overshoot. Ideally, we would choose the value α_k that minimizes $f(x_k - \alpha_k \nabla f(x_k))$. While finding such a minimizer is in general not easy (see Section Lecture 4 for alternatives), for quadratic functions in can be given in closed form.

Linear least squares

Consider a function of the form

$$f(x) = \frac{1}{2}\|Ax - b\|_2^2.$$

The Hessian is symmetric and positive semidefinite, with the gradient given by

$$\nabla f(x) = A^\top (Ax - b).$$

The method of gradient descent proceeds as

$$x_{k+1} = x_k - \alpha_k A^\top (Ax_k - b).$$

To find the best α_k, we compute the minimum of the function

$$\alpha \mapsto \phi(\alpha) = f(x_k - \alpha A^\top (Ax_k - b)). \quad (19.3)$$

If we set $r_k := A^\top (b - Ax_k) = -\nabla f(x_k)$ and compute the minimum of (19.3) by setting the derivative to zero,

$$\begin{aligned}
\phi'(\alpha) &= \frac{d}{d\alpha} f(x_k + \alpha r_k) = \langle \nabla f(x_k + \alpha r_k), r_k \rangle \\
&= \langle A^\top (A(x_k + \alpha r_k) - b), r_k \rangle \\
&= \langle A^\top (Ax_k - b), r_k \rangle + \alpha \langle A^\top A r_k, r_k \rangle
\end{aligned}$$

we get the step length

$$\alpha_k = \frac{r_k^\top r_k}{r_k^\top A^\top A r_k} = \frac{\|r_k\|_2^2}{\|Ar_k\|_2^2}.$$

Note also that when we have r_k and α_k, we can compute the next r_{k+1} as

$$\begin{aligned}
r_{k+1} &= A^\top (b - Ax_{k+1}) \\
&= A^\top (b - A(x_k + \alpha_k r_k)) \\
&= A^\top (b - Ax_k) - \alpha_k A^\top A r_k = r_k - \alpha_k A^\top A r_k.
\end{aligned}$$

The gradient descent algorithm for the linear least squares problem proceeds by first computing $r_0 = A^\top (b - Ax_0)$, and then at each step

$$\begin{aligned}
\alpha_k &= \frac{r_k^\top r_k}{r_k^\top A^\top A r_k} \\
x_{k+1} &= x_k + \alpha_k r_k \\
r_{k+1} &= r_k - \alpha_k A^\top A r_k.
\end{aligned}$$

Does this work? How do we know when to stop? It is worth noting that the residual satisfies $r_k = 0$ if and only if x_k is a stationary point, in our case, a minimizer. One criteria for stopping could then be to check whether $\|r_k\|_2 \le \varepsilon$ for some given tolerance $\varepsilon > 0$. One potential problem with this criterion is that the function can become flat long before reaching a minimum, so an alternative stopping method would be to stop when the difference between two successive points, $\|x_{k+1} - x_k\|_2$, becomes smaller than some $\varepsilon > 0$.

Example 19.1. We plot the trajectory of gradient descent with the data

$$A = \begin{bmatrix} 2 & 3 \\ 10 & 1 \end{bmatrix}, \quad b = \begin{bmatrix} 1 \\ -1 \end{bmatrix}.$$

As can be seen from the plot, we always move in the direction orthogonal to a level set, and stop at a point where we are tangent to a level set.

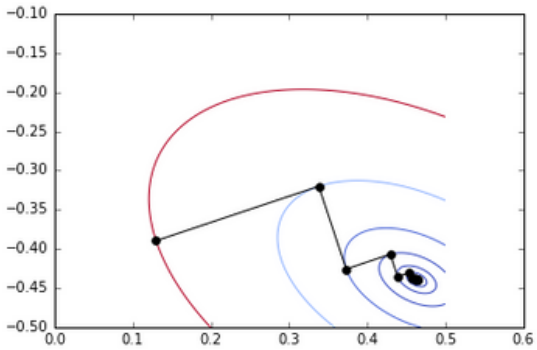


Figure 19.2: Trajectory of gradient descent

Step length selection

While for a quadratic function of the form $||Ax-b||2$ it was possible to find a step length αk by minimizing the function in the direction of steepest descent, in general this may not be possible or even desirable. The step length is often called the learning rate in machine learning. A good step length

• is not too small (so that the algorithm does not take too long);

• is not too large (we might end up at a point with larger function value);

• is easy to compute.

There are conditions (such as the Armijo-Goldstein or the Wolfe conditions) that ensure a sufficient decrease at each step. Another common approach is backtracking: in this method one uses a high initial value of α (for example, $\alpha = 1$), and then decreases it until the sufficient descent condition is satisfied.

20

Convergence

Iterative algorithms for solving a problem of the form

$$\text{minimize} f(x), \quad x \in \mathbb{R}^d \tag{20.1}$$

generate a sequence of vectors $x^0, x^1, \ldots$ in the hope that this sequence converges to a (local or global) minimizer x^* of (20.1). In this lecture we study what it means for a sequence to converge, and how to quantify the speed of convergence. We then study the convergence of of gradient descent for quadratic functions and for convex functions satisfying certain smoothness assumptions.

Convergence of iterative methods

A sequence of vectors $\{x^k\}_{k \in \mathbb{N}_0} \subset \mathbb{R}^d$ converges to x^* with respect to a norm $\|\cdot\|$ as $k \to \infty$, written $x^k \to x$, if the sequence of numbers $\|x^k - x^*\|$ converges to zero. Iterative algorithms will typically not find the exact solution to a problem like (20.1). In fact, computers are not capable of telling very small numbers (say, 2^{-53} in double precision arithmetic) from 0, so finding a numerically exact solution is in general not possible. In addition, in machine learning, high accuracy is not necessary or even desirable due to the unavoidable statistical error.

Definition 20.1. Assume that a sequence of vectors $\{x^k\}_{k \in \mathbb{N}_0}$ converges to x^*. Then the sequence is said to converge

(a) linearly (or Q-linear, Q for Quotient), if there exist an $r \in (0, 1)$ such that for sufficiently large k,

$$\|x^{k+1} - x^*\| \leq r \, \|x^k - x^*\|.$$

(b) superlinearly, if

$$\lim_{k \to \infty} \frac{\|x^{k+1} - x^*\|}{\|x^k - x^*\|} = 0,$$

(c) with order p, if there exists a constant $M > 0$, such that for sufficiently large k,

$$\|x^{k+1} - x^*\| \leq M \, \|x^k - x^*\|^p.$$

The case $p = 2$ is called quadratic convergence

These definitions depend on the choice of a norm, but any two norms on Rd are equivalent, convergence with respect to one norm implies convergence with respect to any other norm. Note that the definitions above start with the assumption that the sequence {xk} converges to x^*. Therefore, for sufficiently large k, $\|x_k - x^*\| < 1$ and if {xk} converges with order of convergence $p > 1$, then

$$\frac{\|x_{k+1} - x^*\|}{\|x_k - x^*\|^p} \leq M \|x_k - x^*\|^{p-1} < M.$$

This shows that convergence of order p implies convergence of any lower order and also superlinear convergence.

Example 20.2. Consider the sequence of numbers $x_k = r^k$ for some r. Clearly, $x_k \to x^* = 0$ as $k \to \infty$ for $0 < r < 1$. Moreover,

$$x_{k+1} = \frac{1}{2r^{k+1}}(2r^k)^2 = 2r^{2k}r^{-k-1} = r^k x_k,$$

which shows that the sequence has rate of convergence r.

