

Probability—the Science of Uncertainty and Data

by Fabián Korynski

Discrete random variables

Probability mass function and expectation

Definition (Random variable) A random variable X is a function of the sample space Ω into the real numbers (or \mathbb{R}^n). Its range can be discrete or continuous.

Definition (Probability mass function (PMF)) The probability law of a discrete random variable X is called its PMF. It is defined as

$$\mathbb{P}(A_i|B) = \frac{\mathbb{P}(A_i)\mathbb{P}(B|A_i)}{\sum_j \mathbb{P}(A_j)\mathbb{P}(B|A_j)}.$$

Probability models and axioms

Definition (Sample space) A sample space Ω is the set of all possible outcomes. The set's elements must be mutually exclusive, collectively exhaustive and at the right granularity.

Definition (Event) An event is a subset of the sample space. Probability is assigned to events.

Definition (Probability axioms) A probability law \mathbb{P} assigns probabilities to events and satisfies the following axioms:

Nonnegativity $\mathbb{P}(A) \geq 0$ for all events A .

Normalization $\mathbb{P}(\Omega) = 1$.

(Countable) additivity For every sequence of events A_1, A_2, \dots such that $A_i \cap A_j = \emptyset$: $\mathbb{P}\left(\bigcup_i A_i\right) = \sum_i \mathbb{P}(A_i)$.

Corollaries (Consequences of the axioms)

$$\mathbb{P}(\emptyset) = 0.$$

For any finite collection of disjoint events A_1, \dots, A_n ,

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

$$\mathbb{P}(A) + \mathbb{P}(A^c) = 1.$$

$$\mathbb{P}(A) \leq 1.$$

If $A \subset B$, then $\mathbb{P}(A) \leq \mathbb{P}(B)$.

$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B).$$

$$\mathbb{P}(A \cup B) \leq \mathbb{P}(A) + \mathbb{P}(B).$$

Example (Discrete uniform law) Assume Ω is finite and consists of n equally likely elements. Also, assume that $A \subset \Omega$ with k elements. Then $\mathbb{P}(A) = \frac{k}{n}$.

Conditioning and Bayes' rule

Definition (Conditional probability) Given that event B has occurred and that $P(B) > 0$, the probability that A occurs is

$$\mathbb{P}(A|B) \triangleq \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}.$$

Remark (Conditional probabilities properties) They are the same as ordinary probabilities. Assuming $\mathbb{P}(B) > 0$:

- $\mathbb{P}(A|B) \geq 0$.
- $\mathbb{P}(\Omega|B) = 1$.
- $\mathbb{P}(B|B) = 1$.

Proposition (Multiplication rule)

$$\mathbb{P}(A_1 \cap A_2 \cap \dots \cap A_n) = \mathbb{P}(A_1) \cdot \mathbb{P}(A_2|A_1) \cdot \mathbb{P}(A_n|A_1 \cap A_2 \cap \dots \cap A_{n-1}).$$

Theorem (Total probability theorem) Given a partition $\{A_1, A_2, \dots\}$ of the sample space, meaning that $\bigcup_i A_i = \Omega$ and the events are disjoint, and for every event B , we have

$$\mathbb{P}(B) = \sum_i \mathbb{P}(A_i)\mathbb{P}(B|A_i).$$

Independence

Definition (Independence of events) Two events are independent if occurrence of one provides no information about the other. We say that A and B are independent if

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

$$\mathbb{P}(B|A) = \mathbb{P}(B) \quad \mathbb{P}(A|B) = \mathbb{P}(A).$$

Remarks

- The definition of independence is symmetric with respect to A and B .
- The product definition applies even if $\mathbb{P}(A) = 0$ or $\mathbb{P}(B) = 0$.

Corollary If A and B are independent, then A and B^c are independent. Similarly for A^c and B , or for A^c and B^c .

Definition (Conditional independence) We say that A and B are independent conditioned on C , where $\mathbb{P}(C) > 0$, if

$$\mathbb{P}(A \cap B|C) = \mathbb{P}(A|C)\mathbb{P}(B|C).$$

Definition (Independence of a collection of events) We say that events A_1, A_2, \dots, A_n are independent if for every collection of distinct indices i_1, i_2, \dots, i_k , we have

$$\mathbb{P}(A_{i_1} \cap \dots \cap A_{i_k}) = \mathbb{P}(A_{i_1}) \cdot \mathbb{P}(A_{i_2}) \cdots \mathbb{P}(A_{i_k}).$$

Counting

This section deals with finite sets with uniform probability law. In this case, to calculate $\mathbb{P}(A)$, we need to count the number of elements in A and in Ω .

Remark (Basic counting principle) For a selection that can be done in r stages, with n_i choices at each stage i , the number of possible selections is $n_1 \cdot n_2 \cdots n_r$.

Definition (Permutations) The number of permutations (orderings) of n different elements is

$$n! = 1 \cdot 2 \cdot 3 \cdots n.$$

Definition (Combinations) Given a set of n elements, the number of subsets with exactly k elements is

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}.$$

Definition (Partitions) We are given an n -element set and nonnegative integers n_1, n_2, \dots, n_r , whose sum is equal to n . The number of partitions of the set into r disjoint subsets, with the i^{th} subset containing exactly n_i elements, is equal to

$$\binom{n}{n_1, \dots, n_r} = \frac{n!}{n_1!n_2!\cdots n_r!}.$$

Remark This is the same as counting how to assign n distinct elements to r people, giving each person i exactly n_i elements.

Properties (Properties of expectation)

- If $X \geq 0$ then $\mathbb{E}[X] \geq 0$.
- If $a \leq X \leq b$ then $a \leq \mathbb{E}[X] \leq b$.
- If $X = c$ then $\mathbb{E}[X] = c$.
- Example Expected value of know r.v.
- If $X \sim \text{Ber}(p)$ then $\mathbb{E}[X] = p$.
- If $X = I_A$ then $\mathbb{E}[X] = \mathbb{P}(A)$.
- If $X \sim \text{Uni}[a, b]$ then $\mathbb{E}[X] = \frac{a+b}{2}$.
- If $X \sim \text{Bin}(n, p)$ then $\mathbb{E}[X] = np$.
- If $X \sim \text{Geo}(p)$ then $\mathbb{E}[X] = \frac{1}{p}$.

Theorem (Expected value rule) Given a random variable X and a function $g : \mathbb{R} \rightarrow \mathbb{R}$, we construct the random variable $Y = g(X)$. Then

$$\sum_y y p_{YY}(y) = \mathbb{E}[Y] = \mathbb{E}[g(X)] = \sum_x g(x)p_X(x).$$

Remark (PMF of $Y = g(X)$) The PMF of $Y = g(X)$ is $p_Y(y) = \sum_{x: g(x)=y} p_X(x)$.

Remark In general $g(\mathbb{E}[X]) \neq \mathbb{E}[g(X)]$. They are equal if $g(x) = ax + b$.

Variance, conditioning on an event, multiple r.v.

Definition (Variance of a random variable) Given a random variable X with $\mu = \mathbb{E}[X]$, its variance is a measure of the spread of the random variable and is defined as

$$\text{Var}(X) \triangleq \mathbb{E}[(X - \mu)^2] = \sum_x (x - \mu)^2 p_X(x).$$

Definition (Standard deviation)

$$\sigma_X = \sqrt{\text{Var}(X)}.$$

Properties (Properties of the variance)

- $\text{Var}(aX) = a^2 \text{Var}(X)$, for all $a \in \mathbb{R}$.
- $\text{Var}(X + b) = \text{Var}(X)$, for all $b \in \mathbb{R}$.
- $\text{Var}(aX + b) = a^2 \text{Var}(X)$.
- $\text{Var}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$.

Example (Variance of known r.v.)

- If $X \sim \text{Ber}(p)$, then $\text{Var}(X) = p(1-p)$.
- If $X \sim \text{Uni}[a, b]$, then $\text{Var}(X) = \frac{(b-a)(b-a+2)}{12}$.
- If $X \sim \text{Bin}(n, p)$, then $\text{Var}(X) = np(1-p)$.
- If $X \sim \text{Geo}(p)$, then $\text{Var}(X) = \frac{1-p}{p}$.

