

Probability Theory for Machine Learning



Outline

- Motivation
- Probability Definitions and Rules
- Probability Distributions
- MLE for Gaussian Parameter Estimation
- MLE and Least Squares

Material

- Pattern Recognition and Machine Learning - Christopher M. Bishop
- All of Statistics – Larry Wasserman
- Wolfram MathWorld
- Wikipedia

Motivation

- Uncertainty arises through:
 - Noisy measurements
 - Finite size of data sets
 - Ambiguity: The word bank can mean (1) a financial institution, (2) the side of a river, or (3) tilting an airplane. Which meaning was intended, based on the words that appear nearby?
 - Limited Model Complexity
- Probability theory provides a consistent framework for the quantification and manipulation of uncertainty
- Allows us to make optimal predictions given all the information available to us, even though that information may be incomplete or ambiguous

Sample Space

- The sample space Ω is the set of possible outcomes of an experiment. Points ω in Ω are called sample outcomes, realizations, or elements. Subsets of Ω are called Events.
- Example. If we toss a coin twice then $\Omega = \{HH, HT, TH, TT\}$. The event that the first toss is heads is $A = \{HH, HT\}$
- We say that events A_1 and A_2 are disjoint (mutually exclusive) if $A_i \cap A_j = \{\}$
 - Example: first flip being heads and first flip being tails

Probability

- We will assign a real number $P(A)$ to every event A , called the probability of A .
- To qualify as a probability, P must satisfy three axioms:
 - Axiom 1: $P(A) \geq 0$ for every A
 - Axiom 2: $P(\Omega) = 1$
 - Axiom 3: If $A_1, A_2, \dots$ are disjoint then

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)$$

Joint and Conditional Probabilities

- Joint Probability
 - $P(X,Y)$
 - Probability of X and Y
- Conditional Probability
 - $P(X|Y)$
 - Probability of X given Y

Independent and Conditional Probabilities

- Assuming that $P(B) > 0$, the **conditional** probability of A given B:
- $P(A|B) = P(AB)/P(B)$
- $P(AB) = P(A|B)P(B) = P(B|A)P(A)$
 - Product Rule
- Two events A and B are **independent** if
- $P(AB) = P(A)P(B)$
 - Joint = Product of Marginals
- Two events A and B are **conditionally independent** given C if they are independent after conditioning on C
- $P(AB|C) = P(B|AC)P(A|C) = P(B|C)P(A|C)$

Example

- 60% of ML students pass the final and 45% of ML students pass both the final and the midterm *
- What percent of students who passed the final also passed the midterm?

* These are made up values.

Example

- 60% of ML students pass the final and 45% of ML students pass both the final and the midterm *
- What percent of students who passed the final also passed the midterm?
- Reworded: What percent of students passed the midterm given they passed the final?
- $P(M|F) = P(M,F) / P(F)$
- $= .45 / .60$
- $= .75$

* These are made up values.

Marginalization and Law of Total Probability

- Marginalization (Sum Rule)

$$p(x) = \sum_y p(x, y)$$

- Law of Total Probability

$$p(x) = \sum_y p(x \mid y) \cdot p(y)$$

Bayes' Rule

$$P(A|B) = P(AB) / P(B) \quad \text{(Conditional Probability)}$$

$$P(A|B) = P(B|A)P(A) / P(B) \quad \text{(Product Rule)}$$

$$P(A|B) = P(B|A)P(A) / \sum P(B|A)P(A) \quad \text{(Law of Total Probability)}$$

$$P(A|B) = \frac{P(A) P(B|A)}{P(B)}$$

$$P(B) = \sum_j P(B | A_j) P(A_j)$$

Bayes' Rule

$$P(A|B) = \frac{P(A) P(B|A)}{P(B)}$$

$$p(\theta|x) = \frac{p(x|\theta)p(\theta)}{p(x)}.$$

$$\text{Posterior} = \frac{\text{Likelihood} * \text{Prior}}{\text{Evidence}}$$

Posterior probability $\propto$ Likelihood $\times$ Prior probability

Example

- Suppose you have tested positive for a disease; what is the probability that you actually have the disease?
- It depends on the accuracy and sensitivity of the test, and on the background (prior) probability of the disease.
- $P(T=1 | D=1) = .95$ (true positive)
- $P(T=1 | D=0) = .10$ (false positive)
- $P(D=1) = .01$ (prior)
- $P(D=1 | T=1) = ?$

Example

- $P(T=1 | D=1) = .95$ (true positive)
- $P(T=1 | D=0) = .10$ (false positive)
- $P(D=1) = .01$ (prior)

