VIP Cheatsheet: Supervised Learning

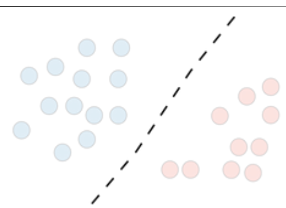
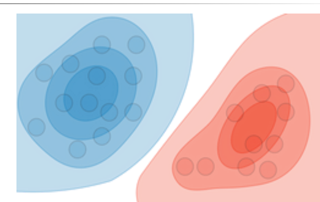
Introduction to Supervised Learning

Given a set of data points $\{x^{(1)},...,x^{(m)}\}$ associated to a set of outcomes $\{y^{(1)},...,y^{(m)}\}$, we want to build a classifier that learns how to predict y from x.

r **Type of prediction** – The different types of predictive models are summed up in the table below:

	Regression	Classifier
Outcome	Continuous	Class
Examples	Linear regression	Logistic regression, SVM, Naive Bayes

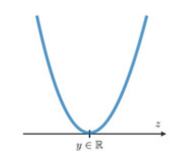
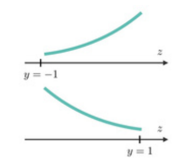
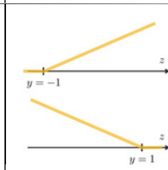
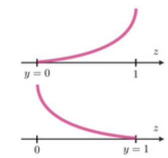
r **Type of model** – The different models are summed up in the table below:

	Discriminative model	Generative model
Goal	Directly estimate $P(y\|x)$	Estimate $P(x\|y)$ to deduce $P(y\|x)$
What's learned	Decision boundary	Probability distributions of the data
Illustration		
Examples	Regressions, SVMs	GDA, Naive Bayes

Notations and general concepts

r **Hypothesis** – The hypothesis is noted h_θ and is the model that we choose. For a given input data $x^{(i)}$, the model prediction output is $h_\theta(x^{(i)})$.

r **Loss function** – A loss function is a function $L:(z,y)\in\mathbb{R}\times Y\longmapsto L(z,y)\in\mathbb{R}$ that takes as inputs the predicted value z corresponding to the real data value y and outputs how different they are. The common loss functions are summed up in the table below:

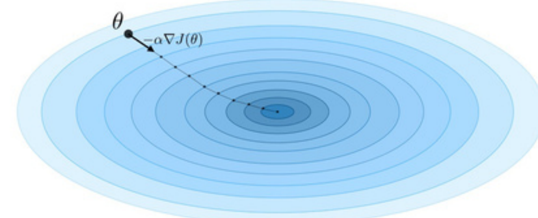
Least squared	Logistic	Hinge	Cross-entropy
$\frac{1}{2}(y-z)^2$	$\log(1+\exp(-yz))$	$\max(0,1-yz)$	$-\big[y\log(z)+(1-y)\log(1-z)\big]$
			
Linear regression	Logistic regression	SVM	Neural Network

r **Cost function** – The cost function J is commonly used to assess the performance of a model, and is defined with the loss function L as follows:

$$J(\theta)=\sum_{i=1}^{m}L(h_\theta(x^{(i)}),y^{(i)})$$

r **Gradient descent** – By noting $\alpha\in\mathbb{R}$ the learning rate, the update rule for gradient descent is expressed with the learning rate and the cost function J as follows:

$$\theta\longleftarrow\theta-\alpha\nabla J(\theta)$$



Remark: Stochastic gradient descent (SGD) is updating the parameter based on each training example, and batch gradient descent is on a batch of training examples.

r **Likelihood** – The likelihood of a model $L(\theta)$ given parameters θ is used to find the optimal parameters θ through maximizing the likelihood. In practice, we use the log-likelihood $\ell(\theta)=\log(L(\theta))$ which is easier to optimize. We have:

$$\theta^{\textrm{opt}}=\underset{\theta}{\textrm{arg max }}L(\theta)$$

r **Newton's algorithm** – The Newton's algorithm is a numerical method that finds θ such that $\ell'(\theta)=0$. Its update rule is as follows:

$$\theta\leftarrow\theta-\frac{\ell'(\theta)}{\ell''(\theta)}$$

Remark: the multidimensional generalization, also known as the Newton-Raphson method, has the following update rule:

$$\theta\leftarrow\theta-\left(\nabla_\theta^2\ell(\theta)\right)^{-1}\nabla_\theta\ell(\theta)$$

