Department of Mathematics
University of Calabria

Business Intelligence and Analytics

(Data Mining)

Estimation Theory

Ph.D. Ettore Ritacco

Department of Mathematics
University of Calabria

Outline

1. Review of probability theory
2. Theorem of total probability
3. Bayes' Theorem
4. Random variables and distributions
5. Random vectors
6. Normal distribution
7. Laws of large numbers
8. Central Limit Theorem
9. Estimation theory
   - Minimum Mean Square Error
   - Maximum Likelihood Estimation
   - Bayesian Estimation
10. Optimization
    - Newton-Raphson method
    - Gradient Descent

Department of Mathematics
University of Calabria

Review of probability theory

- Definitions (informal)
  - Probability is a number assigned to an event
    - It indicates "***how likely***" the event will occur when a random experiment is performed
  - A probability law for a random experiment is a rule that assigns probabilities to the events in the experiment
  - The sample space Ω of a random experiment is the set of all possible outcomes
- Axioms of probability
  - Axiom I: $p(A) \geq 0$
  - Axiom II: $p(\Omega) = 1$
  - Axiom III: $A \cap B = \emptyset \Rightarrow p(A \cup B) = p(A) + p(B)$

Department of Mathematics
University of Calabria

Review of probability theory

- More properties of probability
  - $p(\neg A) = 1 - p(A)$
  - $0 \leq p(A) \leq 1$
  - $p(\emptyset) = 0$
  - $p(A \cup B) = p(A) + p(B) - p(A \cap B)$
  - $A \subset B \Rightarrow p(A) \leq p(B)$

Review of probability theory

- Conditional probability
  - If A and B are two events, the probability of A, when we already know that B has occurred, is:

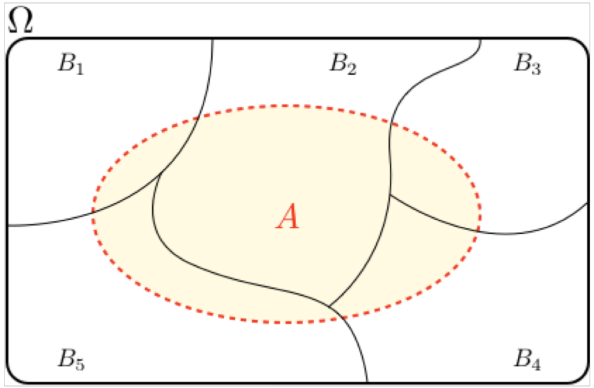
$$p(A|B) = \frac{p(A \cap B)}{p(B)} \quad if \quad p(B) > 0$$

$$\Rightarrow p(A \cap B) = p(A|B) \cdot p(B) = p(B|A) \cdot p(A)$$

    - This conditional probability $p(A|B)$ is read:
      - The "conditional probability of A conditioned on B", or simply
      - "The probability of A given B"
  - Interpretation
    - The new evidence "B has occurred" has the following effects:
      - The original sample space Ω becomes B
      - The event A becomes $A \cap B$
    - $p(B)$ normalizes the probability of events that occur jointly with B

Theorem of total probability

- Let $\{B_1, \dots, B_n\}$ be a partition of Ω, i.e.:
  - $B_i \cap B_j = \emptyset \quad \forall i, j$
  - $\bigcup_{k=1}^{n} B_k = \Omega$
- Then:
  - $A = A \cap \Omega = A \cap (\bigcup_{k=1}^{n} B_k) = \bigcup_{k=1}^{n} A \cap B_k$
- So:
  - $p(A) = p(\bigcup_{k=1}^{n} A \cap B_k)$
    $= \sum_{k=1}^{n} p(A \cap B_k)$
    $= \sum_{k=1}^{n} p(A|B_k) \cdot p(B_k)$



Department of Mathematics
University of Calabria

Bayes' Theorem

- Given the partition $\{B_1, \dots, B_n\}$ of Ω
- Given an occurring event A
- What is the probability of B_j?
- By exploiting the conditional and total probabilities:

Likelihood — Prior probability — Posterior probability — Evidence

$$p(B_j|A) = \frac{p(A \cap B_j)}{p(A)} = \frac{p(A|B_j) \cdot p(B_j)}{p(A)} = \frac{p(A|B_j) \cdot p(B_j)}{\sum_{k=1}^{n} p(A|B_k) \cdot p(B_k)}$$

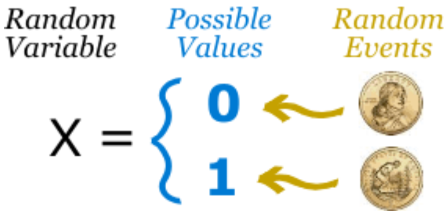
- This is known as Bayes Theorem or Bayes Rule, and is (one of) the most useful relation(s) in probability and statistics

Random variables and distributions

- A random variable is a function that maps the events, in the sample space Ω, into a numerical space:

$$X : \Omega \rightarrow Q$$

  - If $Q \subseteq \mathbb{N}$ then X is discrete
  - If $Q \subseteq \mathbb{R}$ then X is continuous



Random variables and distributions

- The probability of a random variable is a function, often called ***distribution***, that maps the numeric values of the events to the real interval [0,1]:

$$p: Q \to [0,1]$$

- Discrete case:

Random Variable → *Observation*

$$0 \leq p(X = x) \leq 1$$

$$p(X = x) = f(x)$$ ← *Probability mass function*

Cumulative probability distribution → $$p(X \leq x) = F(x) = \sum_{x_i \leq x} f(x_i)$$

$$\sum_{x \in Q} p(X = x) = 1$$

- Continuous case:

$$p(X = x) = 0$$

Distribution function

$$p(X \leq x) = F(x) = \int_{-\infty}^{x} f(s)\, ds$$

Density function

$$p(a \leq X \leq b) = \int_{a}^{b} f(s)\, ds = F(b) - F(a)$$

$$p(-\infty \leq X \leq \infty) = 1$$

Random variables and distributions

- Expected value (average, mean):
  - Discrete case:

$$E_p[X] = \sum_{x \in Q} x \cdot p(X = x)$$

  - Continuous case:

$$E_f[X] = \int_Q x \cdot f(x)\, dx$$

- Variance:

$$Var[X] = E[(X - E[X])^2] = E[X^2] - E[X]^2$$

  - Discrete case:

$$Var_p[X] = \sum_{x \in Q} (x - E[X])^2 \cdot f(x)$$

  - Continuous case:

$$Var_f[X] = \int_Q (x - E[X])^2 \cdot f(x)\, dx$$

Random vectors

- An extension of the concept of random variable
  - A random vector $\bar{X}$ is a function that assigns a vector of real numbers to each outcome in the sample space
- The probability of a random vector observation is a joint probability distribution function:

$$F_{\bar{X}}(\bar{x}) = p[(X_1 \leq x_1) \cap \cdots \cap (X_n \leq x_n)]$$

- whose probability density function (continuous case) is

$$f_{\bar{X}}(\bar{x}) = \frac{\partial^{nF_{\bar{X}}(\bar{x})}}{\partial x_1 \ldots \partial x_n}$$

Department of Mathematics
University of Calabria

Random vectors

- Expected value:

$$\mathrm{E}[\bar{X}] = \bar{\mu} = \left[E[\bar{X}_1], \dots, E[\bar{X}_n]\right]^T = [\mu_1, \dots, \mu_n]^T$$