Bayes' Rule

$$\begin{aligned} \bullet P(D|T) &= P(T|D)P(D) / P(T) \\ &= .95 * .01 / .1085 \\ &= .087 \end{aligned}$$

Law of Total Probability

$$\begin{aligned} \bullet P(T) &= \sum P(T|D)P(D) \\ &= P(T|D=1)P(D=1) + P(T|D=0)P(D=0) \\ &= .95*.01 + .1*.99 \\ &= .1085 \end{aligned}$$

The probability that you have the disease given you tested positive is 8.7%

Random Variable

- How do we link sample spaces and events to data?
- A random variable is a mapping that assigns a real number $X(\omega)$ to each outcome ω
- Example: Flip a coin ten times. Let $X(\omega)$ be the number of heads in the sequence ω . If $\omega = \text{HHTHHTHHTT}$, then $X(\omega) = 6$.

Discrete vs Continuous Random Variables

- Discrete: can only take a countable number of values
- Example: number of heads
- Distribution defined by probability mass function (pmf)
- Marginalization: $p(x) = \sum_y p(x, y)$
- Continuous: can take infinitely many values (real numbers)
- Example: time taken to accomplish task
- Distribution defined by probability density function (pdf)
- Marginalization:

$$p(x) = \int_y p(x, y) dy$$

Probability Distribution Statistics

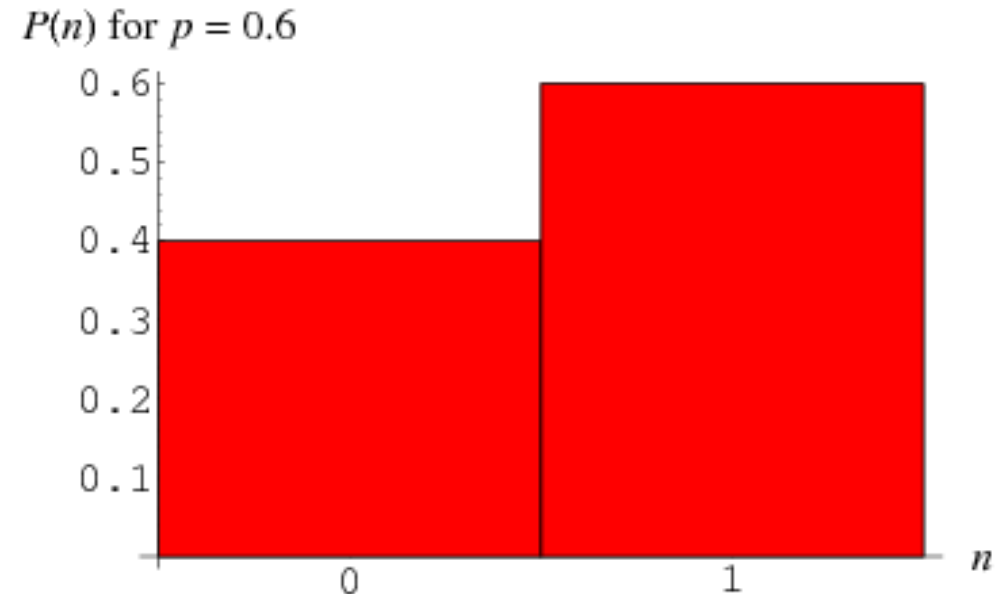
- Mean: $E[x] = \mu = \text{first moment}$
$$= \int_{-\infty}^{\infty} x f(x) dx$$
 Univariate continuous random variable
$$= \sum_{i=1}^{\infty} x_i p_i$$
 Univariate discrete random variable
- Variance: $\text{Var}(X) = E[(X - \mu)^2]$
$$= E[(X - E[X])^2]$$
$$= E[X^2 - 2X E[X] + (E[X])^2]$$
$$= E[X^2] - 2E[X]E[X] + (E[X])^2$$
$$= E[X^2] - (E[X])^2$$
- Nth moment =
$$\int_{-\infty}^{\infty} (x - c)^n f(x) dx$$

Bernoulli Distribution

- Input: $x \in \{0, 1\}$
- Parameter: μ
- Example: Probability of flipping heads ($x=1$)