Proposition (Conditional PMF and expectation, given an event)

Given the event A , with $\mathbb{P}(A) > 0$, we have the following

- $p_{X|A}(x) = \mathbb{P}(X = x|A)$.
- If A is a subset of the range of X , then:

$$p_{X|A}(x) \triangleq p_{X|\{X \in A\}}(x) = \begin{cases} \frac{1}{\mathbb{P}(A)} p_X(x), & \text{if } x \in A, \\ 0, & \text{otherwise.} \end{cases}$$

Proposition (Total expectation rule) Given a partition of disjoint events A_1, \dots, A_n such that $\sum_i \mathbb{P}(A_i) = 1$, and $\mathbb{P}(A_i) > 0$,

$$\mathbb{E}[X] = \mathbb{P}(A_1)\mathbb{E}[X|A_1] + \dots + \mathbb{P}(A_n)\mathbb{E}[X|A_n].$$

Definition (Memorylessness of the geometric random variable) When we condition a geometric random variable X on the event $X > n$ we have memorylessness, meaning that the “remaining time” $X - n$, given that $X > n$, is also geometric with the same parameter.

Formally,

$$p_{X-n|X>n}(i) = p_X(i).$$

Definition (Joint PMF) The joint PMF of random variables X_1, X_2, \dots, X_n is

$$p_{X_1, X_2, \dots, X_n}(x_1, \dots, x_n) = \mathbb{P}(X_1 = x_1, \dots, X_n = x_n).$$

Properties (Properties of joint PMF)

Given a random variable X and a function $g : \mathbb{R} \rightarrow \mathbb{R}$, we construct the random variable $Y = g(X)$.

Then

$$\sum_y y p_{YY}(y) = \mathbb{E}[Y] = \mathbb{E}[g(X)] = \sum_x g(x)p_X(x).$$

Remark (PMF of $Y = g(X)$) The PMF of $Y = g(X)$ is

$$p_Y(y) = \sum_{x: g(x)=y} p_X(x).$$

Remark In general $g(\mathbb{E}[X]) \neq \mathbb{E}[g(X)]$. They are equal if $g(x) = ax + b$.

Variance, conditioning on an event, multiple r.v.

Definition (Variance of a random variable) Given a random variable X with $\mu = \mathbb{E}[X]$, its variance is a measure of the spread of the random variable and is defined as

$$\text{Var}(X) \triangleq \mathbb{E}[(X - \mu)^2] = \sum_x (x - \mu)^2 p_X(x).$$

Properties (Properties of the variance)

Proposition (Multiplication rule) Given jointly discrete random variables X, Y , and whenever the conditional probabilities are defined,

$$p_{X,Y}(x,y) = p_X(x)p_{Y|X}(y|x) = p_Y(y)p_{X|Y}(x|y).$$

Definition (Conditional expectation) Given discrete random variables X, Y and y such that $p_{Y|Y}(y) > 0$ we define

$$\mathbb{E}[X|Y = y] = \sum_x x p_{X|Y}(x|y).$$

Additionally we have

$$\mathbb{E}[g(X)|Y = y] = \sum_x g(x)p_{X|Y}(x|y).$$

Theorem (Total probability and expectation theorems)

If $p_{Y|Y}(y) > 0$, then

$$p_X(x) = \sum_y p_Y(y)p_{X|Y}(x|y),$$

$$\mathbb{E}[X] = \sum_y p_Y(y)\mathbb{E}[X|Y = y].$$

Definition (Independence of a random variable and an event)

Given a partition of disjoint events A_1, \dots, A_n such that $\sum_i \mathbb{P}(A_i) = 1$, and $\mathbb{P}(A_i) > 0$,

$$p_{X|A}(x) \triangleq p_{X|\{X \in A\}}(x) = \begin{cases} \frac{1}{\mathbb{P}(A)} p_X(x), & \text{if } x \in A, \\ 0, & \text{otherwise.} \end{cases}$$

Proposition (Total expectation rule) Given a partition of disjoint events A_1, \dots, A_n such that $\sum_i \mathbb{P}(A_i) = 1$, and $\mathbb{P}(A_i) > 0$,

$$\mathbb{E}[X] = \mathbb{P}(A_1)\mathbb{E}[X|A_1] + \dots + \mathbb{P}(A_n)\mathbb{E}[X|A_n].$$

Definition (Independence of two random variables) Two discrete random variables X and Y are independent if

$$p_{X,Y}(x, y) = p_X(x)p_{Y|X}(y) \text{ for all } x, y.$$

Remark (Independence of a collection of random variables) A collection X_1, X_2, \dots, X_n of random variables are independent if

$$p_{X_1, \dots, X_n}(x_1, \dots, x_n) = p_{X_1}(x_1) \cdots p_{X_n}(x_n), \forall x_1, \dots, x_n.$$

Remark (Independence and expectation) In general,

$$\mathbb{E}[g(X, Y)] \neq g(\mathbb{E}[X], \mathbb{E}[Y]).$$

An exception is for linear functions:

$$\mathbb{E}[aX + bY] = a\mathbb{E}[X] + b\mathbb{E}[Y].$$

We have $\mathbb{E}[X] = \frac{a+b}{2}$ and $\text{Var}(X) = \frac{(b-a)^2}{12}$.

Remark If X and Y are independent random variables,

$$\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y].$$

Remark If X and Y are independent, $\mathbb{E}[g(X)h(Y)] = \mathbb{E}[g(X)]\mathbb{E}[h(Y)]$.

Proposition (Variance of sum of independent random variables)

If X and Y are discrete independent random variables,

$$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y).$$

Continuous random variables

PDF, Expectation, Variance, CDF

Definition (Probability density function (PDF)) A probability density function of a r.v. X is a non-negative real valued function f_X that satisfies the following

$$\bullet \int_{-\infty}^{\infty} f_X(x)dx = 1.$$

Properties (Properties of expectation)

$$\bullet \mathbb{P}(a \leq X \leq b) = \int_a^b f_X(x)dx \text{ for some random variable } X.$$

Definition (Continuous random variable) A random variable X is continuous if its probability law can be described by a PDF f_X .

Remark Continuous random variables satisfy:

$$\bullet \text{For small } \delta > 0, \mathbb{P}(a \leq X \leq a + \delta) \approx f_X(a)\delta.$$

$$\bullet \mathbb{P}(X = a) = 0, \forall a \in \mathbb{R}.$$

Definition (Expectation of a continuous random variable) The expectation of a continuous random variable is

$$\mathbb{E}[X] \triangleq \int_{-\infty}^{\infty} xf_X(x)dx.$$

assuming $\int_{-\infty}^{\infty} |x|f_X(x)dx < \infty$.

Properties (Properties of expectation)

$$\bullet \text{If } X \geq 0 \text{ then } \mathbb{E}[X] \geq 0.$$

$$\bullet \text{If } a \leq X \leq b \text{ then } a \leq \mathbb{E}[X] \leq b.$$

$$\bullet \mathbb{E}[g(X)] = \int_{-\infty}^{\infty} g(x)f_X(x)dx.$$

Definition (Variance of a continuous random variable) Given a continuous random variable X with $\mu = \mathbb{E}[X]$, its variance is

$$\text{Var}(X) = \mathbb{E}[(X - \mu)^2] = \int_{-\infty}^{\infty} (x - \mu)^2 f_X(x)dx.$$

It has the same properties as the variance of a discrete random variable.

Example (Uniform continuous random variable) A Uniform continuous random variable X between a and b , with $a < b$, $(X \sim \text{Uni}(a, b))$ has PDF

$$f_X(x) = \begin{cases} \frac{1}{b-a}, & \text{if } a < x < b, \\ 0, & \text{otherwise.} \end{cases}$$

Example (Exponential random variable) An Exponential random variable X with parameter $\lambda > 0$ ($X \sim Exp(\lambda)$) has PDF

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x}, & \text{if } x \geq 0, \\ 0, & \text{otherwise.} \end{cases}$$

We have $E[X] = \frac{1}{\lambda}$ and $\text{Var}(X) = \frac{1}{\lambda^2}$.