- Variance should consider correlations ➔ Covariance matrix:

$$Cov[\bar{X}] = \Sigma = E[(\bar{X} - \bar{\mu})(\bar{X} - \bar{\mu})^T] =$$

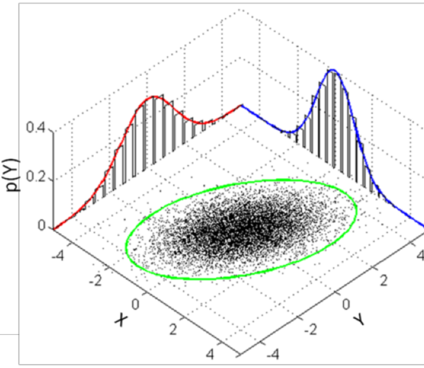
$$= \begin{bmatrix} E[(x_1 - \mu_1)^2] & \dots & E[(x_1 - \mu_1)(x_n - \mu_n)] \\ \dots & \dots & \dots \\ E[(x_n - \mu_n)(x_1 - \mu_1)] & \dots & E[(x_n - \mu_n)^2] \end{bmatrix} = \begin{bmatrix} \sigma_1^2 & \dots & cov_{1,n} \\ \dots & \dots & \dots \\ cov_{n,1} & \dots & \sigma_n^2 \end{bmatrix}$$

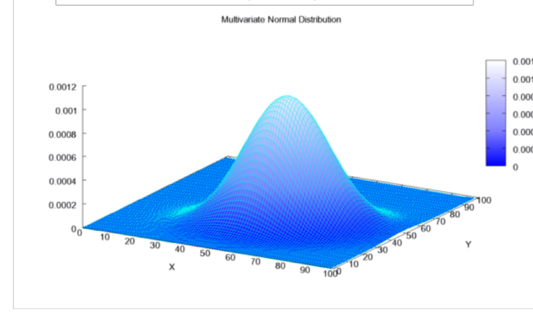
- The covariance matrix indicates the tendency of each pair of features (dimensions in a random vector) to vary together.

- In general, covariance is:

$$Cov[X, Y] = E[(X - E[X]) \cdot (Y - E[Y])] = E[X \cdot Y] - E[X] \cdot E[Y]$$

Normal distributions





- The multivariate Normal (Gaussian) distribution is continuous and defined as:

Mean

$$f_{\bar{X}}(\bar{x}) = \frac{1}{\sqrt{(2\pi)^n|\Sigma|}} \exp\left\{-\frac{1}{2}(\bar{x}-\bar{\mu})^T\Sigma(\bar{x}-\bar{\mu})\right\}$$

Covariance Matrix

- where $|\bar{X}| = n$

- The univariate version is:

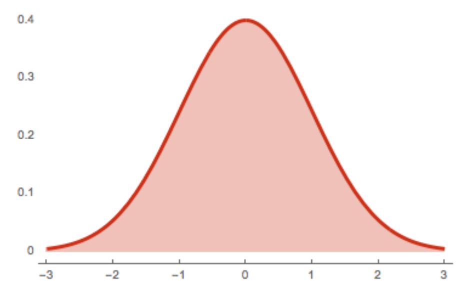
Mean

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$$

Variance

- Expected value $\mathrm{E}[X] = \mu$

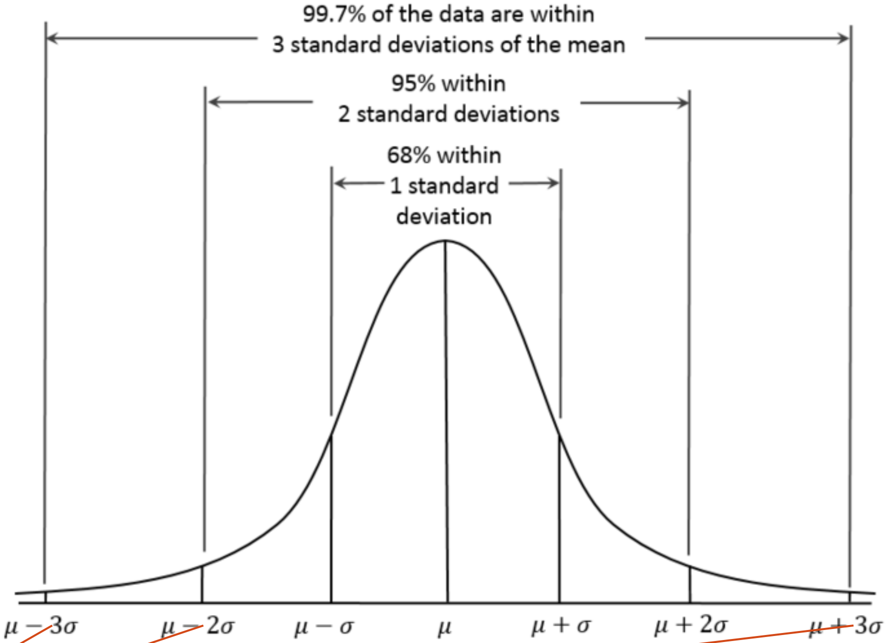
- Variance $Var[X] = \sigma^2$



Department of Mathematics
University of Calabria

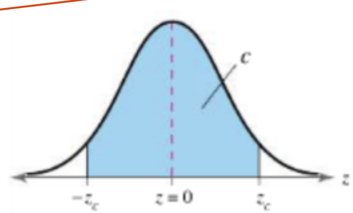
Normal distributions

- Confidence intervals



Critical Values

Level of Confidence c	z_c
0.80	1.28
0.90	1.645
0.95	1.96
0.99	2.575



Department of Mathematics
University of Calabria

Binomial distribution

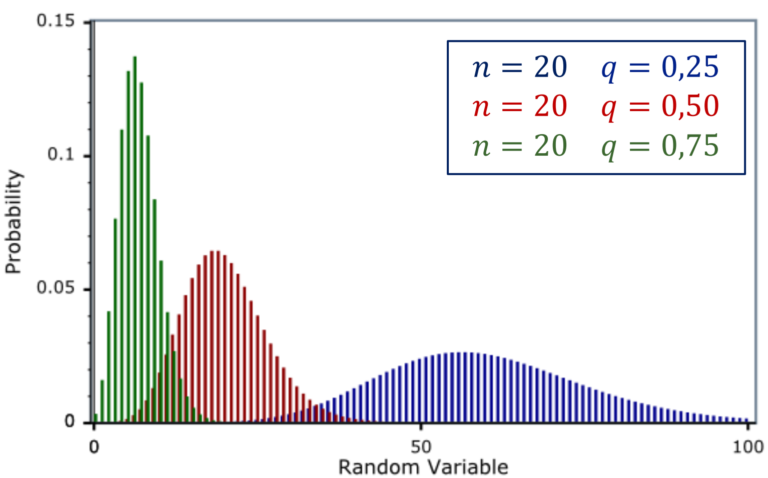
- Probability mass function

$$p(k|n,q) = p(X = k|n,q) = \binom{n}{k} q^k \cdot (1-q)^{n-k}$$

where n and k are integers, q is the probability of a target event and $\binom{n}{k} = \frac{n!}{k!\cdot(n-k)!}$

- Cumulative distribution

$$p(X \leq k) = \sum_{i=0}^{k} \binom{n}{i} q^i \cdot (1-q)^{n-i}$$



- Expected value $E[X] = n \cdot q$
- Variance $Var[X] = n \cdot q \cdot (1-q)$

Laws of large numbers

- The laws of large numbers describe the result of performing the same experiment a large number of times.
- Given a set of independent and identically distributed random variables $\{X_1, \dots, X_n\}$, such that $\forall k\ E[X_k] = \mu$, let define the sample average:

$$S_n = \frac{\sum_{i=1}^{n} X_i}{n}$$

- The **weak law of large numbers** states that the sample average converges *in probability* towards the expected value:

$$\lim_{n\to\infty} p(|S_n - \mu| < \text{const}) = 1$$

- The **strong law of large numbers** states that the sample average converges *almost surely* to the expected value

$$p\left(\lim_{n\to\infty} S_n = \mu\right) = 1$$

Department of Mathematics
University of Calabria

Central Limit Theorem

- Let $\{X_1, \dots, X_n\}$ be a sequence of n independent and identically distributed (i.i.d.) random variables drawn from a distribution of expected value μ and finite variance σ^2
- Let

$$S_n = \frac{\sum_{i=1}^{n} X_i}{n}$$

- Theorem: For large enough n, the distribution of S_n is close to a normal distribution with mean μ and variance $\frac{\sigma^2}{n}$
  - No matter what the shape of the original distribution is!