$$\text{Bern}(x|\mu) = \mu^x (1 - \mu)^{1-x}$$

- Mean = $E[x] = \mu$
- Variance = $\mu(1 - \mu)$



Binomial Distribution

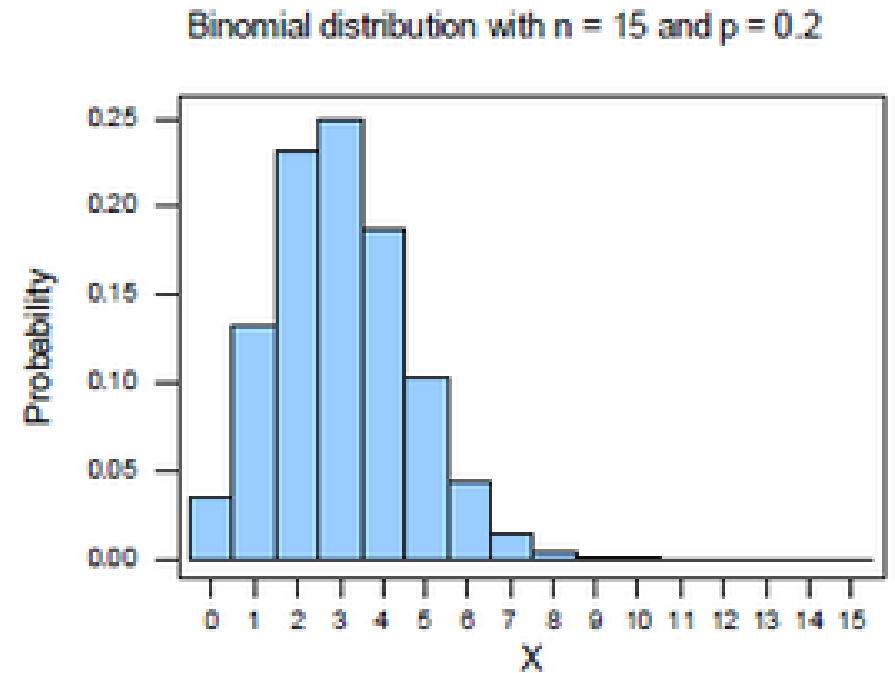
- Input: m = number of successes
- Parameters: N = number of trials

μ = probability of success

- Example: Probability of flipping heads m times out of N independent flips with success probability μ

$$\text{Bin}(m|N, \mu) = \binom{N}{m} \mu^m (1 - \mu)^{N-m}$$

- Mean = $E[x] = N\mu$
- Variance = $N\mu(1 - \mu)$



Multinomial Distribution

- The multinomial distribution is a generalization of the binomial distribution to k categories instead of just binary (success/fail)
- For n independent trials each of which leads to a success for exactly one of k categories, the multinomial distribution gives the probability of any particular combination of numbers of successes for the various categories
- Example: Rolling a die N times

Multinomial Distribution

- Input: $m_1 \dots m_K$ (counts)
- Parameters: N = number of trials
 $\boldsymbol{\mu} = \mu_1 \dots \mu_K$ probability of success for each category, $\sum \boldsymbol{\mu} = 1$

$$\text{Mult}(m_1, m_2, \dots, m_K | \boldsymbol{\mu}, N) = \binom{N}{m_1 m_2 \dots m_K} \prod_{k=1}^K \mu_k^{m_k}$$

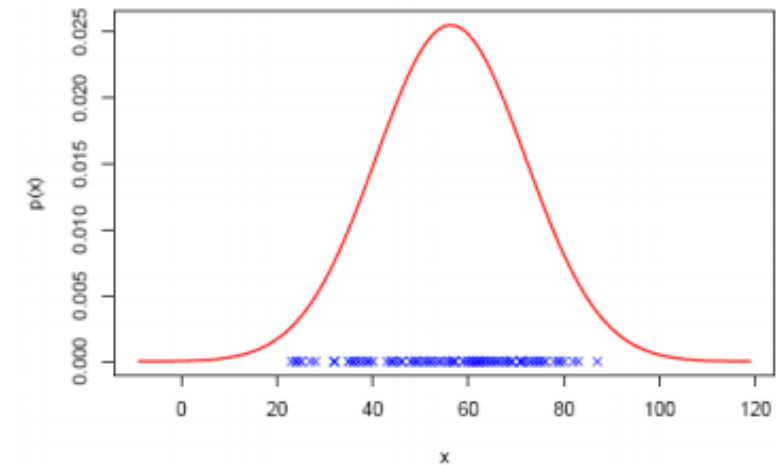
- Mean of m_k : $N\mu_k$
- Variance of m_k : $N\mu_k(1-\mu_k)$

Gaussian Distribution

- Aka the normal distribution
- Widely used model for the distribution of continuous variables
- In the case of a single variable x , the Gaussian distribution can be written in the form