Linear regression

We assume here that y |x; θ ∼ N (μ,σ2)

r **Normal equations** – By noting X the matrix design, the value of θ that minimizes the cost function is a closed-form solution such that:

θ = (XTX)−1XTy

r **LMS algorithm** – By noting α the learning rate, the update rule of the Least Mean Squares (LMS) algorithm for a training set of m data points, which is also known as the Widrow-Hoff learning rule, is as follows:

m

∀j,θ←θ+αy(i)−h(x(i))(i)jj θxj

i=1

∑ [ ]

Remark: the update rule is a particular case of the gradient ascent.

r **LWR** – Locally Weighted Regression, also known as LWR, is a variant of linear regression that weights each training example in its cost function by w(i)(x), which is defined with parameter τ ∈ R as:

x(i)

−x2w(i)(x)=exp()

− 2 τ 2

Classification and logistic regression

r **Sigmoid function** – The sigmoid function g, also known as the logistic function, is defined as follows:

∀z∈R, g(z)= 1]01[1+e− ∈ ,z

r **Logistic regression** – We assume here that y |x;θ Bernoulli(φ). We have the following form:

φ=p(y=1

|x;θ)= 1 =g(θTx)1+exp(

−θT x)

Remark: there is no closed form solution for the case of logistic regressions.

r **Softmax regression** – A softmax regression, also called a multiclass logistic regression, is used to generalize logistic regression when there are more than 2 outcome classes. By convention, we set θK = 0, which makes the Bernoulli parameter φi of each class i equal to:

exp(θ x)

φi

∑T=iKexp(θTjx)j=1

Generalized Linear Models

r **Exponential family** – A class of distributions is said to be in the exponential family if it can be written in terms of a natural parameter, also called the canonical parameter or link function, η, a sufficient statistic T(y) and a log-partition function a(η) as follows:

p(y; η) = b(y) exp(ηT (y)

− a(η))

Remark: we will often have T(y) = y. *Also,* exp(

−a(η)) *can be seen as a normalization parameter that will make sure that the probabilities sum to one.*

Here are the most common exponential distributions summed up in the following table:

Distribution

(η)T(y)a(η)b(y)Bernoullilogφ1−ylog(1+exp(η))1φ

()Gaussianμyη2122√exp−y2π2

Poissonlog(1λ)y

(eη!Geog(1φ)yeη)ymetriclo−log1−1eη

r **Assumptions of GLMs** – Generalized Linear Models (GLM) aim at predicting a random variable y as a function fo x ∈ Rn+1 and rely on the following 3 assumptions:

(1)y

|x;θ ExpFamily(η)(2)h(x)=[;](3)=TθEy|xθηθx

Remark: ordinary least squares and logistic regression are special cases of generalized linear models.

Support Vector Machines

The goal of support vector machines is to find the line that maximizes the minimum distance to the line.

r **Optimal margin classifier** – The optimal margin classifier h is such that:

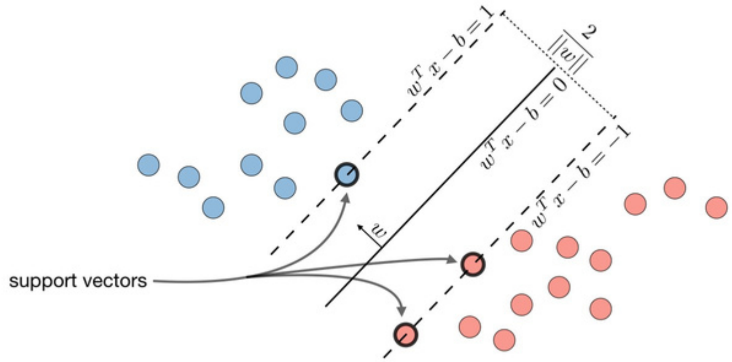
h(x) = sign(wT x

− b)

where (w,b) ∈ Rn × R is the solution of the following optimization problem:

min1

||w||2suchthaty(i)(wTx(i)−b)>12



Remark: the line is defined as wTx − b = 0 .