Estimation theory

- The estimation problem:
  - Let $X = \{X_1, \dots, X_n\}$ be a set of n i.i.d. random variable governed by a probability density function $p(x|\Theta)$, where Θ is unknown
  - Find an estimation of Θ by exploiting the observations of the random variables
  - Three common approaches to solve the problem are:
    - Minimum Mean Squared Error / Least Squares Error
    - Maximum Likelihood estimation
    - Bayesian estimation

Department of Mathematics
University of Calabria

Minimum Mean Squared Error

- Suppose we have a system governed by:

$$Y = f(X|\Phi)$$

- Suppose to run a set of experiments obtaining several observations for X and Y
- Objective:
  - Find $g(X|\Theta)$, an approximation of $f(X|\Phi)$, such that the mean square error

$$E[Y - g(X|\Theta)]^2$$

is minimized

Department of Mathematics
University of Calabria

Minimum Mean Squared Error

- The objective is too hard to automatically achieve
- New objective:
  - Given a chosen function $g(X|\Theta)$, as approximation of $f(X|\Phi)$, find Θ^* such that:

$$\Theta^* = \underset{\Theta}{\operatorname{argmin}}\{E[Y - g(X|\Theta)]^2\}$$

  - Exploiting the observations:

$$\Theta^* = \underset{\Theta}{\operatorname{argmin}}\left\{\sum_{i=1}^{n}\left(y_i - g(x_i|\Theta)\right)^2\right\}$$

- This estimation is also known as least squared error *(LSE)*

Department of Mathematics
University of Calabria

Minimum Mean Squared Error

- Constant case: $g(x|\theta) = \theta$, where $\theta \in \mathbb{R}$
- Then:

$$\theta^* = \underset{\theta \in \mathbb{R}}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} (y_i - \theta)^2 \right\}$$

- Optimization step --- We take derivatives and equate to 0

$$\frac{\partial}{\partial \theta} \sum_{i=1}^{n} (y_i - \theta)^2 = -2 \sum_{i=1}^{n} (y_i - \theta) = -2 \left[\sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \theta \right] = -2 \left[\sum_{i=1}^{n} (y_i) - n \cdot \theta \right] = 0$$

$$\Rightarrow \quad n \cdot \theta = \sum_{i=1}^{n} y_i \quad \Rightarrow \quad \theta^* = \frac{1}{n} \sum_{i=1}^{n} y_i \quad (i.e. \, the \ sample \ mean)$$

Department of Mathematics
University of Calabria

Minimum Mean Squared Error

- Linear case: $g(x|m, q) = m \cdot x + q$, where $m, q \in \mathbb{R}$
- Then

$$\theta^* = \underset{\theta \in \mathbb{R}}{\operatorname{argmin}} \left\{ \sum_{i=1}^{n} (y_i - m \cdot x_i - q)^2 \right\}$$

- Optimization step --- We take derivatives and equate to 0

$$\frac{\partial}{\partial m} \sum_{i=1}^{n} (y_i - m \cdot x_i - q)^2 = -2 \sum_{i=1}^{n} (y_i - m \cdot x_i - q) \cdot x_i = 0$$

$$\frac{\partial}{\partial q} \sum_{i=1}^{n} (y_i - m \cdot x_i - q)^2 = -2 \sum_{i=1}^{n} (y_i - m \cdot x_i - q) = 0$$