Definition (Cumulative Distribution Function (CDF)) The CDF

of a random variable X is $F_X(x) = \mathbb{P}(X \leq x)$. In particular, for a continuous random variable, we have

$$F_X(x) = \int_{-\infty}^x f_X(x) dx,$$

$$f_X(x) = \frac{dF_X(x)}{dx}.$$

Properties (Properties of CDF)

• If $y \geq x$, then $F_X(y) \geq F_X(x)$.

$$\lim_{x \rightarrow -\infty} F_X(x) = 0.$$

$$\lim_{x \rightarrow \infty} F_X(x) = 1.$$

Definition (Normal/Gaussian random variable) A Normal random variable X with mean μ and variance $\sigma^2 > 0$ ($X \sim \mathcal{N}(\mu, \sigma^2)$) has PDF

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

We have $E[X] = \mu$ and $\text{Var}(X) = \sigma^2$.

Remark (Standard Normal) The standard Normal is $\mathcal{N}(0, 1)$.

Proposition (Linearity of Gaussians) Given $X \sim \mathcal{N}(\mu, \sigma^2)$, and if $a \neq 0$, then $aX + b \sim \mathcal{N}(a\mu + b, a^2\sigma^2)$.

Using this $Y = \frac{X-\mu}{\sigma}$ is a standard gaussian.

Conditioning on an event, and multiple continuous r.v.

Definition (Conditional PDF given an event) Given a continuous random variable X and event A with $P(A) > 0$, we define the conditional PDF as the function that satisfies

$$\mathbb{P}(X \in B|A) = \int_B f_{X|A}(x) dx.$$

Definition (Conditional expectation) Given a continuous random variable X and an event A , with $P(A) > 0$:

$$\mathbb{E}[X|A] = \int_{-\infty}^{\infty} x f_{X|A}(x) dx.$$

$$f_{X|X \in A}(x) = \begin{cases} \frac{1}{P(A)} f_X(x), & x \in A, \\ 0, & x \notin A. \end{cases}$$

Definition (Conditional expectation) Given a continuous random variable X and an event A , with $P(A) > 0$:

$$\mathbb{E}[X|A] = \int_{-\infty}^{\infty} g(x) f_{X|A}(x) dx.$$

Definition (Memorylessness of the exponential random variable)

When we condition an exponential random variable X on the event $X > t$ we have memorylessness, meaning that the “remaining time” $X - t$ given that $X > t$ is also geometric with the same parameter i.e.,

$$\mathbb{P}(X - t > x | X > t) = \mathbb{P}(X > x).$$

Theorem (Total probability and expectation theorems) Given a partition of the space into disjoint events A_1, A_2, \dots, A_n such that $\sum_i \mathbb{P}(A_i) = 1$ we have the following:

$$\begin{aligned} F_X(x) &= \mathbb{P}(A_1)F_{X|A_1}(x) + \dots + \mathbb{P}(A_n)F_{X|A_n}(x), \\ f_X(x) &= \mathbb{P}(A_1)f_{X|A_1}(x) + \dots + \mathbb{P}(A_n)f_{X|A_n}(x), \\ \mathbb{E}[X] &= \mathbb{P}(A_1)\mathbb{E}[X|A_1] + \dots + \mathbb{P}(A_n)\mathbb{E}[X|A_n]. \end{aligned}$$

Remark If X and Y are independent, $\mathbb{E}[g(X)h(Y)] = \mathbb{E}[g(X)]\mathbb{E}[h(Y)]$.

Proposition (Variance of sum of independent random variables)
If X and Y are independent continuous random variables,

$$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y).$$

Proposition (Bayes' rule summary)

- For X, Y discrete: $p_{X|Y}(x|y) = \frac{p_X(x)p_{Y|X}(y|x)}{p_Y(y)}$.
- For X, Y continuous: $f_{X|Y}(x|y) = \frac{f_X(x)f_{Y|X}(y|x)}{f_Y(y)}$.
- $F_{X,Y}(x, y) = \mathbb{P}(X \leq x, Y \leq y) = \int_{-\infty}^x \left[\int_{-\infty}^y f_{X,Y}(u, v) du \right] dv.$
- $f_{X,Y}(x) = \frac{\partial^2 F_{X,Y}(x,y)}{\partial x \partial y}.$
- $f_{X,Y}(x) = \frac{\partial^2 F_{X,Y}(x,y)}{\partial y \partial x}.$

Example (Uniform joint PDF on a set S) Let $S \subset \mathbb{R}^2$ with area $s > 0$, then the random variable (X, Y) is uniform over S if it has PDF

$$f_{X,Y}(x, y) = \begin{cases} \frac{1}{s}, & (x, y) \in S, \\ 0, & (x, y) \notin S. \end{cases}$$

Conditioning on a random variable, independence, Bayes' rule

Definition (Conditional PDF given another random variable)

Given jointly continuous random variables X, Y and a value y such that $f_Y(y) > 0$, we define the conditional PDF as

$$f_{X|Y}(x|y) \triangleq \frac{f_{X,Y}(x, y)}{f_Y(y)}.$$

Additionally we define $\mathbb{P}(X \in A|Y = y) = \int_A f_{X|Y}(x|y) dx$.
Proposition (Multiplication rule) Given jointly continuous random variables X, Y , whenever possible we have

$$f_{X,Y}(x, y) = f_X(x)f_{Y|X}(y|x) = f_Y(y)f_{X|Y}(x|y).$$

Definition (Conditional expectation) Given jointly continuous random variables X, Y , and y such that $f_Y(y) > 0$, we define the conditional expected value as

$$\mathbb{E}[X|Y = y] = \int_{-\infty}^{\infty} x f_{X|Y}(x|y) dx.$$

Additionally we have

$$\mathbb{E}[g(X)|Y = y] = \int_{-\infty}^{\infty} g(x) f_{X|Y}(x|y) dx.$$

Theorem (Total probability and total expectation theorems)

1. Find the CDF of Y : $F_Y(y) = \mathbb{P}(Y \leq y) = \mathbb{P}(g(X) \leq y)$.
2. Differentiate the CDF of Y to obtain the PDF: $f_Y(y) = \frac{1}{dy} F'_Y(y) = \frac{1}{dy} \mathbb{P}(g(X) \leq y)$.

Proposition (General formula for monotonic g) Let X be a continuous random variable and g a function that is monotonic wherever $f_X(x) > 0$. The PDF of $Y = g(X)$ is given by

$$f_Y(y) = f_X(h(y)) \left| \frac{dh}{dy}(y) \right|.$$

where $h = g^{-1}$ in the interval where g is monotonic.

Proposition (Expectation of product of independent r.v.) If X and Y are independent continuous random variables,

$$\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y].$$

Remark If X and Y are independent, $\mathbb{E}[g(X)h(Y)] = \mathbb{E}[g(X)]\mathbb{E}[h(Y)]$.

Proposition (Variance of sum of independent random variables)
If X and Y are independent continuous random variables,

$$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y).$$

Proposition (Bayes' rule summary)

- For X, Y discrete: $p_{X|Y}(x|y) = \frac{p_X(x)p_{Y|X}(y|x)}{p_Y(y)}$.
- For X, Y continuous: $f_{X|Y}(x|y) = \frac{f_X(x)f_{Y|X}(y|x)}{f_Y(y)}$.

Proposition (Properties of joint PDFs)

- $f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, y) dy$.
- $F_{X,Y}(x, y) = \mathbb{P}(X \leq x, Y \leq y) = \int_{-\infty}^x \left[\int_{-\infty}^y f_{X,Y}(u, v) du \right] dv.$
- For X discrete, Y continuous: $p_{X|Y}(x|y) = \frac{p_X(x)f_{Y|X}(y|x)}{p_Y(y)}$.
- For X continuous, Y discrete: $f_{X|Y}(x|y) = \frac{f_X(x)p_{Y|X}(y|x)}{p_Y(y)}$.

Derived distributions

Proposition (Discrete case) Given a discrete random variable X and a function g , the r.v. $Y = g(X)$ has PMF

$$p_Y(y) = \sum_{x: g(x)=y} p_X(x).$$

Corollary (Linear function of normal r.v.) If $X \sim \mathcal{N}(\mu, \sigma^2)$ and $Y = aX + b$, with $a \neq 0$, then $p_Y(y) = p_X\left(\frac{y-b}{a}\right)$.