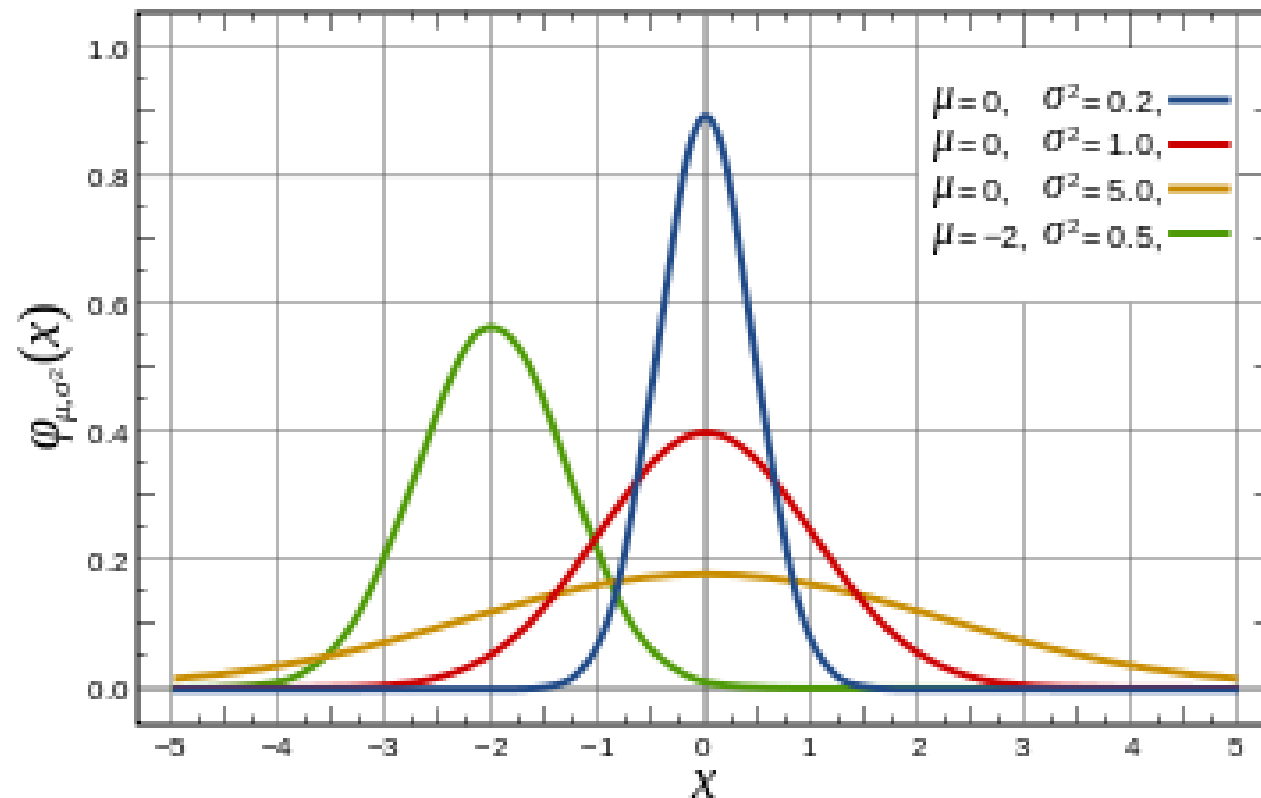
$$\mathcal{N}(x|\mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp \left\{ -\frac{1}{2\sigma^2} (x - \mu)^2 \right\}$$

- where μ is the mean and σ^2 is the variance



Gaussian Distribution

- Gaussians with different means and variances

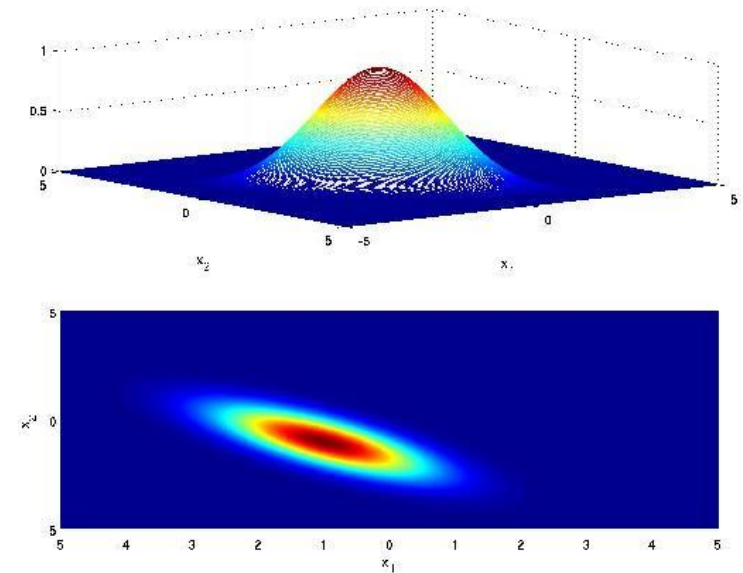


Multivariate Gaussian Distribution

- For a D-dimensional vector $\mathbf{x}$, the multivariate Gaussian distribution takes the form