Minimum Mean Squared Error

- This is a complete system of equations (2 equations and 2 variables), whose solution is:

$$m^* = \frac{n\sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \sum_{i=1}^n y_i}{n\sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i\right)^2} = \frac{Cov[X,Y]}{Var[X]}$$

$$\sum_{i=1}^{n}(y_i - m^* \cdot x_i - q) = \sum_{i=1}^{n} y_i - m^* \sum_{i=1}^{n} x_i - n \cdot q$$

$$= \frac{1}{n}\sum_{i=1}^{n} y_i - m^* \cdot \frac{1}{n}\sum_{i=1}^{n} x_i - q = 0 \quad \Rightarrow \quad q^* = E[Y] - m^* E[X]$$

Minimum Mean Squared Error

- Multivariate linear case:

$$g(\bar{X}|\bar{A}) = \bar{X} \cdot \bar{A}$$

where $\bar{X} \in \mathbb{R}^{n\times[m+1]}$ and $\bar{A} \in \mathbb{R}^{m+1}$

- In expanded form:

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{1,1} & \cdots & x_{1,m} \\ 1 & x_{2,1} & \cdots & x_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n,1} & \cdots & x_{n,m} \end{bmatrix} \cdot \begin{bmatrix} a_0 \\ a_1 \\ \vdots \\ a_m \end{bmatrix}$$

Minimum Mean Squared Error

- Then:

$$\bar{A}^* = \underset{\bar{A}\in\mathbb{R}^{m+1}}{\operatorname{argmin}}\{\|\bar{Y} - \bar{X}\cdot\bar{A}\|_2^2\}$$

- Optimization step --- We take derivatives and equate to 0

$$\nabla\|\bar{Y} - \bar{X}\cdot\bar{A}\|_2^2 = -2\cdot\bar{X}^T\cdot(\bar{Y} - \bar{X}\cdot\bar{A}) = 0$$

$$\Rightarrow (\bar{X}^T\cdot\bar{X})\cdot\bar{A} = \bar{X}^T\cdot\bar{Y}$$

$$\Rightarrow \bar{A}^* = (\bar{X}^T\cdot\bar{X})^{-1}\cdot\bar{X}^T\cdot\bar{Y}$$

- The term $(\bar{X}^T\cdot\bar{X})^{-1}\cdot\bar{X}^T$ is known as the pseudo-inverse of $\bar{X}$

Minimum Mean Squared Error

- If $\bar{X}^T \cdot \bar{X}$ is a singular matrix (non invertible) the objective can be modified in:

$$\bar{A}^* = \underset{\bar{A}\in\mathbb{R}^{m+1}}{\operatorname{argmin}}\{||\bar{Y} - \bar{X} \cdot \bar{A}||_2^2 + \alpha||A||_2^2\}$$

where α is a *regularization* parameter

- The estimation then is:

$$\bar{A}^* = (\bar{X}^T \cdot \bar{X} + \alpha \cdot I)^{-1} \cdot \bar{X}^T \cdot \bar{Y}$$

which is normally known as *regularized LSE* or *ridge-regression* solution

Maximum Likelihood Estimation

- Maximum Likelihood Estimation (MLE) is one of the most used parametric estimation method
- Let $\{X_1, \dots, X_n\}$ be i.i.d. random variables whose observations are $\{x_1, \dots, x_n\}$
- Let $p(x|\Theta)$ be a distribution that approximate the function that governs the data
- Goal:

Likelihood (→ $p(X|\Theta)$)

$$\Theta^* = \underset{\Theta}{\operatorname{argmax}}\, p(X|\Theta)$$

$$= \underset{\Theta}{\operatorname{argmax}} \prod_{i=1}^{n} p(x_i|\Theta) \quad (\textit{since the observations are independent})$$

Department of Mathematics
University of Calabria

Maximum Likelihood Estimation

- For the sake of simplicity (and numerical calculus), likelihood is typically expressed in logarithmic form:

$$llk(\Theta|X) = \log \prod_{i=1}^{n} p(x_i|\Theta) = \sum_{i=1}^{n} \log p(x_i|\Theta)$$

- As before the optimization step can be performed by taking the derivatives

Department of Mathematics
University of Calabria

Maximum Likelihood Estimation

- Gaussian case:

$$p(x_i|\Theta = \{\mu, \sigma^2\}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x_i-\mu)^2}{2\sigma^2}\right\}$$