Convergence of gradient descent for least squares

Throughout this section, $\|\cdot\|$ refers to the ℓ_2-norm. We study the convergence of gradient descent for the least squares problem

$$\text{minimize } f(x) = \frac{1}{2}\|Ax-b\|_2, \quad (20.2)$$

where $A \in \mathbb{R}^{m\times n}$ with $m \geq n$ is a matrix of full rank. The function $f(x)$ is convex, since it is a quadratic function with positive semi-definite Hessian $A^T A$. Gradient descent produces a sequence of vectors by the rule

$$x_{k+1} = x_k + \alpha_k r_k,$$

where the step length α_k and the residual r_k are given by

$$\alpha_k = \frac{\|r_k\|_2^2}{\|Ar_k\|^2}, \qquad r_k = A^T(b - Ax_k) = -\nabla f(x_k).$$

At the minimizer x^*, the residual is $r = -\nabla f(x^*) = 0$ and if the sequence x_k converges to x^*, the norms of the residuals converge to 0. Conversely, the residual is related to the difference $x_k - x^*$ by

$$r_k = A^T(b - Ax_k) = A^T b - Ax_k - (b - Ax^*) = A^T A(x_k - x^*). \quad (20.3)$$

Therefore

$$\|x_k - x^*\| = \|(A^T A)^{-1} r_k\| \leq \|(A^T A)^{-1}\| \|r_k\|$$

where $\|Bx\| \leq \|B\| \|x\|$ for $x \neq 0$ is the operator norm of a matrix with respect to the 2-norm. Consequently, if the sequence r_k converges to zero, so does the sequence $\|x_k - x^*\|$. A reasonable criterion to stop the algorithm is therefore when the residual norm $\|r_k\|$ is smaller than a predefined tolerance $\varepsilon > 0$.

The following theorem (whose proof we omit) shows that the gradient descent method for linear least squares converges linearly with respect to the $A^T A$ norm. The statement involves the condition number of A[1]. This quantity is defined as $\kappa(A) = \|A\| \|A^{-1}\|$.

[1] The concept of condition number, introduced by Alan Turing while in Manchester, is one of the most important ideas in numerical analysis, as it is indispensable in studying the performance of numerical algorithms.

where $A^\dagger$ is the pseudoinverse of A. If $A \in \mathbb{R}^{m \times n}$ with $m \geq n$ and linearly independent columns, it is defined as $A^\dagger = (A^\top A)^{-1} A^\top$. The condition number is always greater than or equal to one.

Theorem 20.3. The error in the $k+1$-th iterate is bounded by

$$\|x_{k+1} - x^*\| \leq \left(\frac{\kappa^2(A) - 1}{\kappa^2(A) + 1} \right) \|x_k - x^*\|.$$

In particular, the gradient descent algorithm converges linearly. We can deduce Theorem 20.3 as a special case of a more general convergence result from convex function satisfying certain smoothness assumptions.

Notes

21

Gradient Descent

In this lecture we will derive a convergence result for gradient descent applied to a convex function $f \in C^1(\mathbb{R}^d)$. Convergence results in optimization are often stated in terms of the difference $f(x_k) - f^*$, where $f^* = f(x^*)$ and x^* is a minimizer of f. By the convexity of f, we have

$$f(x_k) - f(x^*) \leq \langle \nabla f(x_k), x_k - x^* \rangle \leq \| \nabla f(x_k) \| \cdot \| x_k - x^* \|, \quad (21.1)$$

which allows to relate the convergence of $f(x_k) - f^*$ to the convergence of $\|x_k - x^*\|$. In order to guarantee good convergence rates, we need some additional smoothness and boundedness conditions.

Gradient descent for smooth convex functions

The most common smoothness condition in optimization is Lipschitz continuity, applied to a function or to its gradient.

Definition 21.1. A function $f : \mathbb{R}^d \to \mathbb{R}^k$ is called Lipschitz continuous with Lipschitz constant $L > 0$, if for all $x, y \in \mathbb{R}^d$,

$$\| f(x) - f(y) \| \leq L \| x - y \|.$$

A function $f : \mathbb{R}^d \to \mathbb{R}$ is called β-smooth for some $\beta > 0$ if the gradient is Lipschitz continuous:

$$\| \nabla f(x) - \nabla f(y) \| \leq \beta \cdot \| x - y \|$$

for all $x, y \in \mathbb{R}^d$.

Lipschitz continuity lies somewhere between continuity and differentiability. If $f \in C^1(\mathbb{R}^d)$ is Lipschitz continuous with Lipschitz constant L, then $\| \nabla f(x) \| \leq L$. Similarly, β-smoothness implies that $\| \nabla^2 f(x) \| \leq \beta$. Recall from Lecture 14 that a function $f \in C^1(\mathbb{R}^d)$ is convex if and only if

$$f(y) + \langle \nabla f(y), x - y \rangle \leq f(x) \quad (21.2)$$

for all $x, y \in \mathbb{R}^d$. The following result shows that β-smoothness is equivalent to a quadratic upper bound on the difference between the function value $f(x)$ and its linear lower bound (21.2).

Lemma 21.2. Let $f \in C^1(\mathbb{R}^d)$ be β-smooth and convex. Then, for any $x, y \in \mathbb{R}^d$,

$$f(x) \leq f(y) + \langle \nabla f(y), x - y \rangle + \frac{\beta}{2} \| x - y \|^2. \quad (21.3)$$

Conversely, if a convex function $f \in C^1(\mathbb{R}^d)$ satisfies (21.3), then for all $x, y \in \mathbb{R}^d$,

$$\frac{1}{\beta}\|\nabla f(x) - \nabla f(y)\|^2 \le \langle x - y, \nabla f(x) - \nabla f(y) \rangle . \tag{21.4}$$

In particular, f is β-smooth.

Proof. The first inequality follows from the convexity assumption. For the second inequality, represent $f(x) - f(y)$ as an integral:

$$f(x) - f(y) = \int_0^1 \langle \nabla f(y + t(x - y)), x - y \rangle \, dt.$$

We can then write

$$\begin{aligned} f(x) - f(y) - \langle \nabla f(y), x - y \rangle &= \int_0^1 \langle \nabla f(y + t(x - y)) - \nabla f(y), x - y \rangle \, dt \\ &\le \int_0^1 \|\nabla f(y + t(x - y)) - \nabla f(y)\| \|x - y\| \, dt \\ &\le \int_0^1 \beta t \|x - y\|^2 \, dt \\ &= \frac{\beta}{2}\|x - y\|^2 , \end{aligned}$$

where the first inequality follows from applying Cauchy-Schwartz, and the second from the assumption of β-smoothness.

For the second claim, assume that f satisfies the bound (21.3). For any $x, y, z \in \mathbb{R}^d$ we have

$$\begin{aligned} f(x) - f(y) &= (f(x) - f(z)) + (f(z) - f(y)) \\ &\le \langle \nabla f(x), x - z \rangle + \langle \nabla f(y), z - y \rangle + \frac{\beta}{2}\|z - y\|^2 , \end{aligned}$$

where we used the convexity of f to bound $f(z) - f(y)$. If we now set $z = y - (1/\beta)(\nabla f(y) - \nabla f(x))$ and simplify the resulting expression, we get

$$f(x) - f(y) \le \langle \nabla f(x), x - y \rangle - \frac{1}{2\beta}\|\nabla f(x) - \nabla f(y)\|^2 .$$

Adding this expression to the same one with the roles of x and y exchanged, we get

$$0 \le \langle \nabla f(x) - \nabla f(y), x - y \rangle - \frac{1}{\beta}\|\nabla f(x) - \nabla f(y)\|^2 .$$

The fact that this implies β-smoothness of f follows from the Cauchy-Schwartz inequality. □

We can, and will, use (21.3) and (21.4) as alternative definitions of β-smoothness.

Theorem 21.3. Let $f \in C^1(\mathbb{R}^d)$ be a function that is β-smooth and convex. Then for any $x^0 \in \mathbb{R}^d$, the iterates $\{x^k\}$ generated by gradient descent with constant step length $1/\beta$ satisfy

$$f(x^k) - f^* \le \frac{2\beta}{k}\|x^0 - x^*\|^2 ,$$

where $f^* = f(x^*)$ and x^* is a minimizer of f.