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \exp \left\{ -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right\}$$

- where $\boldsymbol{\mu}$ is a D-dimensional mean vector
- $\boldsymbol{\Sigma}$ is a $D \times D$ covariance matrix
- $|\boldsymbol{\Sigma}|$ denotes the determinant of $\boldsymbol{\Sigma}$



Inferring Parameters

- We have data X and we assume it comes from some distribution
- How do we figure out the parameters that ‘best’ fit that distribution?
 - Maximum Likelihood Estimation (MLE)

$$\hat{\pi}_{MLE} = \underset{\pi}{\operatorname{argmax}} P(X|\pi)$$

- Maximum a Posteriori (MAP)

$$\hat{\pi}_{MAP} = \underset{\pi}{\operatorname{argmax}} P(\pi|X)$$

See ‘Gibbs Sampling for the Uninitiated’ for a straightforward introduction to parameter estimation: <http://www.umiacs.umd.edu/~resnik/pubs/LAMP-TR-153.pdf>

I.I.D.

- Random variables are independent and identically distributed (i.i.d.) if they have the same probability distribution as the others and are all mutually independent.
- Example: Coin flips are assumed to be IID

MLE for parameter estimation

- The parameters of a Gaussian distribution are the mean (μ) and variance (σ^2)

$$\mathcal{N}(x|\mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp \left\{ -\frac{1}{2\sigma^2} (x - \mu)^2 \right\}$$

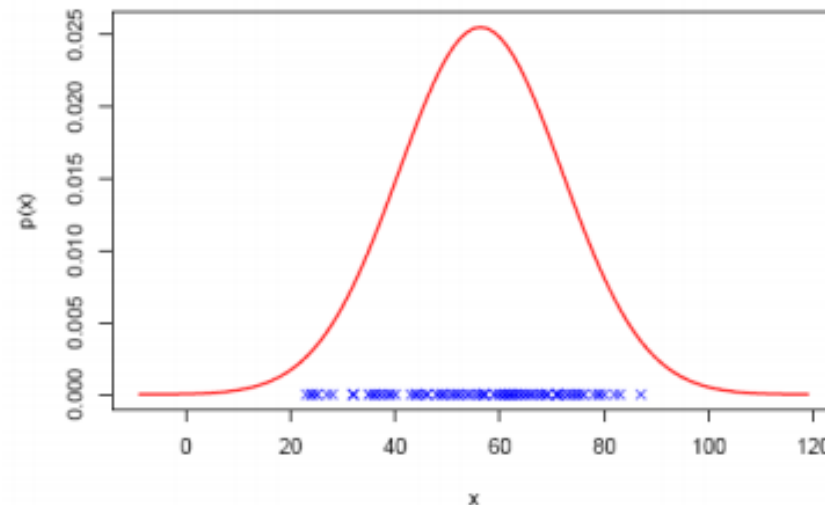
- We'll estimate the parameters using MLE
- Given observations $x_1, \dots, x_N$, the likelihood of those observations for a certain μ and σ^2 (assuming IID) is

Likelihood =
$$p(x_1, \dots, x_N | \mu, \sigma^2) = \prod_{n=1}^N \frac{1}{\sqrt{2\pi}\sigma} \exp \left\{ \frac{-(x_n - \mu)^2}{2\sigma^2} \right\}$$

MLE for parameter estimation

$$\mathcal{N}(x|\mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp \left\{ -\frac{1}{2\sigma^2} (x - \mu)^2 \right\}$$

Likelihood = $p(x_1, \dots, x_N | \mu, \sigma^2) = \prod_{n=1}^N \frac{1}{\sqrt{2\pi}\sigma} \exp \left\{ \frac{-(x_n - \mu)^2}{2\sigma^2} \right\}$



What's the distribution's mean and variance?

MLE for Gaussian Parameters

$$\text{Likelihood} = p(x_1, \dots, x_N | \mu, \sigma^2) = \prod_{n=1}^N \frac{1}{\sqrt{2\pi}\sigma} \exp \left\{ \frac{-(x_n - \mu)^2}{2\sigma^2} \right\}$$

- Now we want to maximize this function wrt μ
- Instead of maximizing the product, we take the log of the likelihood so the product becomes a sum

$$\text{Log Likelihood} = \log p(x_1, \dots, x_N | \mu, \sigma^2) = \sum_{n=1}^N \text{Log} \frac{1}{\sqrt{2\pi}\sigma} \exp \left\{ \frac{-(x_n - \mu)^2}{2\sigma^2} \right\}$$

- We can do this because log is monotonically increasing
- Meaning

$$\max L(\theta) = \max \log L(\theta)$$

MLE for Gaussian Parameters

- Log Likelihood simplifies to:

$$\mathcal{L}(\mu, \sigma) = -\frac{1}{2}N \log(2\pi\sigma^2) - \sum_{n=1}^N \frac{(x_n - \mu)^2}{2\sigma^2}$$

- Now we want to maximize this function wrt μ
- How?