Example (General function of continuous r.v.) Given a continuous random variable X and $Y = g(X)$ we follow the two-step procedure:

$$f_Y(y) = \frac{1}{|a|} f_X\left(\frac{y-b}{a}\right).$$

Proposition (Linear function of continuous r.v.) Given a continuous random variable X and $Y = aX + b$, with $a \neq 0$, we have

Corollary (General formula for monotonic g) If $X \sim \mathcal{N}(\mu, \sigma^2)$ and $Y = aX + b$, with $a > 0$, then $Y \sim \mathcal{N}(a\mu + b, a^2\sigma^2)$.

Example (General function of a continuous r.v.) If X is a continuous random variable and g is any function, to obtain the pdf of $Y = g(X)$ we follow the two-step procedure:

Proposition (Linear function of normal r.v.) If $X \sim \mathcal{N}(\mu, \sigma^2)$ and $Y = aX + b$, then $p_Y(y) = p_X\left(\frac{y-b}{a}\right)$.

Example (General function of a continuous r.v.) If X is a continuous random variable and g is any function, to obtain the pdf of $Y = g(X)$ we follow the two-step procedure:

Proposition (General formula for monotonic g) Let X be a continuous random variable and g a function that is monotonic wherever $f_X(x) > 0$. The PDF of $Y = g(X)$ is given by

Proposition (General formula for monotonic g) Let X be a continuous random variable and g a function that is monotonic wherever $f_X(x) > 0$. The PDF of $Y = g(X)$ is given by

Proposition (General formula for monotonic g) Let X be a continuous random variable and g a function that is monotonic wherever $f_X(x) > 0$. The PDF of $Y = g(X)$ is given by

Proposition (General formula for monotonic g) Let X be a continuous random variable and g a function that is monotonic wherever $f_X(x) > 0$. The PDF of $Y = g(X)$ is given by

Proposition (General formula for monotonic g) Let X be a continuous random variable and g a function that is monotonic wherever $f_X(x) > 0$. The PDF of $Y = g(X)$ is given by

Proposition (General formula for monotonic g) Let X be a continuous random variable and g a function that is monotonic wherever $f_X(x) > 0$. The PDF of $Y = g(X)$ is given by

Proposition (General formula for monotonic g) Let X be a continuous random variable and g a function that is monotonic wherever $f_X(x) > 0$. The PDF of $Y = g(X)$ is given by

Proposition (General formula for monotonic g) Let X be a continuous random variable and g a function that is monotonic wherever $f_X(x) > 0$. The PDF of $Y = g(X)$ is given by

Proposition (General formula for monotonic g) Let X be a continuous random variable and g a function that is monotonic wherever $f_X(x) > 0$. The PDF of $Y = g(X)$ is given by

Proposition (General formula for monotonic g) Let X be a continuous random variable and g a function that is monotonic wherever $f_X(x) > 0$. The PDF of $Y = g(X)$ is given by

Sums of independent r.v., covariance and correlation

Proposition (Discrete case) Let X, Y be discrete independent random variables and $Z = X + Y$, then the PMF of Z is

$$p_Z(z) = \sum_x p_X(x)p_Y(z-x).$$

Proposition (Continuous case) Let X, Y be continuous independent random variables and $Z = X + Y$, then the PDF of Z is

$$f_Z(z) = \int_{-\infty}^{\infty} f_X(x)f_Y(z-x)dx.$$

Proposition (Sum of independent normal r.v.) Let $X \sim \mathcal{N}(\mu_x, \sigma_x^2)$ and $Y \sim \mathcal{N}(\mu_y, \sigma_y^2)$ independent. Then $Z = X + Y \sim \mathcal{N}(\mu_x + \mu_y, \sigma_x^2 + \sigma_y^2)$.

Definition (Covariance) We define the covariance of random variables X, Y as

$$\text{Cov}(X, Y) \triangleq E[(X - E[X])(Y - E[Y])].$$

Properties (Properties of covariance)

- If X, Y are independent, then $\text{Cov}(X, Y) = 0$.

$$\text{Cov}(X, X) = \text{Var}(X).$$

$$\text{Cov}(aX + b, Y) = a\text{Cov}(X, Y).$$

$$\text{Cov}(X, Y + Z) = \text{Cov}(X, Y) + \text{Cov}(X, Z).$$

$$\text{Cov}(X, Y) = E[XY] - E[X]E[Y].$$

Proposition (Variance of a sum of r.v.)

$$\text{Var}(X_1 + \dots + X_n) = \sum_i \text{Var}(X_i) + \sum_{i \neq j} \text{Cov}(X_i, X_j).$$

Definition (Correlation coefficient) We define the correlation coefficient of random variables X, Y , with $\sigma_X, \sigma_Y > 0$, as

$$\rho(X, Y) \triangleq \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y}.$$

Properties (Properties of the correlation coefficient)

- $-1 \leq \rho \leq 1$.
- If X, Y are independent, then $\rho = 0$.
- $|\rho| = 1$ if and only if $X - E[X] = c(Y - E[Y])$.
- $\rho(aX + b, Y) = \text{sign}(a)\rho(X, Y)$.

Conditional expectation and variance, sum of random number of r.v.

Definition (Conditional expectation as a random variable) Given random variables X, Y the conditional expectation $E[X|Y]$ is the random variable that takes the value $E[X|Y = y]$ whenever $Y = y$.

Theorem (Law of iterated expectations)

$$E[E[X|Y]] = E[X].$$

The Central Limit Theorem

Theorem (Central Limit Theorem (CLT)) Given a sequence of independent random variables $\{X_1, X_2, \dots\}$ with $E[X_i] = \mu$ and $\text{Var}(X_i) = \sigma^2$, we define

$$Z_n = \frac{1}{\sigma\sqrt{n}} \sum_{i=1}^n (X_i - \mu).$$

Then, for every z , we have

$$\lim_{n \rightarrow \infty} P(Z_n \leq z) = P(Z \leq z),$$

where $Z \sim \mathcal{N}(0, 1)$.

Corollary (Normal approximation of a binomial) Let $X \sim \text{Bin}(n, p)$ with n large. Then S_n can be approximated by $Z \sim \mathcal{N}(np, np(1-p))$.

Remark (De Moivre-Laplace 1/2 approximation) Let $X \sim \text{Bin}(n, p)$ then $P(X = i) = P(i - \frac{1}{2} \leq X \leq i + \frac{1}{2})$ and we can use the CLT to approximate the PMF of X .

Definition (Conditional variance as a random variable) Given

random variables X, Y the conditional variance $\text{Var}(X|Y)$ is the random variable that takes the value $\text{Var}(X|Y = y)$ whenever $Y = y$.

Theorem (Law of total variance)

$$\text{Var}(X) = E[\text{Var}(X|Y)] + \text{Var}(E[X|Y]).$$

Proposition (Sum of a random number of independent r.v.)

Let N be a nonnegative integer random variable. Let X_1, X_2, \dots, X_N be i.i.d. random variables.

$$\text{Let } Y = \sum_i X_i. \text{ Then}$$

$$E[Y] = E[N]E[X],$$

$$\text{Var}(Y) = E[N]\text{Var}(X) + (E[X])^2\text{Var}(N).$$

CONVERGENCE OF RANDOM VARIABLES
Inequalities, convergence, and the Weak Law of Large Numbers

Theorem (Markov inequality) Given a random variable $X \geq 0$ and, for every $a > 0$ we have

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

Theorem (Chebyshev inequality) Given a random variable X with $E[X] = \mu$ and $\text{Var}(X) = \sigma^2$, for every $\epsilon > 0$ we have

$$P(|X - \mu| \geq \epsilon) \leq \frac{\sigma^2}{\epsilon^2}.$$

Theorem (Weak Law of Large Number (WLLN)) Given a sequence of i.i.d. random variables $\{X_1, X_2, \dots\}$ with $E[X_i] = \mu$ and $\text{Var}(X_i) = \sigma^2$, we define

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

for every $\epsilon > 0$ we have

$$\lim_{n \rightarrow \infty} P(|M_n - \mu| \geq \epsilon) = 0.$$

Definition (Convergence in probability) A sequence of random variables $\{Y_i\}$ converges in probability to the random variable Y if

$$\lim_{n \rightarrow \infty} P(|Y_i - Y| \geq \epsilon) = 0,$$

for every $\epsilon > 0$.