Proof. Observe first that

$$\begin{aligned}\|x^{k+1} - x^*\|^2 &= \|x^k - (1/\beta)\nabla f(x^k) - x^*\|^2 \\ &= \|x^k - x^*\|^2 - \frac{2}{\beta}\langle \nabla f(x^k), x^k - x^*\rangle + \frac{1}{\beta^2}\|\nabla f(x^k)\|^2 \\ &\overset{(21.4)}{\leq} \|x^k - x^*\|^2 - \frac{1}{\beta^2}\|\nabla f(x^k)\|^2 \leq \|x^k - x^*\|^2\end{aligned}$$

where in addition to (21.4) we used the fact that $\nabla f(x^*) = 0$. In particular, since $\|x^k - x^*\|$ is non-increasing, we get from (21.1) that

$$f(x^k) - f^* \leq \|\nabla f(x^k)\| \cdot \|x^k - x^*\| \leq \|\nabla f(x^k)\| \cdot \|x^0 - x^*\|. \tag{21.5}$$

Using (21.3) with $y = x^k$ and $x = x^k - (1/\beta)\nabla f(x^k)$, we get

$$f(x^{k+1}) \leq f(x^k) - \frac{1}{2\beta}\|\nabla f(x^k)\|^2.$$

Set $\Delta_k = f(x^k) - f^*$, so that $\Delta_{k+1} \leq \Delta_k$. Then

$$\Delta_{k+1} \leq \Delta_k - \frac{1}{2\beta}\|\nabla f(x^k)\|^2 \overset{(21.5)}{\leq} \Delta_k - \frac{1}{2\beta}\frac{\Delta_k^2}{\|x^0 - x^*\|^2}, \tag{21.6}$$

where we used (21.5) to lower-bound $\|\nabla f(x^k)\|$ in the second inequality. In particular, we see that $\Delta_{k+1} \leq \Delta_k$. We can rearrange the inequality (21.6) to

$$\Delta_{k+1} + \frac{\Delta_k^2}{2\beta\|x^0 - x^*\|^2} \leq \Delta_k \Rightarrow \frac{1}{\Delta_{k+1}} \geq \frac{1}{\Delta_k}\frac{1}{1 - \frac{\Delta_k}{2\beta\|x^0 - x^*\|^2}} \Rightarrow \frac{1}{\Delta_{k+1}} \geq \frac{1}{\Delta_k} + \frac{1}{2\beta\|x^0 - x^*\|^2},$$

where for the first implication we divided both sides by $\Delta_k\Delta_{k+1}$, and for the second implication we used that $\Delta_k/\Delta_{k+1} \geq 1$. Applying the same bound recursively to $\frac{1}{\Delta_k}, \frac{1}{\Delta_{k-1}}, \ldots, \frac{1}{\Delta_1}$, we get

$$\frac{1}{\Delta_{k+1}} \geq \frac{k+1}{2\beta\|x^0 - x^*\|^2} \Rightarrow \Delta_{k+1} \leq \frac{2\beta\|x^0 - x^*\|^2}{k+1}.$$

Taking the inverse, and shifting the index from $k+1$ to k, we get

$$f(x^k) - f^* \leq \Delta_k \leq \frac{2\beta\|x^0 - x^*\|^2}{k},$$

as claimed. □

We can get even better convergence when assuming strong convexity.

Definition 21.4. A function $f \in C^1(\mathbb{R}^d)$ is called α-strongly convex for some $\alpha > 0$ if for every $x, y \in \mathbb{R}^d$

$$f(y) + \langle \nabla f(y), x - y\rangle + \frac{\alpha}{2}\|x - y\|^2 \leq f(x).$$

If $\alpha = 0$ this is just the derivative characterization of convexity. Note that a function f is α-strongly convex if and only if the function $f(x) - \alpha\|x\|^2/2$ is convex.

Remark 21.5. The difference

$$D_f(x,y) := f(x) - f(y) - \langle \nabla f(y), x - y \rangle$$

is called the Bregman divergence associated to f. To say that a function $f \in C^1(\mathbb{R}^d)$ is α-strongly convex and β-smooth is to say that for any $x, y \in \mathbb{R}^d$,

$$\frac{\alpha}{2}\|x - y\|_2^2 \leq D_f(x,y) \leq \frac{\beta}{2}\|x - y\|_2^2.$$

This means that we can, locally, upper and lower bound the function by quadratic functions. In particular, $\beta \geq \alpha$.

Theorem 21.6. Let $f \in C^1(\mathbb{R}^d)$ be a function that is α-strongly convex and β-smooth. Then for any $x^0 \in \mathbb{R}^d$, the iterates of gradient descent with constant step length $2/(\alpha+\beta)$ satisfy

$$\|x^{k+1} - x^*\|_2 \leq \left(\frac{\kappa - 1}{\kappa + 1}\right) \|x^k - x^*\|_2,$$

where $\kappa = \beta/\alpha$.

Example 21.7. Assume that $\alpha = \beta$ (and therefore $\kappa = 1$) in Theorem 21.6. Then

$$f(x) = \frac{\alpha}{2}\|x\|_2^2.$$

The gradient is $\nabla f(x) = \alpha x$. Starting with $x^0 \in \mathbb{R}^d$, gradient descent with step length $2/(\alpha+\beta) = 1/\alpha$ gives

$$x^1 = x^0 - x^0 = 0 = x^*,$$

so that this converges in a single iteration.

Example 21.8. Let $f(x) = \frac{1}{2}\|Ax - b\|_2^2$ for $A \in \mathbb{R}^{m \times n}$ with $m \geq n$, and assume A has full rank. The difference between the function and its approximation is

$$\frac{1}{2}\|Ax - b\|_2^2 - \frac{1}{2}\|Ay - b\|_2^2 - (Ay - b)^T A(x - y) = \frac{1}{2}\|A(x - y)\|_2^2. \tag{21.7}$$

The largest and smallest singular values of the matrix A are defined as

$$\sigma_1(A) = \max_{x \,:\, \|x\|_2 = 1} \|Ax\|_2, \qquad \sigma_n(A) = \min_{x \,:\, \|x\|_2 = 1} \|Ax\|_2.$$

The term (21.7) is therefore bounded from above and below by the squares of the largest singular value and by the smallest singular value of A:

$$\sigma_1^2(A)\|x - y\|_2^2 \geq \|A(x - y)\|_2^2 \geq \sigma_n^2(A)\|x - y\|_2^2.$$

A well-known characterization of the condition number of a matrix is $\kappa(A) = \sigma_1(A)/\sigma_n(A)$, and from this we recover the convergence result from Lecture 20.

The proof of Theorem 21.6 relies on the following auxiliary result.

Lemma 21.9. Let f be α-strongly convex and β-smooth. Then for any $x, y \in \mathbb{R}^d$,

$$\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq \frac{\alpha\beta}{\alpha+\beta}\|x - y\|_2^2 + \frac{1}{\alpha+\beta}\|\nabla f(x) - \nabla f(y)\|_2^2.$$

Proof. Set $\phi(x) := f(x) - \frac{\alpha}{2}\|x\|^2$. Since f is α-strongly convex, $\phi(x)$ is convex. Moreover, $\nabla\phi(x) = \nabla f(x) - \alpha x$. We therefore get

$$\begin{aligned}\phi(x) - \phi(y) - \langle \nabla\phi(y), x - y\rangle &= f(x) - f(y) - \langle \nabla f(y), x - y\rangle - \frac{\alpha}{2}\left(\|x\|^2 - \|y\|^2 + \langle y, x - y\rangle\right) \\ &\overset{\text{Lemma 21.2}}{\leq} \frac{\beta}{2}\|x - y\|^2 - \frac{\alpha}{2}\left(\|x\|^2 - \|y\|^2 - 2\langle y, x\rangle\right) \\ &= \frac{\beta - \alpha}{2}\|x - y\|^2. \end{aligned} \tag{21.8}$$

From Lemma 21.2 it follows that ϕ is β -smooth and satisfies the inequality

$$\langle \nabla\phi(x) - \nabla\phi(y), x - y\rangle \geq \frac{1}{\beta}\|\nabla\phi(x) - \nabla\phi(y)\|^2.$$