MLE for Gaussian Parameters

- Log Likelihood simplifies to:

$$\mathcal{L}(\mu, \sigma) = -\frac{1}{2}N \log(2\pi\sigma^2) - \sum_{n=1}^N \frac{(x_n - \mu)^2}{2\sigma^2}$$

- Now we want to maximize this function wrt μ
- Take the derivative, set to 0, solve for μ

$$\hat{\mu} = \frac{1}{N} \sum_{n=1}^N x_n$$

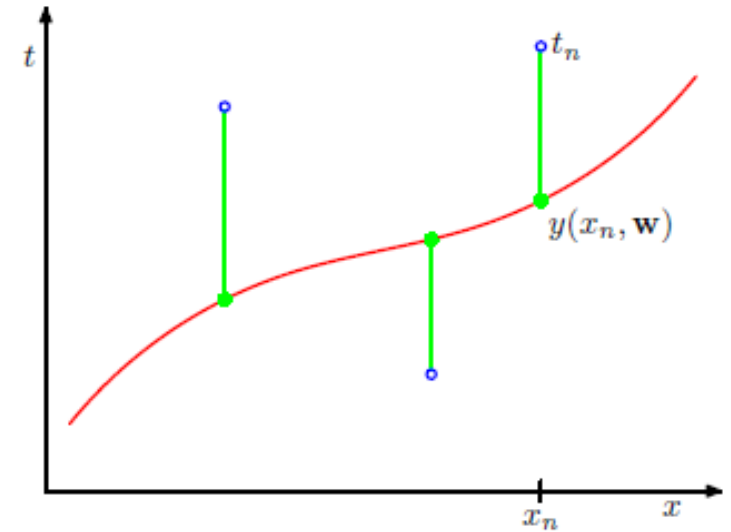
$$\hat{\sigma}^2 = \frac{1}{N} \sum_{n=1}^N (x_n - \hat{\mu})^2$$

To see proofs for these derivations: http://www.statlect.com/normal_distribution_maximum_likelihood.htm

Maximum Likelihood and Least Squares

- Suppose that you are presented with a sequence of data points $(X_1, T_1), \dots, (X_n, T_n)$, and you are asked to find the “best fit” line passing through those points.
- In order to answer this you need to know precisely how to tell whether one line is “fitter” than another
- A common measure of fitness is the squared-

error
$$\sum_{n=1}^N [t^{(n)} - y^{(n)}]^2$$

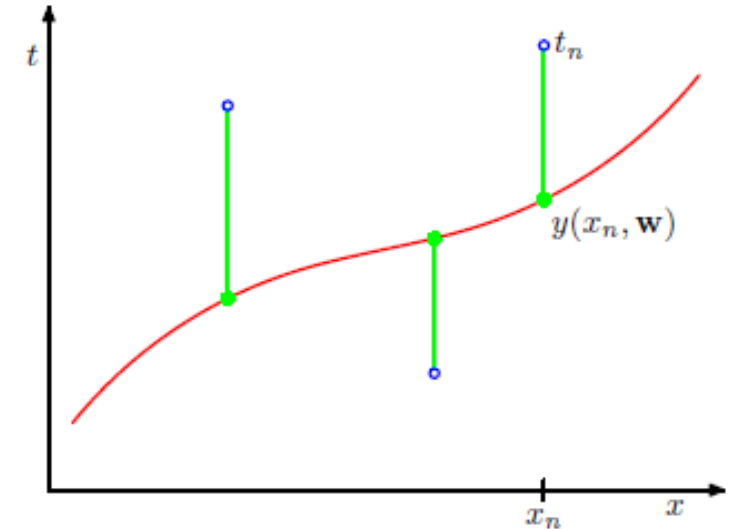


For a good discussion of Maximum likelihood estimators and least squares see http://people.math.gatech.edu/~ecroot/3225/maximum_likelihood.pdf

Maximum Likelihood and Least Squares

$y(x, \mathbf{w})$ is estimating the target t

Red line
$$y(x, \mathbf{w}) = w_0 + w_1x + w_2x^2 + \dots + w_Mx^M = \sum_{j=0}^M w_jx^j$$