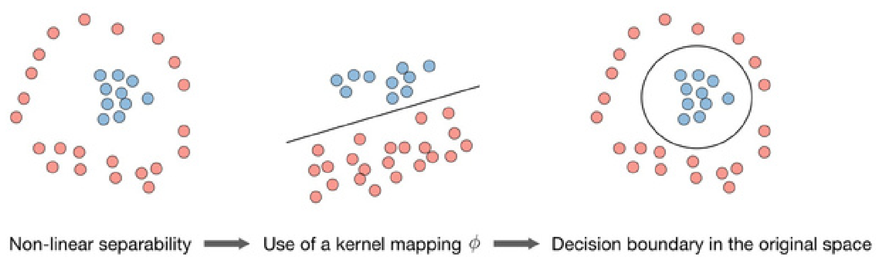
r **Hinge loss** – The hinge loss is used in the setting of SVMs and is defined as follows:

L(z,y) = [1 − yz]+ = max(0,1 − yz)

r **Kernel** – Given a feature mapping φ, we define the kernel K to be defined as:

K(x,z) = φ(x)T

(φ(z))finedby()=exp−||x−z||2Inpractice,thekernelKdeKx,z2σ2iscalledtheGaussiankernelandiscommonlyused.



Remark: we say that we use the "kernel trick" to compute the cost function using the kernel because we actually don't need to know the explicit mapping φ, which is often very complicated. Instead, only the values K(x,z) are needed.

r **Lagrangian** – We define the Lagrangian

L(w,b)

∑asfollows:lL(w,b)=f(w)+βihi(w)i=1

Remark: the coefficients βi are called the Lagrange multipliers.

Generative Learning

A generative model first tries to learn how the data is generated by estimating P(x|y), which we can then use to estimate P(y|x) by using Bayes' rule.

Gaussian Discriminant Analysis

r **Setting** – The Gaussian Discriminant Analysis assumes that y and x|y = 0 and x|y = 1 are such that:

y Bernoulli(φ)

x|y=0 N(μ0,Σ) and x|y=1 N(μ1,Σ)

r **Estimation** – The following table sums up the estimates that we find when maximizing the likelihood:

φ μĵ(j=0,1) Σ̂

m m

1 1x(i)m 1 y = 1

∑={(i)j}(x(i)μ)(x(i){y(i)=1}m−y(i)−μT(i))m1(i)myi=1i=1{y=j}i=1

Naive Bayes

r **Assumption** – The Naive Bayes model supposes that the features of each data point are all indep endent:

n

P(x|y)=P(x1,x2,...|y)=P(x1|y)P(x2|y)...= P(xi|y)

i=1

r **Solutions** – Maximizing the log-likelihood gives the following solutions, with k ∈ {0,1}, l ∈ [[1,L]]

P(y=k)=1 (j)(j)

x{(i) #{j|y=kandxi=l}#j|y=k}andP(xi=l|y=k)=m #{j|y(j) = k}

Remark: Naive Bayes is widely used for text classification and spam detection.

Tree-based and ensemble methods

These methods can be used for both regression and classification problems.

r **CART** – Classification and Regression Trees (CART), commonly known as decision trees, can be represented as binary trees. They have the advantage to be very interpretable.

r **Random forest** – It is a tree-based technique that uses a high number of decision trees built out of randomly selected sets of features. Contrary to the simple decision tree, it is highly uninterpretable but its generally good performance makes it a popular algorithm.

Remark: random forests are a type of ensemble methods.

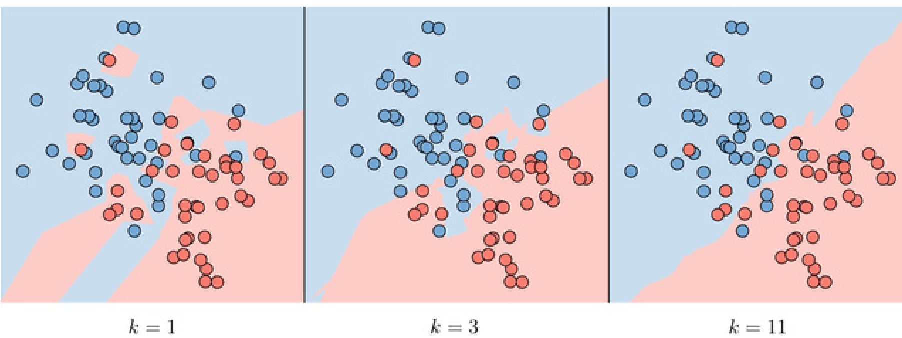
r **Boosting** – The idea of boosting methods is to combine several weak learners to form a stronger one. The main ones are summed up in the table below:

Adaptive boosting	Gradient boosting
- High weights are put on errors to improve at the next boosting step - Known as Adaboost	- Weak learners trained on remaining errors

Other non-parametric approaches

r **k-nearest neighbors** – The k-nearest neighbors algorithm, commonly known as k-NN, is a non-parametric approach where the response of a data point is determined by the nature of its k neighbors from the training set. It can be used in both classification and regression settings.

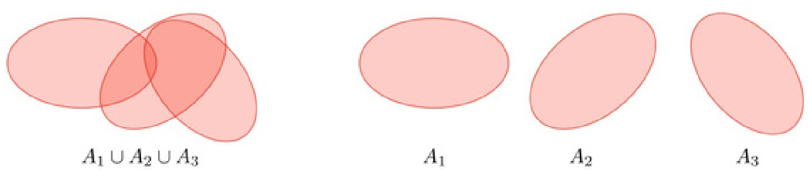
Remark: The higher the parameter k, the higher the bias, and the lower the parameter k, the higher the variance.



Learning Theory

r **Union bound** – Let $A_1, ..., A_k$ be k events. We have:

$$P(A_1\cup...\cup A_k)\leqslant P(A_1)+...+P(A_k)$$



r **Hoeffding inequality** – Let $Z_1,..,Z_m$ be m iid variables drawn from a Bernoulli distribution of parameter ϕ. Let $\widehat{\phi}$ be their sample mean and $\gamma>0$ fixed. We have:

$$P(|\phi-\widehat{\phi}|>\gamma)\leqslant 2\exp(-2\gamma^2m)$$

Remark: this inequality is also known as the Chernoff bound.

r **Training error** – For a given classifier h, we define the training error $\widehat{\epsilon}(h)$, also known as the empirical risk or empirical error, to be as follows:

$$\widehat{\epsilon}(h)=\frac{1}{m}\sum_{i=1}^m1_{\{h(x^{(i)})\neq y^{(i)}\}}$$

r **Probably Approximately Correct (PAC)** – PAC is a framework under which numerous results on learning theory were proved, and has the following set of assumptions:

- the training and testing sets follow the same distribution
- the training examples are drawn independently

r **Shattering** – Given a set $S=\{x^{(1)},...,x^{(d)}\}$, and a set of classifiers $\mathcal{H}$, we say that $\mathcal{H}$ shatters S if for any set of labels $\{y^{(1)},...,y^{(d)}\}$, we have:

$$\exists h\in\mathcal{H}, \quad \forall i\in[\![1,d]\!],\quad h(x^{(i)})=y^{(i)}$$

r **Upper bound theorem** – Let $\mathcal{H}$ be a finite hypothesis class such that $|\mathcal{H}|=k$ and let δ and the sample size m be fixed. Then, with probability of at least $1-\delta$, we have:

$$\epsilon(\widehat{h})\leqslant\left(\min_{h\in\mathcal{H}}\epsilon(h)\right)+2\sqrt{\frac{1}{2m}\log\left(\frac{2k}{\delta}\right)}$$

r **VC dimension** – The Vapnik-Chervonenkis (VC) dimension of a given infinite hypothesis class $\mathcal{H}$, noted $\textrm{VC}(\mathcal{H})$ is the size of the largest set that is shattered by $\mathcal{H}$.

Remark: the VC dimension of $\mathcal{H}$ = {set of linear classifiers in 2 dimensions} is 3.

r **Theorem (Vapnik)** – Let $\mathcal{H}$ be given, with $\textrm{VC}(\mathcal{H})=d$ and m the number of training examples. With probability at least $1-\delta$, we have:

$$\epsilon(\widehat{h})\leqslant\left(\min_{h\in\mathcal{H}}\epsilon(h)\right)+O\left(\sqrt{\frac{d}{m}\log\left(\frac{m}{d}\right)+\frac{1}{m}\log\left(\frac{1}{\delta}\right)}\right)$$