Replacing ϕ in this expression, we get

$$\langle \nabla f(x) - \nabla f(y) - \alpha x + \alpha y, x - y\rangle \geq \frac{1}{\beta}\|\nabla f(x) - \nabla f(y) - \alpha x + \alpha y\|^2.$$

The left-hand side of this inequality gives

$$\langle \nabla f(x) - \nabla f(y) - \alpha x + \alpha y, x - y\rangle = \langle \nabla f(x) - \nabla f(y), x - y\rangle - \alpha\|x - y\|^2. \tag{21.9}$$

The right-hand side, on the other hand, gives

$$\begin{aligned}\frac{1}{\beta}\|\nabla f(x) - \nabla f(y) - \alpha x + \alpha y\|^2 &= \frac{1}{\beta}\|\nabla f(x) - \nabla f(y)\|^2 + \frac{\alpha^2}{\beta}\|x - y\|^2 - \frac{2\alpha}{\beta}\langle \nabla f(x) - \nabla f(y), x - y\rangle \\ &\leq \|\nabla f(x) - \nabla f(y)\|^2 - \frac{2\alpha}{\beta}\langle \nabla f(x) - \nabla f(y), x - y\rangle. \end{aligned} \tag{21.10}$$

Collecting the terms in (27.2) and (21.10) involving $\|\nabla f(x) - \nabla f(y)\|^2$ on the left, and the terms involving $\langle \nabla f(x) - \nabla f(y), x - y\rangle$ on the right, we get

$$\frac{\beta - \alpha}{\beta}\|\nabla f(x) - \nabla f(y)\|^2 \leq \alpha\beta\|x - y\|^2 - 2\alpha\langle \nabla f(x) - \nabla f(y), x - y\rangle.$$

Multiplying this expression with $\frac{\beta}{\alpha(\beta - \alpha)}$ gives the desired inequality. □

Proof of Theorem 21.6. Set $\eta = \frac{2}{\alpha + \beta}$. Since x^* is fixed, we will omit this term whenever it appears in the following (so that we can think of $\|x^k - x^*\|$ as the error in the kth step, whenever we see $f(x^k)$).

$$\begin{aligned}\|x^{k+1} - x^*\|^2 &= \|x^k - \eta\nabla f(x^k) - x^*\|^2 = \|(x^k - x^*) - \eta\nabla f(x^k)\|^2 \\ &= \|x^k - x^*\|^2 - 2\eta\langle \nabla f(x^k), x^k - x^*\rangle + \eta^2\|\nabla f(x^k)\|^2 \\ &\leq \left(\frac{\beta - \alpha}{\beta + \alpha}\right)^2\|x^k - x^*\|^2 - \frac{4}{(\beta + \alpha)^2}\|\nabla f(x^k)\|^2 + \frac{4}{(\beta + \alpha)^2}\|\nabla f(x^k)\|^2 \\ &= \left(\frac{\beta - \alpha}{\beta + \alpha}\right)^2\|x^k - x^*\|^2. \end{aligned}$$

where in the inequality we used Lemma 21.9 to bound the term $\langle \nabla f(x^k), x^k - x^*\rangle$, and simplified the resulting expression. The claim now follows by dividing the numerator and the denominator by α. □

Using the bound

$$\left(\frac{\kappa-1}{\kappa+1}\right)^2 \leq e^{-4/(\kappa+1)},$$

and the inequality $f(x^k) - f^* \leq (\beta/2)\|x^k - x^*\|^2$ from the β-smoothness assumption, we get the convergence bound

$$f(x^k) - f^* \leq \frac{\beta\|x^0 - x^*\|^2}{2} \cdot e^{-\frac{4}{\kappa+1}k},$$

which is a considerable improvement over the $1/k$ convergence bound when only assuming β-smoothness. In particular, the number of iterations to reach accuracy ε is of order $O(\log(1/\varepsilon))$.

Notes

The convergence of gradient descent under various smoothness assumptions is a classic theme in convex optimization. The presentation in this chapter is based on [4]. A standard reference for many of the tools used in the analysis of gradient descent (and a wealth of additional information) [19].

22

Extensions of Gradient Descent

We have seen that for a β-smooth convex function $f \in C^1(\mathbb{R}^d)$, the sequence of iterates $\{x^k\}_{k \geq 0}$ generated by gradient descent with steplength satisfies $1/\beta$

$$f(x^{k+1}) - f(x^*) = O(1/k),$$

where x^* is a minimizer of f. This implies that we need a number of iterations of order $O(1/\varepsilon)$ to reach accuracy ε. If in addition the function is α-strongly convex, then we get linear convergence with ratio $(\kappa - 1)/(\kappa+1)$, where $\kappa = \beta/\alpha$. For the convergence of the function value, this implies

$$f(x^{k+1}) - f(x^*) = O(e^{-4k/(\kappa+1)})$$

shows that only O(log(1/ε)) iterations are needed to reach accuracy ε. In this lecture we will have a look at two extensions of gradient descent. The first is an accelerated version, while the second extension covers common situations in which the function to be minimized is not differentiable.

Acceleration

Accelerated gradient descent, proposed by Y. Nesterov, begins with initial values $y^0 = x^0 = x^{-1} \in \mathbb{R}^d$ and the proceeds as follows for $k \geq 0$:

$$y^k = x^k + \frac{k-1}{k+2}(x^k - x^{k-1})$$
$$x^{k+1} = y^k - \eta \nabla f(y^k)$$

The method can be interpreted as carrying over some momentum from previous iterates: instead of only taking into account the current iterate, the gradient step is based on a combination of the current and the previous step. This method has favourable convergence properties.

Theorem 22.1. Let $f \in C^1(\mathbb{R}^d)$ be convex and β-smooth. Then accelerated gradient descent with step length $1/\beta$ converges to a minimizer x^* of f with rate

$$f(x^k) - f(x^*) \leq \frac{2\|x^0 - x^*\|^2}{\beta (k+1)^2}.$$

There are lower bounds that show that this rate is optimal for gradient methods.

Proximal gradients

The objective functions arising in machine learning often have the form

$$f(x) = g(x) + h(x),$$

where both g and h are convex, but only g is differentiable. The term $h(x)$ is typically a regularization term.

Example 22.2. Consider the function

$$f(x) = \frac{1}{2}\|Ax - b\|_2^2 + \lambda \|x\|_1$$

where $\|x\| = \sum_{i=1}^{d} |x_i|$ and $\lambda > 0$. The problem of minimizing this function is often referred to as the LASSO problem in statistics. The purpose of the $\|x\|_1$ term is to promote sparse solutions. These are solutions that have only few entries that are significantly larger than 0.

Definition 22.3. Let f be a convex function. The subdifferential of f at x is the set

$$\partial f(x) = \{g \in \mathbb{R}^d : \forall y \in \mathbb{R}^d, f(x) + g^T(y - x) \leq f(y)\}.$$

The elements of $\partial f(x)$ are called subgradients.

If f is differentiable at x, then there exists only one subgradient, which coincides with the gradient.

Example 22.4. Consider the function $f(x) = |x|$. The differential is then

$$\partial f(x) = \begin{cases} 1 & x > 0 \\ [-1, 1] & x = 0 \\ -1 & x < 0 \end{cases}$$

Example 22.5. The subdifferential is additive, in the sence that if A and B are matrices and $f(x) = g(Ax) + h(Bx)$ then

$$\partial f(x) = A^T(\partial g)(Ax) + B^T(\partial h)(Bx).$$

A special case is the 1-norm. Here, we can write

$$\|x\|_1 = \sum_{i=1}^{d} |\Pi_i(x)|,$$

where $\Pi_i(x) = x_i$ is the projection on the i-th coordinate. It follows that the subdifferential of the 1-norm can be described as

$$\partial \|x\|_1 = \{z : z_i = \operatorname{sign}(x_i) \text{ if } x_i = 0, z_j \in [-1, 1] \text{ if } x_j = 0\}.$$

Using the subdifferential, we have the following optimality condition.