- Error/Loss/Cost/Objective function measures the squared error

Green lines
$$\ell(\mathbf{w}) = \sum_{n=1}^N [t^{(n)} - y^{(n)}]^2$$

- Least Square Regression
 - Minimize $L(\mathbf{w})$ wrt $\mathbf{w}$

Maximum Likelihood and Least Squares

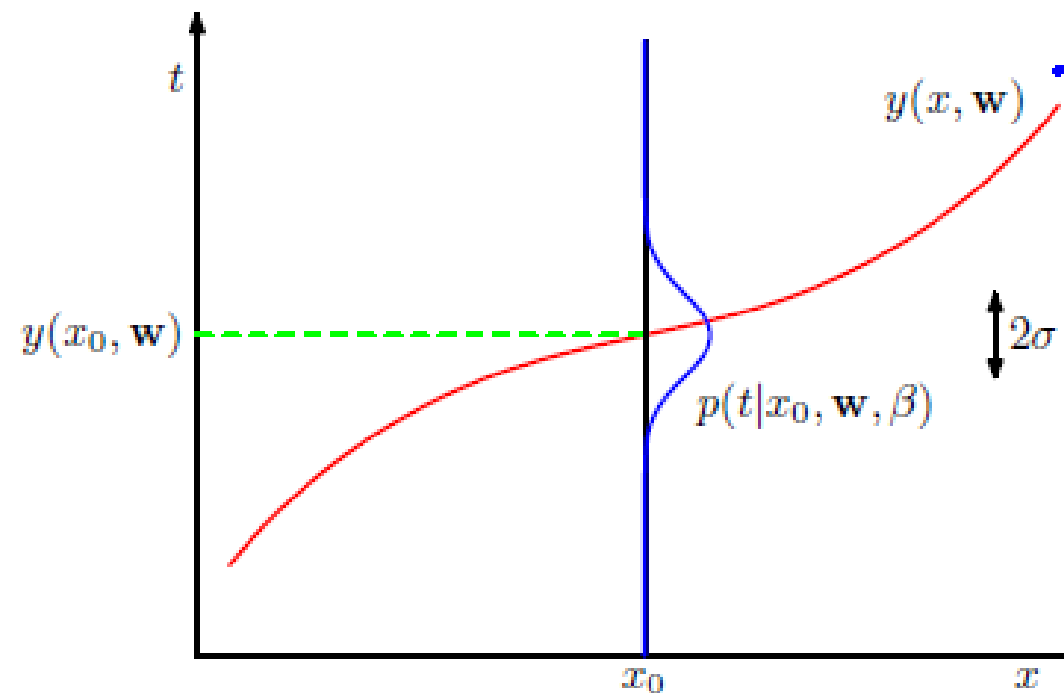
- Now we approach curve fitting from a probabilistic perspective
- We can express our uncertainty over the value of the target variable using a probability distribution
- We assume, given the value of x , the corresponding value of t has a Gaussian distribution with a mean equal to the value $y(x, \mathbf{w})$

$$p(t|x, \mathbf{w}, \beta) = \mathcal{N}(t|y(x, \mathbf{w}), \beta^{-1})$$

β is the precision parameter (inverse variance)