Properties (Properties of convergence in probability) If $X_n \rightarrow a$ and $Y_n \rightarrow b$ in probability, then

$$\bullet \quad X_n + Y_n \rightarrow a + b.$$

$$\bullet \quad \text{If } g \text{ is a continuous function, then } g(X_n) \rightarrow g(a).$$

$$\bullet \quad E[X_n] \text{ does not always converge to } a.$$

Super VIP CheatSheet: Machine Learning

Afshine AMIDI and Shervine AMIDI

October 6, 2018

Contents

1 Supervised Learning	4	Machine Learning Tips and Tricks	10
1.1 Introduction to Supervised Learning	4.1 Metrics	10	
1.2 Notations and general concepts	4.1.1 Classification	10	
1.3 Linear models	4.1.2 Regression	10	
1.3.1 Linear regression	4.2 Model selection	11	
1.3.2 Classification and logistic regression	4.3 Diagnostics	11	
1.3.3 Generalized Linear Models			
1.4 Support Vector Machines			
1.5 Generative Learning			
1.5.1 Gaussian Discriminant Analysis			
1.5.2 Naive Bayes			
1.6 Tree-based and ensemble methods			
1.7 Other non-parametric approaches			
1.8 Learning Theory			
2 Unsupervised Learning	6		
2.1 Introduction to Unsupervised Learning			
2.2 Clustering			
2.2.1 Expectation-Maximization			
2.2.2 k -means clustering			
2.2.3 Hierarchical clustering			
2.2.4 Clustering assessment metrics			
2.3 Dimension reduction			
2.3.1 Principal component analysis			
2.3.2 Independent component analysis			
3 Deep Learning	8		
3.1 Neural Networks			
3.2 Convolutional Neural Networks			
3.3 Recurrent Neural Networks			
3.4 Reinforcement Learning and Control			

1 Supervised Learning

1.1 Introduction to Supervised Learning

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

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

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

1.2 Notations and general concepts

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 $h_\theta(x^{(i)})$.

Loss function – A loss function is a function $L : (z, y) \in \mathbb{R} \times Y \mapsto L(z, y) \in \mathbb{R}$ 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:

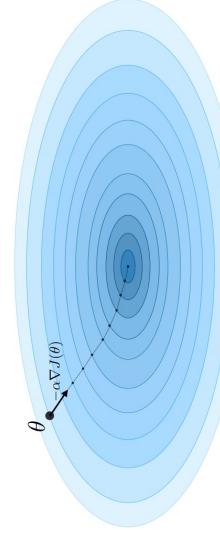
Least squared	Logistic	Hinge	Cross-entropy
$\frac{1}{2}(y - z)^2$	$\log(1 + \exp(-yz))$	$\max(0, 1 - yz)$	$-\left[y \log(z) + (1 - y) \log(1 - z)\right]$
Linear regression	Logistic regression	SVM	Neural Network

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:

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

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:

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



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.

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:

$$\theta^{\text{opt}} = \arg \max_{\theta} L(\theta)$$

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:

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

1.3 Linear models

1.3.1 Linear regression

We assume here that $y|z; \theta \sim \mathcal{N}(\mu, \sigma^2)$

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

$$\theta = (X^T X)^{-1} X^T y$$

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:

$$\forall j, \quad \theta_j \leftarrow \theta_j + \alpha \sum_{i=1}^m [y^{(i)} - h_\theta(x^{(i)})] x_j^{(i)}$$

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

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 $\tau \in \mathbb{R}$ as:

$$w^{(i)}(x) = \exp\left(-\frac{(x^{(i)} - x)^2}{2\tau^2}\right)$$

1.3.2 Classification and logistic regression

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

$$\forall z \in \mathbb{R}, \quad g(z) = \frac{1}{1 + e^{-z}} \in]0, 1[$$

Logistic regression – We assume here that $y|x; \theta \sim \text{Bernoulli}(\phi)$. We have the following form:

$$\phi = p(y=1|x; \theta) = \frac{1}{1 + \exp(-\theta^T x)} = g(\theta^T x)$$

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

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 $\theta_K = 0$, which makes the Bernoulli parameter ϕ_i of each class i equal to:

$$\phi_i = \frac{\exp(\theta_i^T x)}{\sum_{j=1}^K \exp(\theta_j^T x)}$$

1.3.3 Generalized Linear Models

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 $a(\eta)$ as follows:

$$p(y; \eta) = b(y) \exp(\eta T(y) - a(\eta))$$

Remark: we will often have $T(y) = y$. Also, $\exp(-a(\eta))$ 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	$T(y)$	$a(\eta)$	$b(y)$
Bernoulli	$\log\left(\frac{\phi}{1-\phi}\right)$	y	$\log(1 + \exp(\eta))$	1
Gaussian	μ	y	$\frac{\eta^2}{2}$	$\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{\eta^2}{2}\right)$
Poisson	$\log(\lambda)$	y	e^η	$\frac{1}{y!}$
Geometric	$\log(1 - \phi)$	y	$\log\left(\frac{e^\eta}{1 - e^\eta}\right)$	1

Assumptions of GLMs – Generalized Linear Models (GLM) aim at predicting a random variable y as a function fo $x \in \mathbb{R}^{n+1}$ and rely on the following 3 assumptions:

$$(1) \quad [y|x; \theta \sim \text{ExpFamily}(\eta)] \quad (2) \quad [h_\theta(x) = E[y|x; \theta]] \quad (3) \quad [\eta = \theta^T x]$$

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

1.4 Support Vector Machines

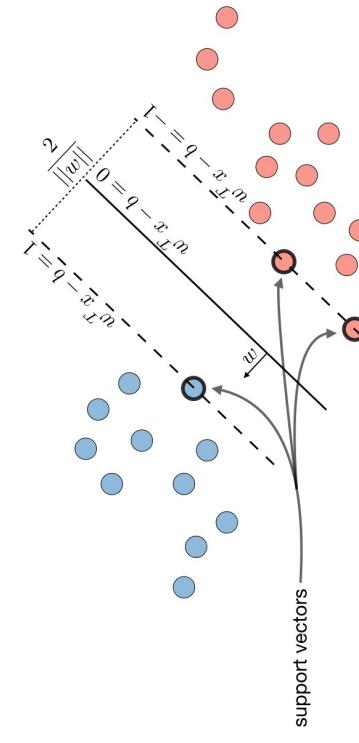
The goal of support vector machines is to find the line that maximizes the minimum distance to the line.

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

$$h(x) = \text{sign}(w^T x - b)$$

where $(w, b) \in \mathbb{R}^n \times \mathbb{R}$ is the solution of the following optimization problem:

$$\min \frac{1}{2} \|w\|^2 \quad \text{such that} \quad y^{(i)}(w^T x^{(i)} - b) \geq 1$$



Remark: the line is defined as $w^T x - b = 0$.

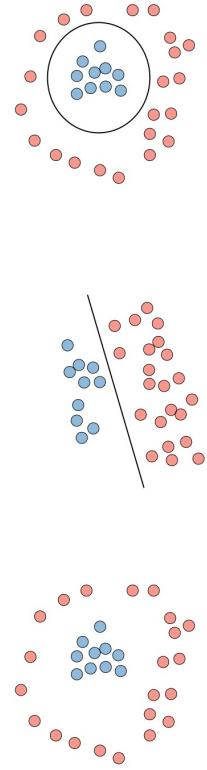
Hinge loss – The hinge loss is used in the setting of SVMs and is defined as follows:

$$L(z, y) = [1 - yz]_+ = \max(0, 1 - yz)$$

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

$$K(x, z) = \phi(x)^T \phi(z)$$

In practice, the kernel K defined by $K(x, z) = \exp\left(-\frac{\|x - z\|^2}{2\sigma^2}\right)$ is called the Gaussian kernel and is commonly used.



Non-linear separability → Use of a kernel mapping ϕ → Decision boundary in the original space

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.

□ **Lagrangian** – We define the Lagrangian $\mathcal{L}(w, b)$ as follows:

$$\mathcal{L}(w, b) = f(w) + \sum_{i=1}^l \beta_i h_i(w)$$

Remark: the coefficients β_i are called the Lagrange multipliers.

1.5 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.