Theorem 22.6. Let $f : \mathbb{R}^d \to \mathbb{R}$ be a convex function. Then x is a global minimizer of f if and only if $0 \in \partial f(x)$.

For composite functions $f(x) = g(x) + h(x)$ with $g \in C^1(\mathbb{R}^d)$, this means that

$$\partial f(x) = \nabla g(x) + \partial h(x).$$

There are different possible strategies for generalizing gradient descent to make it work with non-smooth functions. One would be to simply pick a subgradient at each step and follow that direction. Note that this may not be a descent direction. One common strategy for composite functions is to perform a gradient descent step based on the smooth function g, and then project onto the subgradient of h. Projection onto the subgradient is done via the proximal operator

$$\text{prox}_h(x) = \arg\min_y \frac{1}{2}\|x - y\|^2 + h(y).$$

Note that $x^* = \text{prox}_h(x)$ satisfies $x^* - x \in \partial h(x^*)$. The proximal gradient method for minimizing a function of the form $f(x) = g(x) + h(x)$ starts with a vector x^0, and then for $k \geq 0$ proceeds by computing

$$x^{k+1} = \text{prox}_{\eta h}(x^k - \eta \nabla g(x^k)).$$

Example 22.7. Recall the image inpainting problem from Lecture 13. An image can be viewed as an $m \times n$ matrix U, with each entry u_{ij} corresponding to a light intensity (for greyscale images), or a colour vector, represented by a triple of red, green and blue intensities (usually with values between 0 and 255 each). For simplicity the following discussion assumes a greyscale image. For computational purposes, the matrix of an image is often viewed as an mn-dimensional vector u, with the columns of the matrix stacked on top of each other. In the image inpainting problem, one aims to learn the true value of missing or corrupted entries of an image. There are different approaches to this problem. A conceptually simple approach is to replace the image with the closest image among a set of images satisfying typical properties. But what are typical properties of a typical image? Some properties that come to mind are:

- Images tend to have large homogeneous areas in which the colour doesn't change much;
- Images have approximately low rank, when interpreted as matrices.

Total variation image analysis takes advantage of the first property. The total variation or TV-norm is the sum of the norm of the horizontal and vertical differences,

$$\|U\|_{TV} = \sum_{i=1}^{m}\sum_{j=1}^{n}\sqrt{(u_{i+1,j} - u_{i,j})^2 + (u_{i,j+1} - u_{i,j})^2},$$

where we set entries with out-of-bounds indices to 0.

Now let U be an image with entries u_{ij}, and let $\Omega \subset [m]\times[n] = \{(i,j) \mid 1 \leq i \leq m, 1 \leq j \leq n\}$ be the set of indices where the original image and the corrupted image coincide (all the other entries are missing). One could attempt to find the image with the smallest TV-norm that coincides with the known pixels u_{ij} for $(i, j) \in \Omega$. This is an optimization problem of the form

$$\text{minimize } \|X\|_{TV} \text{ subject to } x_{ij} = u_{ij} \text{ for } (i,j) \in \Omega.$$

Alternatively (see Exercise 8.3), one can solve a regularized problem

$$\text{minimize } \|A(X) - b\|^2 + \lambda \|X\|_{TV},$$

where A represents the linear map projects X onto the entries indexed by Ω. This problem can be solved using proximal gradient methods.

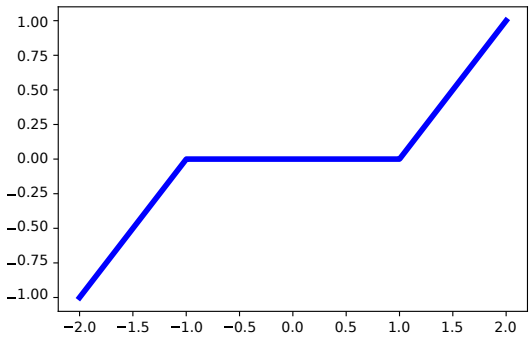


Figure 22.1: Soft thresholding

Even though it seems that a proximal gradient method would require solving an optimization problem within an optimization problem, closed form expressions are known in many cases.

Example 22.8. Consider the function h(|x|)= ||forsome 0.Then

$$\text{prox}_{h}(x) = T_{\lambda}(x) := \begin{cases} x - \lambda & x \geq \lambda \\ 0 & x \in [-\lambda, \lambda] \\ x + \lambda & x \leq -\lambda \end{cases}$$

This is known as soft thresholding (see Figure 22.1).

If a function h has the form

$$h(x) = \sum_{i=1}^{d} h_i(x_i)$$

then the proximal mapping associated to h has the form

$$\text{prox}_{h}(x) = (\text{prox}_{h_1}(x_1), \ldots, \text{prox}_{h_d}(x_d))^T.$$

It follows that if h(x) = λ ||x||1, then we can apply the proximal operator by applying the soft thresholding operator
Tλ to each coordinate of x.

For the proximal gradient method it is possible to obtain similar convergence results as for gradient descent.

Notes

Accelerated gradient descent goes back to Nesterov's work [20]. A more in depth analysis can be found in [19] and [4]. An interesting interpretation of accelerated gradient descent in terms of differential equations is given in [26]. The proximal operator is discussed in detail in Chapter 6 of [1].

23

Stochastic Gradient Descent

In this lecture we introduce Stochastic Gradient Descent (SGD), a probabilistic version of gradient descent that has been around since the 1950s, and that has become popular in the context of data science and machine learning. To motivate the algorithm, consider a set of functions $H=\{h:w\in R^d w\}$, where each such function depends on d parameters. Also consider a smooth loss functions L, or a smooth approximation of a loss function. Given samples $(x, \quad) \quad \subset X \times Y$ $\{ \text{hlii=1 ,definethefunctions}$

$$f_i(w) = L(h_w(x_i), y \quad i$$

$$\in \{1, \ldots, n\}.$$

y The problem of finding functions that minimize the empirical risk is

$$\underset{w}{\text{minimize}} \frac{1}{n} \sum_{i=1}^{n} f_i(w).$$

$$\in R^d$$

The fi are often assumed to be convex and smooth. In addition one often considers a regularization term R(w). In what follows, we abstract from the machine learning context and consider purely the associated optimization problem. Hence, as usual when dealing only with optimization problems, we switch notation and denote the variables to be optimized over by x.

Stochastic Gradient Descent

We consider an objective function of the form

$$f(x) = \frac{1}{n} \sum_{i=1}^{n} f_i(x). \tag{23.1}$$

In what follows we assume that the functions fi are convex and differentiable. If n is large, then computing the gradient can be very expensive. However, and considering the machine learning context, where f(x) is an estimator of the generalization risk $E_\xi[f_\xi(x)]$ of a family of functions f_ξ parametrized by a random vector ξ, we can shift the focus to finding an unbiased estimator of the gradient. Quite trivially, choosing an index j uniformly at random and computing the gradient of fj(x) gives such an unbiased estimator by definition:

$$\sum_{i=1}^{n} f_i(x)$$

$$117$$

$$E_U[f_U(x)] = \frac{1}{n} \sum_{i} \quad),$$

where P {U=j}=1/nforj∈[n]∈b x }.TheStochasticGradientDescent(SGD)algorithmd
proceeds as follows. Begin with R. At each step , compute an unbiased estimator gk of the
gradient at xk:

$$E[g^k \mid x^k] = \nabla f(x^k).$$

Next, take a step in direction g^k:

$$x^{k+1} = x^k - \eta_k g^k,$$

where ηk is a step length (or learning rate in machine learning jargon).