Maximum Likelihood and Least Squares

$$p(t|x, \mathbf{w}, \beta) = \mathcal{N}(t|y(x, \mathbf{w}), \beta^{-1})$$



Maximum Likelihood and Least Squares

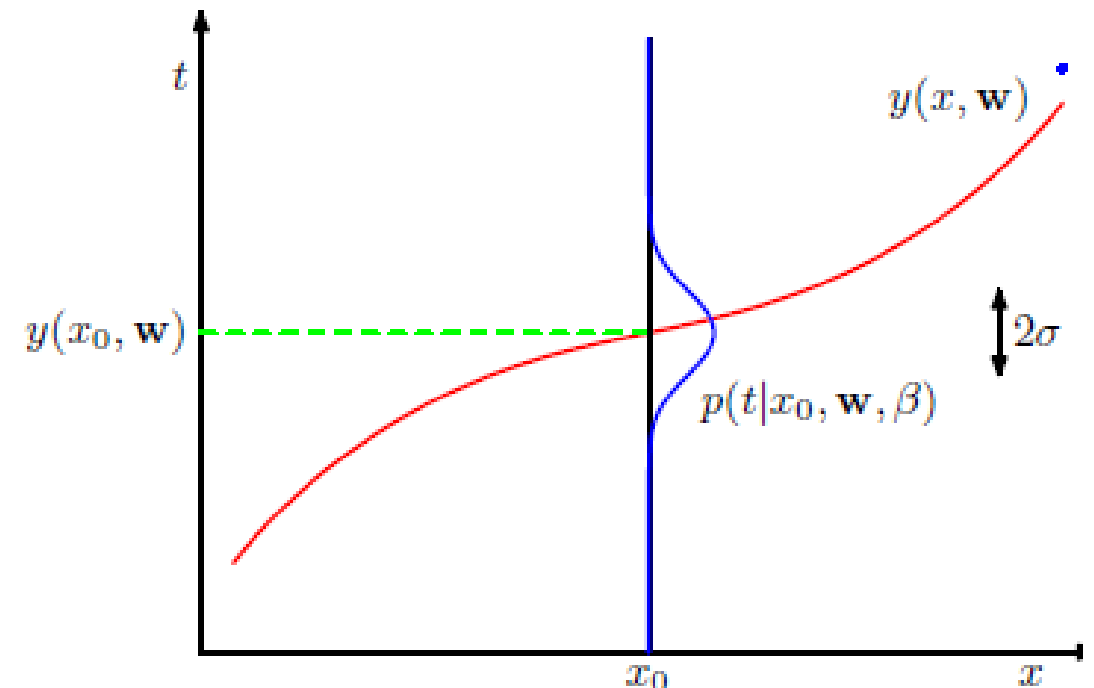
- We now use the training data $\{x, t\}$ to determine the values of the unknown parameters w and β by maximum likelihood

$$p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta) = \prod_{n=1}^N \mathcal{N}(t_n | y(x_n, \mathbf{w}), \beta^{-1})$$

- Log Likelihood

$$\ln p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta) = -\frac{\beta}{2} \sum_{n=1}^N \{y(x_n, \mathbf{w}) - t_n\}^2 + \frac{N}{2} \ln \beta - \frac{N}{2} \ln(2\pi)$$

$$p(t|x, \mathbf{w}, \beta) = \mathcal{N}(t | y(x, \mathbf{w}), \beta^{-1})$$



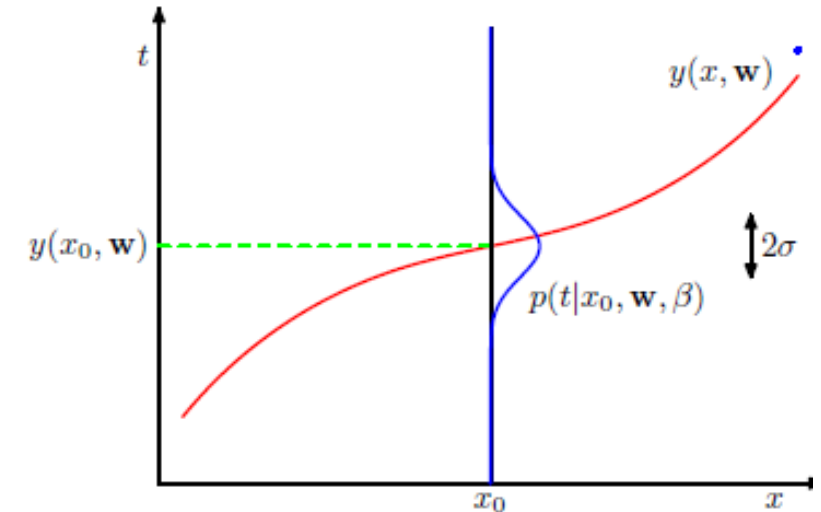
Maximum Likelihood and Least Squares

- Log Likelihood

$$\ln p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \beta) = -\frac{\beta}{2} \sum_{n=1}^N \{y(x_n, \mathbf{w}) - t_n\}^2 + \frac{N}{2} \ln \beta - \frac{N}{2} \ln(2\pi)$$

- Maximize Log Likelihood wrt to $\mathbf{w}$
- Since last two terms, don't depend on $\mathbf{w}$, they can be omitted.
- Also, scaling the log likelihood by a positive constant $\beta/2$ does not alter the location of the maximum with respect to $\mathbf{w}$, so it can be ignored

- Result: Maximize $-\sum_{n=1}^N \{y(x_n, \mathbf{w}) - t_n\}^2$



Maximum Likelihood and Least Squares

- MLE

- Maximize $-\sum_{n=1}^N \{y(x_n, \mathbf{w}) - t_n\}^2$

- Least Squares

- Minimize $\sum_{n=1}^N [t^{(n)} - y^{(n)}]^2$

- Therefore, maximizing likelihood is equivalent, so far as determining $\mathbf{w}$ is concerned, to minimizing the sum-of-squares error function
- Significance: sum-of-squares error function arises as a consequence of maximizing likelihood under the assumption of a Gaussian noise distribution