- The log likelihood is:

$$\sum_{i=1}^{n} \log p(x_i|\Theta) = \sum_{i=1}^{n} \log\left(\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x_i-\mu)^2}{2\sigma^2}\right\}\right)$$

$$= \sum_{i=1}^{n} \log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) - \frac{(x_i-\mu)^2}{2\sigma^2}$$

$$= -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2$$

Department of Mathematics
University of Calabria

Maximum Likelihood Estimation

- Optimization step --- We take derivatives and equate to 0

$$\mu^* = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Sample mean

Sample variance

$$\sigma^{2^*} = \frac{1}{n}\sum_{i=1}^{n} (x_i - \mu^*)^2$$

Department of Mathematics
University of Calabria

Bayesian Estimation

- Bayesian estimation differs from MLE by considering Θ as a random variable, not a fixed value
- Maximum A Posteriori (MAP) estimation:

$$\Theta^* = \underset{\Theta}{\operatorname{argmax}}\{p(\Theta|X)\} = \underset{\Theta}{\operatorname{argmax}}\left\{\frac{p(X|\Theta)\cdot p(\Theta)}{p(X)}\right\}$$

$$= \underset{\Theta}{\operatorname{argmax}}\{p(X|\Theta)\cdot p(\Theta)\}$$

Likelihood

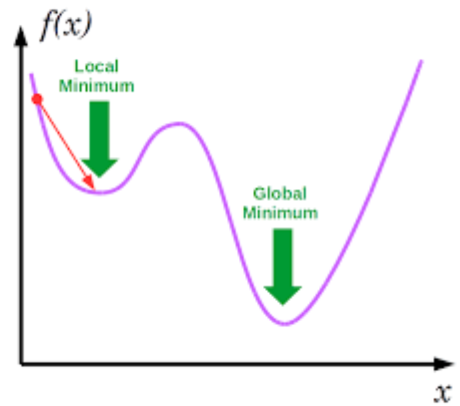
A Priori knowledge about the parameter

Bayesian Estimation

- The Map estimator enables the embedding of prior knowledge about the parameters Θ in terms of $p(\Theta)$
  - With limited data, $p(\Theta)$ is dominant
  - With sufficient data, $p(\Theta)$ balances the likelihood with the background knowledge
  - For large data repositories, $p(\Theta)$ approximates the MLE approach

Optimization

- All the optimization steps seen so far are based on exact derivatives
- There are cases where derivatives are intractable due to the size of the problem
- Typically, we need find heuristics and we have to be content with optimal (non optima) solutions
  - Newton-Raphson method (Root-finding algorithm)
  - Gradient Descent (Finding local minimum)



Department of Mathematics
University of Calabria

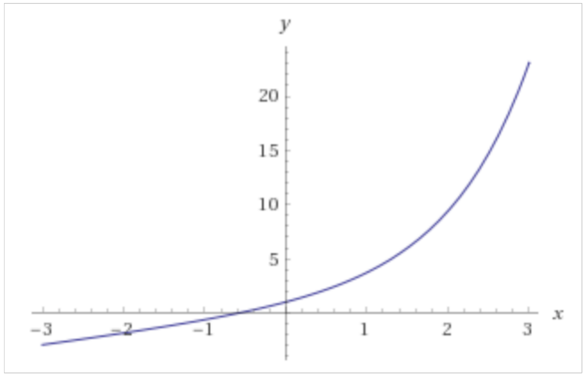
Newton-Raphson method

- Newton-Raphson method is an heuristic for solving the problem of finding approximations of the roots of a function:

$$f(x) = 0$$

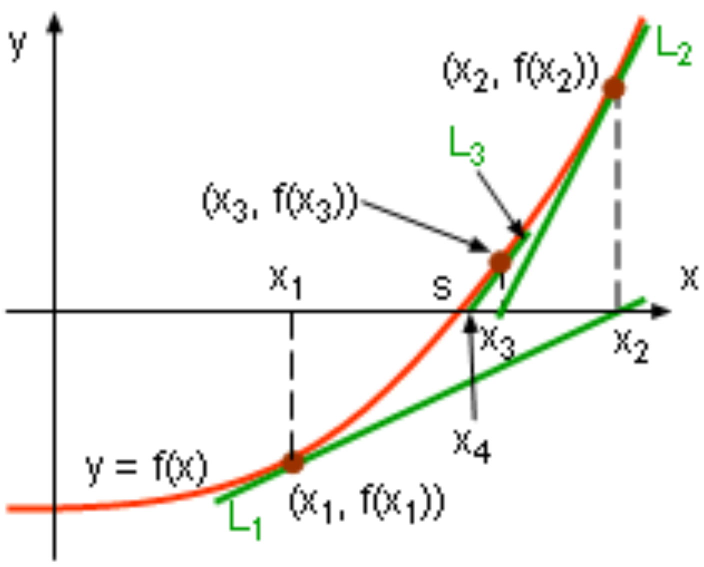
  - For example:

$$x + e^x = 0$$



Newton-Raphson method

- The idea is to exploit the derivative of the function to follow the tangent starting from a random initial point



- The algorithm is: update x until convergence:

$$x^{new} = x^{old} - \frac{f(x^{old})}{f'(x^{old})}$$

- In our example $f'(x^{old}) = 1 + e^{x^{old}}$ (very simple!)

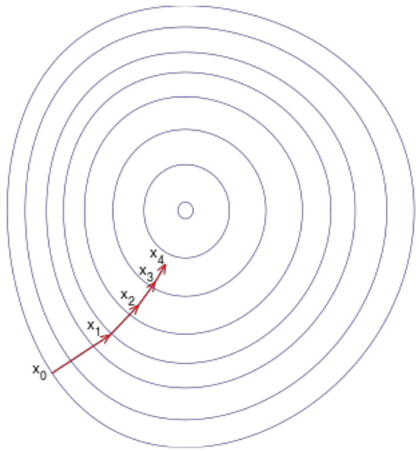
Department of Mathematics
University of Calabria

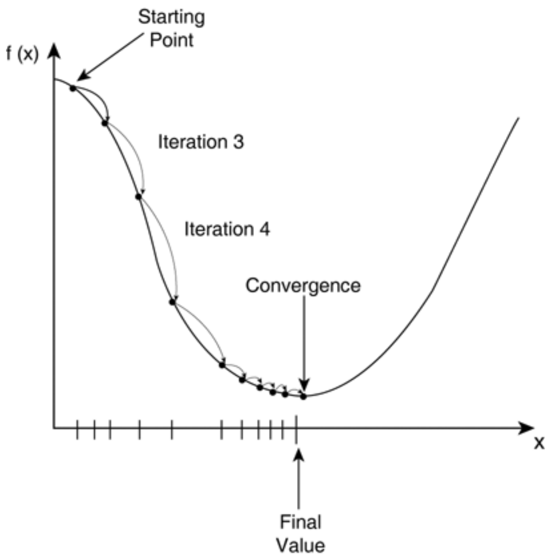
Gradient Descent

- Gradient Descent is an iterative algorithm for searching for minimal points
- Let $F(\bar{x})$ be a multivariate and differentiable in a neighborhood of a point $\bar{a}$
  - $F(\bar{x})$ decreases *fastest* if one goes from $\bar{a}$ in the direction of the negative gradient
  - The algorithm is: update $\bar{a}$ until convergence:

$$\bar{a}^{new} = \bar{a}^{old} - \eta \nabla F\left(\bar{a}^{old}\right)$$

The parameter η is called learning rate and determines the behavior of the optimization





Department of Mathematics
University of Calabria

Gradient Descent

- Possible behaviors according η
  - Let assume that our error function is:

$$e(w) = \frac{1}{2} \cdot C \cdot w^2 \qquad (where\ C\ is\ a\ constant)$$