While there are many variants of stochastic gradient descent, we consider the simplest version in which gk is chosen by picking one of the gradients ∇fi(x) uniformly at random, and we refer to this as SGD with uniform sampling. A commonly used generalization is mini-batch sampling, where one chooses a small set of indices I ⊂ {1, . . . , n} at random, instead of only one. We also restrict to the smooth setting without a regularization term; in the non-smooth setting one would apply a proximal operator. Since SGD involves random choices, convergence results are stated in terms of the expected value.LetUbearandomvariablewithdistributionP i
} for∈[n].Then

$$E_U[\nabla f_U(x)] = \sum_{i=1}^n \frac{1}{n} \nabla f_i(x) = \nabla f(x),$$

so that ∇f_U is an unbiased estimator of ∇f. Assuming that has a unique minimizer x^*, we define the empirical variance at the optimal point x^* as

$$\sigma^2 = E_U[\| \nabla f_U(x^*) \|^2] = \frac{1}{n} \sum_{i=1}^n \| \nabla f_i(x^*) \|^2. \tag{23.2}$$

We can now state the convergence result for stochastic gradient descent.

Theorem 23.1. Assume the function f is α-strongly convex and that the fi are convex and β-smooth for
i ∈ [n] and 4β > α. Assume f has a unique minimizer x^* and define the variance as in (23.2). Then for any staring point x0, the sequence of iterates {xk} generated by SGD with uniform sampling and step length η = 1/(2β) satisfies

$$E[\|x^k - x^*\|^2] \le \left(1 - \frac{\alpha}{2\beta}\right)^k \|x^0 - x^*\|^2 + \frac{2\sigma^2}{\alpha\beta}.$$

Proof. As in the analysis of gradient descent, we get

$$\|x^{k+1} - x^*\|^2 = \|x^k - x^*\|^2 - 2\eta \langle \nabla f_U(x^k), x^k - x^* \rangle + \eta^2 \| \nabla f_U(x^k) \|^2.$$

Taking the expectation conditional on x^k we get

$$E[\|x^{k+1} - x^*\|^2 \mid x^k] = \|x^k - x^*\|^2 - 2\eta \langle \nabla f(x^k), x^k - x^* \rangle + \eta^2 E[\| \nabla f_U(x^k) \|^2 \mid x^k], \tag{23.3}$$

where we used the fact that the expectation satisfies $E[\nabla f_U(x)] = \nabla f(x)$. For the last term we use the bound

$$\begin{aligned} E[\| \nabla f_U(x^k) \|^2 \mid x^k] &= E[\| \nabla f_U(x^k) - \nabla f_U(x^*) + \nabla f_U(x^*) \|^2] \\ &\le 2E[\| \nabla f_U(x^k) - \nabla f_U(x^*) \|^2] + 2E[\| \nabla f_U(x^*) \|^2] \\ &\le 2E[\| \nabla f_U(x^k) - \nabla f_U(x^*) \|^2] + 2\sigma^2. \end{aligned}$$

Using the characterization of β smoothness from Lecture 21, namely

$$\frac{1}{\beta}\|\nabla f_i(x) - \nabla f_i(y)\|^2 \leq f_i(x) - f_i(y) - \langle \nabla f_i(y), x - y \rangle, \tag{23.4}$$

we get that

$$\begin{aligned}
\mathbb{E}\left[\|\nabla f_U(x^k) - \nabla f_U(x^*)\|^2 \mid x^k\right] &= \frac{1}{n}\sum_{i=1}^n \|\nabla f_i(x^k) - \nabla f_i(x^*)\|^2 \\
&\overset{(23.4)}{\leq} \frac{2\beta}{n}\sum_{i=1}^n \left(f_i(x^k) - f_i(x^*) - \langle \nabla f_i(x^*), x^k - x^* \rangle\right) \\
&= 2\beta\left(f(x^k) - f(x^*) - \langle \nabla f(x^*), x^k - x^* \rangle\right) \\
&= 2\beta\left(f(x^k) - f(x^*)\right),
\end{aligned}$$

where we used that $\nabla f(x^*) = 0$ for the last equality. Hence, we get the bound

$$\mathbb{E}\left[\|\nabla f_U(x^k)\|^2 \mid x^k\right] \leq 2\beta\left(f(x^k) - f(x^*)\right) + \sigma^2.$$

Plugging this into (23.3), we get

$$\mathbb{E}\left[\|x^{k+1} - x^*\|^2 \mid x^k\right] \leq \|x^k - x^*\|^2 - (2\eta - 2\eta^2\beta)\left(f(x^k) - f(x^*)\right) + \eta^2\sigma^2.$$

Using the α-strong convexity, we get the bound

$$\langle \nabla f(x^k), x^k - x^* \rangle \geq f(x^k) - f(x^*) + \frac{\alpha}{2}\|x^k - x^*\|^2 \geq \frac{\alpha}{2}\|x^k - x^*\|^2,$$

since $f(x^k) \geq f(x^*)$, so that we get

$$\mathbb{E}\left[\|x^{k+1} - x^*\|^2 \mid x^k\right] \leq (1 - \eta\alpha(1 - \eta\beta))\|x^k - x^*\|^2 + 2\eta^2\sigma^2.$$

With step length $\eta = 1/(2\beta)$, and taking the expected value over all previous iterates, we get

$$\mathbb{E}\left[\|x^{k+1} - x^*\|^2\right] \leq \left(1 - \frac{\alpha}{4\beta}\right)\mathbb{E}\left[\|x^k - x^*\|^2\right] + \frac{\sigma^2}{2\beta^2}.$$

Applying this bound recursively (and moving the index down), we get

$$\begin{aligned}
\mathbb{E}\left[\|x^k - x^*\|^2\right] &\leq \left(1 - \frac{\alpha}{4\beta}\right)^k \|x^0 - x^*\|^2 + \frac{\sigma^2}{2\beta^2}\sum_{j=0}^{k-1}\left(1 - \frac{\alpha}{4\beta}\right)^j \\
&\leq \left(1 - \frac{\alpha}{4\beta}\right)^k \|x^0 - x^*\|^2 + \frac{2\sigma^2}{\alpha\beta},
\end{aligned}$$

where we used that $4\beta > \alpha$. □

Example 23.2. Consider the problem of logistic regression, where the aim is to minimize the objective function

$$f(w) = \sum_i n \log(1 + \exp(x_i^T w)) - y_i x_i^T w$$

over a vector of weights w. This problem arises in the context of a binary classification problem with data pairs (x_i, y_i) and $y_i \in \{0, 1\}$. Setting

$$p := \frac{e^{x^T w}}{1 + e^{x^T w}},$$

the resulting classifier is the function

$$h(w) = \begin{cases} 1, & p > 1/2 \\ 0, & p \leq 1/2 \end{cases}$$

The function f is convex (Exercise 7.6(a)), and the gradient is

$$\nabla f(x) = X^T(y - p(w)),$$

where $X \in R^{n \times d}$ is the matrix with the x_i^T as rows, $y = (y_1, \ldots, y_n)^T$, and $p(w) \in R^n$ has coordinates

$$p_i(w) = \frac{\exp(x_i^T w)}{1 + \exp(x_i^T w)}, \quad 1 \leq i \leq n.$$

We can apply different versions of gradient descent to this problem. Figure 1 shows the typical paths of gradient descent and of stochastic gradient descent for a problem with 100 data points. Note that using a naive approach to computing the gradient, one would need to compute 100 gradients at each step. Stochastic gradient descent, on the other hand, fails to converge due to the variance of the gradient estimator (see Figure 2).

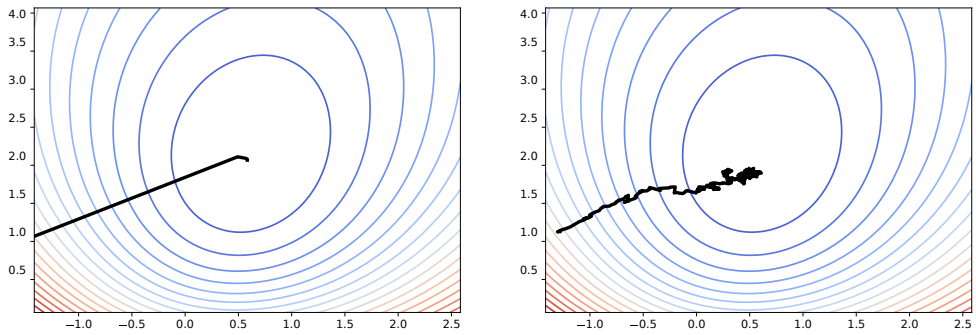


Figure 23.1: The path of gradient descent and of stochastic gradient descent with constant step length.