1.5.1 Gaussian Discriminant Analysis

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

$$y \sim \text{Bernoulli}(\phi)$$

$$x|y = 1 \sim \mathcal{N}(\mu_1, \Sigma)$$

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

$\hat{\phi}$	$\hat{\mu}_j$ ($j = 0, 1$)	$\hat{\Sigma}$
$\frac{1}{m} \sum_{i=1}^m \mathbf{1}_{\{y^{(i)}=1\}}$	$\frac{\sum_{i=1}^m \mathbf{1}_{\{y^{(i)}=j\}} x^{(i)}}{\sum_{i=1}^m \mathbf{1}_{\{y^{(i)}=j\}}}$	$\frac{1}{m} \sum_{i=1}^m (x^{(i)} - \mu_{y^{(i)}})(x^{(i)} - \mu_{y^{(i)}})^T$

1.5.2 Naive Bayes

□ **Assumption** – The Naive Bayes model supposes that the features of each data point are all independent:

$$P(x|y) = P(x_1, x_2, \dots | y) = P(x_1|y)P(x_2|y)\dots = \prod_{i=1}^n P(x_i|y)$$

□ **Solutions** – Maximizing the log-likelihood gives the following solutions, with $k \in \{0, 1\}$,

$$P(y = k) = \frac{1}{m} \times \#\{j | y^{(j)} = k\}$$

$$P(x_i = l | y = k) = \frac{\#\{j | y^{(j)} = k\}}{\#\{j | y^{(j)} = k\}}$$

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

1.6 Tree-based and ensemble methods

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

□ **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.

□ **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.

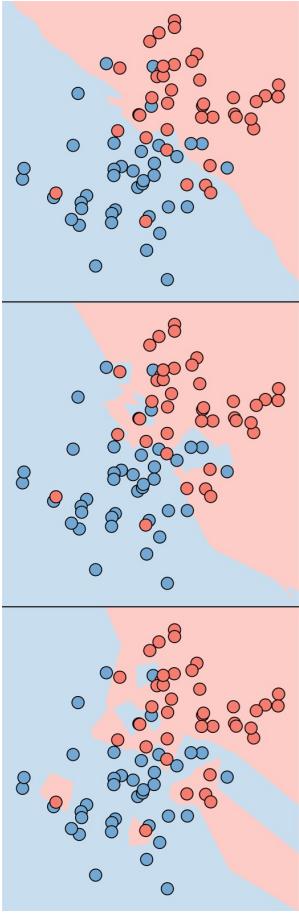
□ **Boosting** – The idea of boosting methods is to combine several weak learners to form a stronger one. The main ones are summed up in the table below:

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

1.7 Other non-parametric approaches

□ **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.

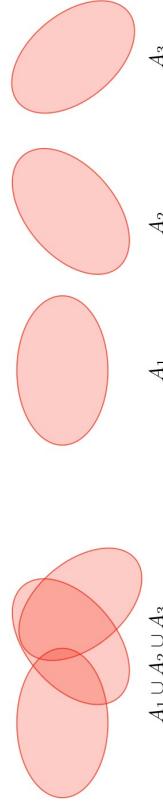
Remark: The higher the parameter k , the higher the bias, and the lower the variance.



1.8 Learning Theory

□ **Union bound** – Let A_1, \dots, A_k be k events. We have:

$$P(A_1 \cup \dots \cup A_k) \leq P(A_1) + \dots + P(A_k)$$



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

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

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

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

$$\hat{\epsilon}(h) = \frac{1}{m} \sum_{i=1}^m \mathbb{1}_{\{h(x^{(i)}) \neq y^{(i)}\}}$$

□ **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

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

2 Unsupervised Learning

2.1 Introduction to Unsupervised Learning

- **Motivation** – The goal of unsupervised learning is to find hidden patterns in unlabeled data $\{x^{(1)}, \dots, x^{(m)}\}$.
- **Jensen's inequality** – Let f be a convex function and X a random variable. We have the following inequality:

$$E[f(X)] \geq f(E[X])$$

2.2 Clustering

2.2.1 Expectation-Maximization

- **Latent variables** – Latent variables are hidden/unobserved variables that make estimation problems difficult, and are often denoted z . Here are the most common settings where there are latent variables:

Setting	Latent variable z	$x z$	Comments
Mixture of k Gaussians	Multinomial(ϕ)	$N(\mu_j, \Sigma_j)$	$\mu_j \in \mathbb{R}^n, \phi \in \mathbb{R}^k$
Factor analysis	$\mathcal{N}(0, I)$	$N(\mu + \Lambda z, \psi)$	$\mu_j \in \mathbb{R}^n$

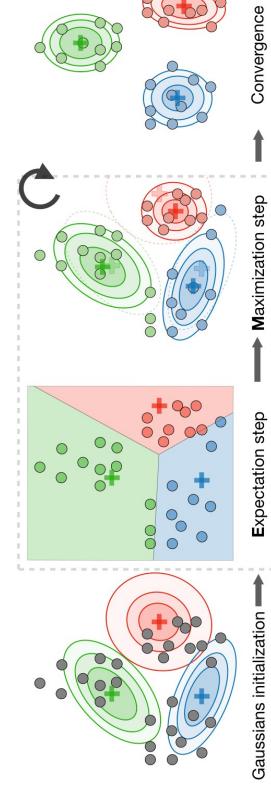
□ **Algorithm** – The Expectation-Maximization (EM) algorithm gives an efficient method at estimating the parameter θ through maximum likelihood estimation by repeatedly constructing a lower-bound on the likelihood (E-step) and optimizing that lower bound (M-step) as follows:

- E-step: Evaluate the posterior probability $Q_i(z^{(i)})$ that each data point $x^{(i)}$ came from a particular cluster $z^{(i)}$ as follows:

$$Q_i(z^{(i)}) = P(z^{(i)}|x^{(i)}; \theta)$$

- M-step: Use the posterior probabilities $Q_i(z^{(i)})$ as cluster specific weights on data points $x^{(i)}$ to separately re-estimate each cluster model as follows:

$$\theta_i = \underset{\theta}{\operatorname{argmax}} \sum_i \int_{z^{(i)}} Q_i(z^{(i)}) \log \left(\frac{P(x^{(i)}, z^{(i)}; \theta)}{Q_i(z^{(i)})} \right) dz^{(i)}$$

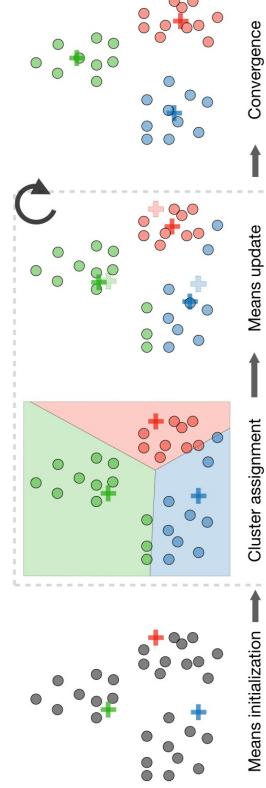


2.2.2 k -means clustering

We note $c^{(i)}$ the cluster of data point i and μ_j the center of cluster j .

- **Algorithm** – After randomly initializing the cluster centroids $\mu_1, \mu_2, \dots, \mu_k \in \mathbb{R}^n$, the k -means algorithm repeats the following step until convergence:

$$c^{(i)} = \arg \min_j \|x^{(i)} - \mu_j\|^2 \quad \text{and} \quad \mu_j = \frac{\sum_{i=1}^m 1_{\{c^{(i)}=j\}} x^{(i)}}{\sum_{i=1}^m 1_{\{c^{(i)}=j\}}}$$



□ **Distortion function** – In order to see if the algorithm converges, we look at the distortion function defined as follows:

$$J(c, \mu) = \sum_{i=1}^m \|x^{(i)} - \mu_{c^{(i)}}\|^2$$

2.2.3 Hierarchical clustering

- **Algorithm** – It is a clustering algorithm with an agglomerative hierarchical approach that build nested clusters in a successive manner.

- **Types** – There are different sorts of hierarchical clustering algorithms that aims at optimizing different objective functions, which is summed up in the table below:

Ward linkage	Average linkage	Complete linkage
Minimize within cluster distance	Minimize average distance between cluster pairs	Minimize maximum distance of between cluster pairs

2.2.4 Clustering assessment metrics

In an unsupervised learning setting, it is often hard to assess the performance of a model since we don't have the ground truth labels as was the case in the supervised learning setting.