Extensions

The version of SGD described here is the most basic one. There are many possible extensions to the methods. The include considering different sampling schemes, including mini-batching and importance sampling. These sampling strategies have the effect of reducing the variance σ2. In addition, improvements can be made in the step length selection and when dealing with non-smooth function, where the proximal operator comes into play.

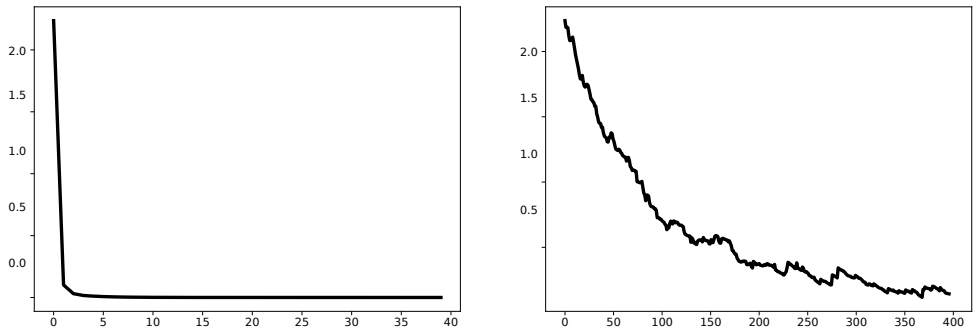


Figure 23.2: Convergence of gradient descent and SGD.

Notes

The origins of stochastic gradient descent go back to the work of Robbins and Monro in 1951 [21]. The algorithm has been rediscovered many times, and gained popularity due to its effectiveness in training deep neural networks on large data sets, where gradient computations are very expensive. Despite its simplicity, a systematic and rigorous analysis has not been available until recently. The presentation in this chapter is based loosely on the papers [11] and [2]. A more general and systematic analysis of SGD that includes non-smooth objectives is given in [10]. These works also discuss general sampling techniques, not just uniform sampling.

Part III

Deep Learning

24

Neural Networks

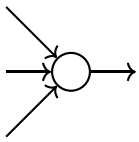
Neural Networks are a powerful class of functions with a wide range of applications in machine learning and data science. Originally introduced as simplified models of neurons in the brain, nowadays the biological motivation plays a less prominent role. Instead, the popularity of neural networks owes to their ability to combine generality with computational tractability: while neural networks can approximate most reasonable functions to arbitrary accuracy, their structure is still simple enough so that they can be trained efficiently by gradient descent.

Connectivity and Activation

We begin by considering the problem of binary classification using a linear function. Given a vector of weights $w \in \mathbb{R}^d$ and a bias term $b \in \mathbb{R}$, define the classifier

$$h_{w,b}(x) = \mathbb{1}\{w^T x + b \geq 0\} = \begin{cases} 1 & w^T x + b \geq 0 \\ 0 & w^T x + b < 0 \end{cases}$$

We already encountered this problem when studying linear support vector machines. Visually, we can represent this classifier by a node that takes d inputs $(x_1, \ldots, x_d)$, and outputs 0 or 1:



Such a unit is called a perceptron. One interpretation is that the node represents a neuron that fires if a certain linear combination of inputs, $w^T x$, exceeds a threshold $-b$. It is sometimes useful to approximate the indicator with a smooth function, and a convenient candidate is the sigmoid

$$\sigma(x) = \frac{1}{1 + e^{-x}}.$$

We can then replace the function $h_{w,b}$ with the smooth function

$$g(x) = \sigma(w^T x + b).$$

A convenient property of the sigmoid function is that the derivative has the form

$$\sigma'(x) = \sigma(x)(1 - \sigma(x)),$$

so that the gradient of g can be computed as

$$\nabla g(x) = (\sigma w^T x + b)(1 - \sigma(w^T x + b))\, w.$$

Other activation functions that are commonly used are the hyperbolic tangent, tanh(x), and the rectifiable linear unit (ReLU), max{x, 0}. Figure 24.1 illustrates these different activations functions.

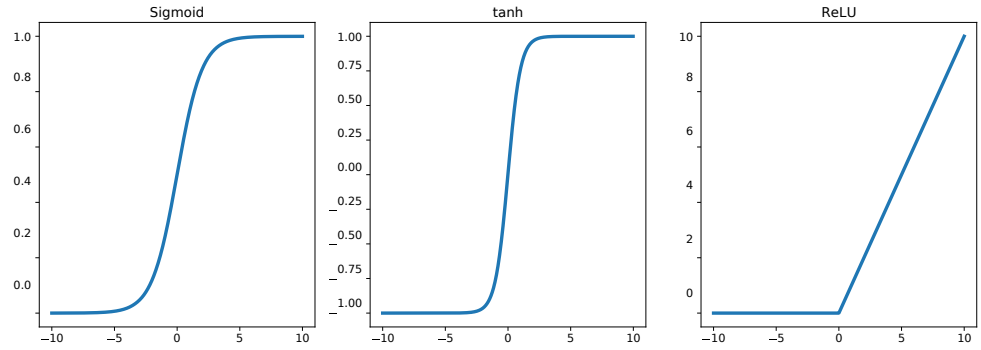


Figure 24.1: Activation functions

A feedforward neural network arises by combining various perceptrons by feeding the outputs of a series of perceptrons into a new perceptrons, see Figure 25.5.

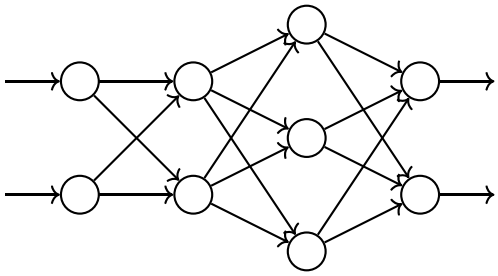


Figure 24.2: A fully connected neural network.

We interpret a neural network as consisting of different layers. To the k-th layer we associated a linear map $W_k : \mathbb{R}^{d_{k-1}} \to \mathbb{R}^{d_k}$ and a bias vector $b_k \in \mathbb{R}^{d_k}$. One then applies the activation function componentwise, to obtain a map

$$\sigma(W_k x + b_k).$$

The first layer is the input layer, while the last layer is the output layer. Hence, a neural network with ℓ layers is a function of the form $F_\ell \circ \cdots \circ F_1(x)$, where the F_k are recursively defined as

$$F_1(x) = \sigma(W_1 x + b_1),$$
$$F_{k+1}(x) = \sigma(W_{k+1}(F_k(x)) + b_{k+1}),$$

and $W_k \in \mathbb{R}^{d_i \times d_{k-1}}, b_k \in \mathbb{R}^{d_k}$ for $1 \le k \le \ell$ (with $d_0 = d$).The layersbetweentheinputandoutput layer are called the hidden layers.

A neural network is therefore just a parametrized function that depends on a possibly large number of parameters. If we fix the architecture, that is, the number of layers ℓ and the number of nodes

d = (d0, d1, . . . , d`), where di represents the number of nodes in each layer, we get a a hypothesis class

$$H=\{F:Rd\rightarrow Rd`\ d:FisaNNwithformatd\}.$$

We can use neural networks for binary classification tasks, for example, by setting d` = 1 (only one output layer), and declaring an output to be of class 1 if F (x) > 1/2 and 0 otherwise. Alternatively, we can have two outputs and classify according to whether the first output is larger than the second. We can train the neural network on data (xi, yi), i $\in$ {1, . . . , n}, using our favourite loss function. Neural networks are particularly attractive for two reasons:

1. The class of neural networks is rich enough to capture almost all functions of interest (see Lecture 25).
2. The structure is simple enough to train using gradient descent, by a computational implementation of the chain rule known as backpropagation.

We discuss the computational aspect, backpropagation, first.

Backpropagation

Denote by W and b the concatenation of all the weight matrices and bias vectors. We denote by $(W_k)_{ij}$ the j)-th entry of the k-th matrix, and by b^k_i the i-th entry of the k-th bias vector. Given n data points $\{x_i\}_{i=1}^n$ with corresponding outputs $\{y_i\}_{i=1}^n$ and a smooth loss function L, the task is to minimize the function

$$f(W,b) = \frac{1}{n}\sum_{i=1}^{n} L(F`(x_i),y_i).$$

Gradiend descent, or stochastic gradient descent, requires evaluating the gradient of a function of the form

$$f_i(W,b) = L(F`(x_i),y_i).$$

We will describe a method for computing the gradient of such a function efficiently. In what follows, set $x=x_i$ and. Also write $y=y_i$ a $=x$, and for $k \in \{1,\ldots,`\}$,

$$z_k = W_k a_{k-1}+b_k, a_k=\sigma(z_k). \quad (24.1)$$

In particular, a` = F`(xi) is the output of the neural network on input x. Moreover, set C = C(W , b) = L(a`, y) for the loss function.

For every layer k and coordinate $j \in \{1,\ldots,d_k\}$, define the sensitivities

$$\delta^k_j := \frac{\partial C}{\partial z^k_j},$$

where z_k is z^k_j δ_{kj} he-th coordinate of. Thus δ^k_j measures the sensitivity of the loss function to the input at the j-th node of the k-th layer. Denote by $\delta_k \in R^{d_k}$ the vector of δ^k_j for $j \in \{1,\ldots,d_k\}$. The partial derivatives of C can be computed in terms of these quantities. In what follows, we denote by $x \circ y$ the componentwise product, that is, the vector with entries $x_i y_i$.

Proposition 24.1. For a neural network with ℓ layers and $k \in \{1, \dots, \ell\}$, we have

$$\frac{\partial C}{\partial w^k_{ij}} = \delta^k_i a^{k-1}_j, \qquad \frac{\partial C}{\partial b^k_i} = \delta^k_i \tag{24.2}$$

for $i, j \in \{1, \dots, d_k\}$. Moreover, the sensitivities δ^k_i can be computed as follows:

$$\delta^\ell = \sigma'(z^\ell) \circ \nabla_a L(a^\ell, y), \qquad \delta^k = \sigma'(z^k) \circ (W^{k+1})^T \delta^{k+1} \tag{24.3}$$

for $k \in \{1, \dots, \ell - 1\}$.

Proof. We begin by showing (24.2). For δ^ℓ_i, note that by the chain rule, we have

$$\delta^\ell_i = \frac{\partial L(a^\ell, y)}{\partial z^\ell_i} = \sum_{j=1}^{d_\ell} \frac{\partial L(a^\ell, y)}{\partial a^\ell_j} \frac{\partial a^\ell_j}{\partial z^\ell_i} = \frac{\partial L(a^\ell, y)}{\partial a^\ell_i} \cdot \sigma'(z^\ell_i).$$

For $k < \ell$, we compute δ^k_i in terms of the δ^{k+1}_j as follows:

$$\delta^k_i = \frac{\partial C}{\partial z^k_i} = \sum_{j=1}^{d_{k+1}} \frac{\partial C}{\partial z^{k+1}_j} \frac{\partial z^{k+1}_j}{\partial z^k_i} = \sum_{j=1}^{d_{k+1}} \delta^{k+1}_j \frac{\partial z^{k+1}_j}{\partial z^k_i}.$$

For the summands in the last expression we use

$$z^{k+1}_j = \sum_{s=1}^{d_k} w^{k+1}_{js} \sigma(z^k_s) + b^{k+1}_j,$$

so that the derivatives evaluate to

$$\frac{\partial z^{k+1}_j}{\partial z^k_i} = w^{k+1}_{ji} \cdot \sigma'(z^k_i).$$

Putting everything together, we arrive at

$$\delta^k_i = \sum_{j=1}^{d_{k+1}} \delta^{k+1}_j \cdot w^{k+1}_{ji} \cdot \sigma'(z^k_i) = \sigma'(z^k_i) \cdot ((W^{k+1})^T \sigma^{k+1})_i.$$

The expressions for the partial derivatives of C with respect to the weights W and the bias b are computed in a straight-forward way using the chain rule. More precisely, at the k-th layer write

$$z^k_i = \sum_{j=1}^{d_{k-1}} w^k_{ij} a^{k-1}_j + b^k_i.$$

The claimed expressions for the derivatives then follow by applying the chain rule,

$$\frac{\partial C}{\partial w^k_{ij}} = \frac{\partial C}{\partial z^k_i} \frac{\partial z^k_i}{\partial w^k_{ij}} = \delta^k_i a^{k-1}_j,$$

and similarly for the derivative with respect to b^k_i. □

For the common quadratic loss function

$$L(a`, y) = \frac{1}{2}\|a` - y\|^2,$$

we get

$$\nabla_{a`} L(a`, y) = a` - y,$$

which can be computed easily from the function value $F(x)$ and y. Other differentiable loss functions may lead to different terms. Given initial weights W and bias terms b, we can compute the values a_k and z_k using a forward pass, that is, by applying (24.1) recursively. We can then compute the sensitivities δ_k using (24.3), and the partial derivatives of the loss function using (24.2), starting at layer $`$. This way of computing the gradients is called backpropagation. We note that choosing the sigmoid δ as activation function, the computation of the derivative $\sigma'(x)$ is easy. The whole process of computing the gradient of a neural network thus reduces to a simple sequence of matrix-vector product operations and sigmoids.

Example 24.2. Consider the following setting with $n = 10$ points. We train a neural network with four layers and dimensions $d = (d_0, d_1, d_2, d_3) = (2, 2, 3, 2)$ using stochastic gradient descent. The neural network outputs a vector in $\mathbb{R}^2$ and classifies an input point according to whether the first coordinate is greater or smaller than the second coordinate. Figure 24.3 shows the decision boundary by the neural network based on 10 training points, and the display on the right shows the error per iteration of stochastic gradient descent.

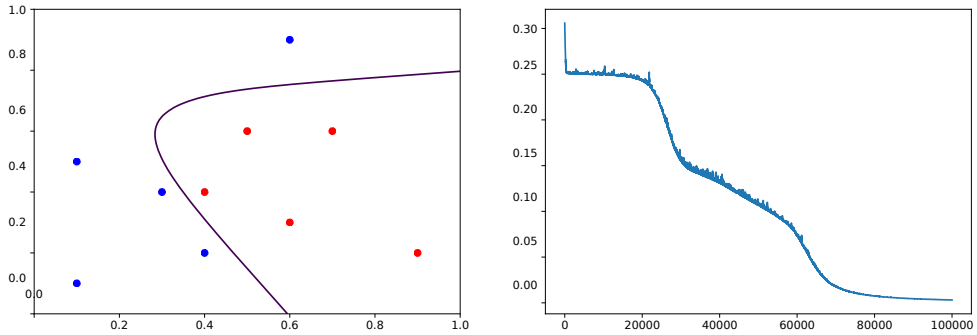


Figure 24.3: Training a neural network for classification and the error of stochastic gradient descent.

Notes

The study of neural networks has its origin in pioneering work by McCulloch and Pitts in the 1940s. Another seminal work in the history of artificial neural networks is the work by Rosenblatt on the Perceptron [22]. The idea of backpropagation was explicitly named in [23] and is closely related to automatic differentiation. Automatic differentiation has been discovered independently many times, and an interesting overview is given here: [12]. The content of this lecture is based on the excellent tutorial [13]. A comprehensive modern treatment of the subject can be found in the book [8].