- **Silhouette coefficient** – By noting a and b the mean distance between a sample and all other points in the same class, and between a sample and all other points in the next nearest cluster, the silhouette coefficient s for a single sample is defined as follows:

$$s = \frac{b - a}{\max(a, b)}$$

□ Calinski-Harabasz index – By noting k the number of clusters, B_k and W_k the between and within-clustering dispersion matrices respectively defined as

$$B_k = \sum_{j=1}^k n_{c(i)} (\mu_{c(i)} - \mu)(\mu_{c(i)} - \mu)^T, \quad W_k = \sum_{i=1}^m (x^{(i)} - \mu_{c(i)}) (x^{(i)} - \mu_{c(i)})^T$$

the Calinski-Harabasz index $s(k)$ indicates how well a clustering model defines its clusters, such that the higher the score, the more dense and well separated the clusters are. It is defined as follows:

$$s(k) = \frac{\text{Tr}(B_k)}{\text{Tr}(W_k)} \times \frac{N - k}{k - 1}$$

2.3 Dimension reduction

2.3.1 Principal component analysis

It is a dimension reduction technique that finds the variance maximizing directions onto which to project the data.

□ Eigenvector, eigenvalue – Given a matrix $A \in \mathbb{R}^{n \times n}$, λ is said to be an eigenvalue of A if there exists a vector $z \in \mathbb{R}^n \setminus \{0\}$, called eigenvector, such that we have:

$$Az = \lambda z$$

□ Spectral theorem – Let $A \in \mathbb{R}^{n \times n}$. If A is symmetric, then A is diagonalizable by a real orthogonal matrix $U \in \mathbb{R}^{n \times n}$. By noting $\Lambda = \text{diag}(\lambda_1, \dots, \lambda_n)$, we have:

$$\exists \Lambda \text{ diagonal}, \quad A = U \Lambda U^T$$

Remark: the eigenvector associated with the largest eigenvalue is called principal eigenvector of matrix A .

□ Algorithm – The Principal Component Analysis (PCA) procedure is a dimension reduction technique that projects the data on k dimensions by maximizing the variance of the data as follows:

- Step 1: Normalize the data to have a mean of 0 and standard deviation of 1.

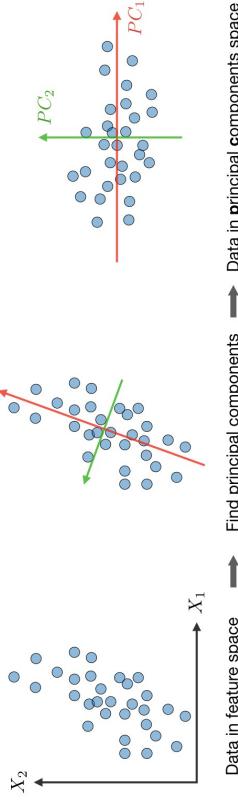
$$x_j^{(i)} \leftarrow \frac{x_j^{(i)} - \mu_j}{\sigma_j} \quad \text{where} \quad \mu_j = \frac{1}{m} \sum_{i=1}^m x_j^{(i)} \quad \text{and} \quad \sigma_j^2 = \frac{1}{m} \sum_{i=1}^m (x_j^{(i)} - \mu_j)^2$$

- Step 2: Compute $\Sigma = \frac{1}{m} \sum_{i=1}^m x^{(i)} x^{(i)T} \in \mathbb{R}^{n \times n}$, which is symmetric with real eigenvalues.

- Step 3: Compute $u_1, \dots, u_k \in \mathbb{R}^n$ the k orthogonal principal eigenvectors of Σ , i.e. the orthogonal eigenvectors of the k largest eigenvalues.

- Step 4: Project the data on $\text{span}_{\mathbb{R}}(u_1, \dots, u_k)$. This procedure maximizes the variance among all k -dimensional spaces.

□ Calinski-Harabasz index – By noting k the number of clusters, B_k and W_k the between and within-clustering dispersion matrices respectively defined as



2.3.2 Independent component analysis

It is a technique meant to find the underlying generating sources.

□ Assumptions – We assume that our data x has been generated by the n -dimensional source vector $s = (s_1, \dots, s_n)$, where s_i are independent random variables, via a mixing and non-singular matrix A as follows:

$$x = As$$

The goal is to find the unmixing matrix $W = A^{-1}$ by an update rule.

□ Bell and Sejnowski ICA algorithm – This algorithm finds the unmixing matrix W by following the steps below:

- Write the probability of $x = As = W^{-1}s$ as:

$$p(x) = \prod_{i=1}^n p_s(w_i^T x) \cdot |W|$$

Remark: write the log likelihood given our training data $\{x^{(i)}, i \in [1, m]\}$ and by noting g the sigmoid function as:

$$l(W) = \sum_{i=1}^m \left(\sum_{j=1}^n \log \left(g'(w_j^T x^{(i)}) \right) + \log |W| \right)$$

Therefore, the stochastic gradient ascent learning rule is such that for each training example $x^{(i)}$, we update W as follows:

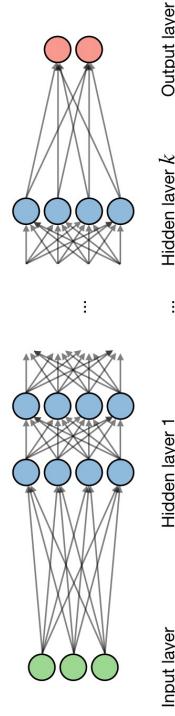
$$W \leftarrow W + \alpha \begin{pmatrix} \left(\frac{1 - 2g(w_1^T x^{(i)})}{1 - 2g(w_2^T x^{(i)})} \right) x^{(i)T} + (W^T)^{-1} \\ \vdots \\ \left(\frac{1 - 2g(w_n^T x^{(i)})}{1 - 2g(w_1^T x^{(i)})} \right) x^{(i)T} + (W^T)^{-1} \end{pmatrix}$$

3 Deep Learning

3.1 Neural Networks

Neural networks are a class of models that are built with layers. Commonly used types of neural networks include convolutional and recurrent neural networks.

□ Architecture – The vocabulary around neural networks architectures is described in the figure below:

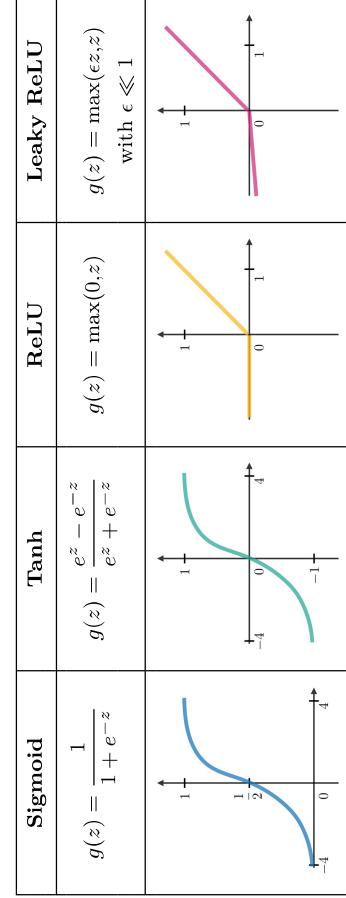


By noting i the i^{th} layer of the network and j the j^{th} hidden unit of the layer, we have:

$$z_j^{[i]} = w_j^{[i]T} x + b_j^{[i]}$$

where we note w , b , z the weight, bias and output respectively.

□ Activation function – Activation functions are used at the end of a hidden unit to introduce non-linear complexities to the model. Here are the most common ones:



As a result, the weight is updated as follows:

$$w \leftarrow w - \eta \frac{\partial L(z,y)}{\partial w}$$

□ Updating weights – In a neural network, weights are updated as follows:

- Step 1: Take a batch of training data.
- Step 2: Perform forward propagation to obtain the corresponding loss.
- Step 3: Backpropagate the loss to get the gradients.
- Step 4: Use the gradients to update the weights of the network.

□ Dropout – Dropout is a technique meant at preventing overfitting the training data by dropping out units in a neural network. In practice, neurons are either dropped with probability p or kept with probability $1 - p$.

3.2 Convolutional Neural Networks

□ Convolutional layer requirement – By noting W the input volume size, F the size of the convolutional layer neurons, P the amount of zero padding, then the number of neurons N that fit in a given volume is such that:

$$N = \frac{W - F + 2P}{S} + 1$$

□ Batch normalization – It is a step of hyperparameter γ, β that normalizes the batch $\{x_i\}$. By noting μ_B, σ_B^2 the mean and variance of that we want to correct to the batch, it is done as follows:

$$x_i \leftarrow \gamma \frac{x_i - \mu_B}{\sqrt{\sigma_B^2 + \epsilon}} + \beta$$

It is usually done after a fully connected/convolutional layer and before a non-linearity layer and aims at allowing higher learning rates and reducing the strong dependence on initialization.

□ Cross-entropy loss – In the context of neural networks, the cross-entropy loss $L(z,y)$ is commonly used and is defined as follows:

$$L(z,y) = - \left[y \log(z) + (1-y) \log(1-z) \right]$$

□ Types of gates – Here are the different types of gates that we encounter in a typical recurrent neural network:

Input gate	Forget gate	Output gate	Gate
Write to cell or not?	Erase a cell or not?	Reveal a cell or not?	How much writing?

□ LSTM – A long short-term memory (LSTM) network is a type of RNN model that avoids the vanishing gradient problem by adding 'forget' gates.

3.4 Reinforcement Learning and Control

The goal of reinforcement learning is for an agent to learn how to evolve in an environment.

□ Markov decision processes – A Markov decision process (MDP) is a 5-tuple $(S, A, \{P_{sa}\}, \gamma, R)$ where:

- S is the set of states
- A is the set of actions
- $\{P_{sa}\}$ are the state transition probabilities for $s \in S$ and $a \in A$
- $\gamma \in [0, 1]$ is the discount factor
- $R : S \times A \rightarrow \mathbb{R}$ or $R : S \rightarrow \mathbb{R}$ is the reward function that the algorithm wants to maximize

□ Policy – A policy π is a function $\pi : S \rightarrow A$ that maps states to actions.

Remark: we say that we execute a given policy π if given a state s we take the action $a = \pi(s)$.

□ Value function – For a given policy π and a given state s , we define the value function V^π as follows:

$$V^\pi(s) = E \left[R(s_0) + \gamma R(s_1) + \gamma^2 R(s_2) + \dots | s_0 = s, \pi \right]$$

□ Bellman equation – The optimal Bellman equations characterizes the value function V^{π^*} of the optimal policy π^* :

$$V^{\pi^*}(s) = R(s) + \max_{a \in A} \gamma \sum_{s' \in S} P_{sa}(s') V^{\pi^*}(s')$$

Remark: we note that the optimal policy π^ for a given state s is such that:*

$$\pi^*(s) = \arg\max_{a \in A} \sum_{s' \in S} P_{sa}(s') V^*(s')$$

□ Value iteration algorithm – The value iteration algorithm is in two steps:

- We initialize the value:

$$V_0(s) = 0$$

- We iterate the value based on the values before:

$$V_{i+1}(s) = R(s) + \max_{a \in A} \left[\sum_{s' \in S} \gamma P_{sa}(s') V_i(s') \right]$$

4 Machine Learning Tips and Tricks

4.1 Metrics

Given a set of data points $\{x^{(1)}, \dots, x^{(m)}\}$, where each $x^{(i)}$ has n features, associated to a set of outcomes $\{y^{(1)}, \dots, y^{(m)}\}$, we want to assess a given classifier that learns how to predict y from x .

4.1.1 Classification

In a context of a binary classification, here are the main metrics that are important to track to assess the performance of the model.

□ Confusion matrix – The confusion matrix is used to have a more complete picture when assessing the performance of a model. It is defined as follows:

		Predicted class	
		+	-
Actual class	+	TP True Positives	FN False Negatives
	-	FP False Positives	TN True Negatives

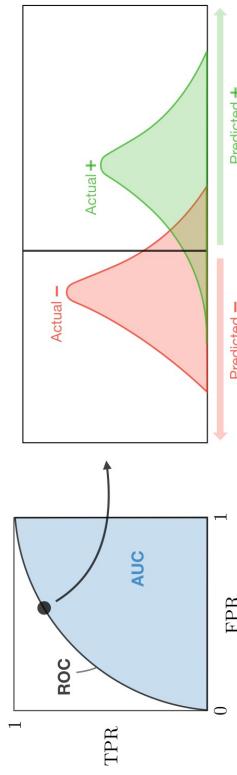
□ Main metrics – The following metrics are commonly used to assess the performance of classification models:

Metric	Formula	Interpretation
Accuracy	$\frac{TP + TN}{TP + TN + FP + FN}$	Overall performance of model
Precision	$\frac{TP}{TP + FP}$	How accurate the positive predictions are
Recall	$\frac{TP}{TP + FN}$	Coverage of actual positive sample
Sensitivity	$\frac{TP}{TP + FN}$	Coverage of actual negative sample
Specificity	$\frac{TN}{TN + FP}$	
F1 score	$\frac{2TP}{2TP + FP + FN}$	Hybrid metric useful for unbalanced classes

□ ROC – The receiver operating curve, also noted ROC, is the plot of TPR versus FPR by varying the threshold. These metrics are summed up in the table below:

Metric	Formula	Equivalent
True Positive Rate TPR	$\frac{TP}{TP + FN}$	Recall, sensitivity
False Positive Rate FPR	$\frac{FP}{TN + FP}$	1-specificity

□ AUC – The area under the receiving operating curve, also noted AUC or AUROC, is the area below the ROC as shown in the following figure:



4.1.2 Regression

□ Basic metrics – Given a regression model f , the following metrics are commonly used to assess the performance of the model:

Total sum of squares	Explained sum of squares	Residual sum of squares
$SS_{\text{tot}} = \sum_{i=1}^m (y_i - \bar{y})^2$	$SS_{\text{reg}} = \sum_{i=1}^m (f(x_i) - \bar{y})^2$	$SS_{\text{res}} = \sum_{i=1}^m (y_i - f(x_i))^2$

□ Coefficient of determination – The coefficient of determination, often noted R^2 or r^2 , provides a measure of how well the observed outcomes are replicated by the model and is defined as follows:

$$R^2 = 1 - \frac{SS_{\text{res}}}{SS_{\text{tot}}}$$

□ Main metrics – The following metrics are commonly used to assess the performance of regression models, by taking into account the number of variables n that they take into consideration:

Mallow's Cp	AIC	BIC	Adjusted R^2
$\frac{SS_{\text{res}} + 2(n+1)\hat{\sigma}^2}{m}$	$2[(n+2) - \log(L)]$	$\log(m)(n+2) - 2\log(L)$	$1 - \frac{(1 - R^2)(m-1)}{m-n-1}$

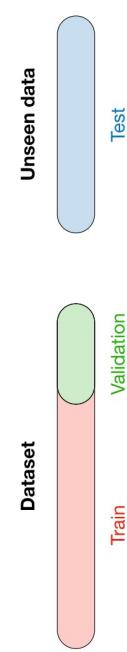
where L is the likelihood and $\hat{\sigma}^2$ is an estimate of the variance associated with each response.

4.2 Model selection

□ **Vocabulary** – When selecting a model, we distinguish 3 different parts of the data that we have as follows:

	Training set	Validation set	Testing set
- Model is trained	- Model is assessed	- Model gives predictions	- Unseen data
- Usually 80% of the dataset	- Usually 20% of the dataset	- Also called hold-out or development set	

Once the model has been chosen, it is trained on the entire dataset and tested on the unseen test set. These are represented in the figure below:



□ **Cross-validation** – Cross-validation, also noted CV, is a method that is used to select a model that does not rely too much on the initial training set. The different types are summed up in the table below:

	<i>k</i> -fold	Leave- <i>p</i> -out
- Training on $k - 1$ folds and assessment on the remaining one	- Training on $n - p$ observations and assessment on the p remaining ones	- Case $p = 1$ is called leave-one-out

The most commonly used method is called *k*-fold cross-validation and splits the training data into *k* folds to validate the model on one fold while training the model on the $k - 1$ other folds, all of this *k* times. The error is then averaged over the *k* folds and is named cross-validation error.

Fold	Dataset	Validation error	Cross-validation error
1		ϵ_1	
2		ϵ_2	$\frac{\epsilon_1 + \dots + \epsilon_k}{k}$
:		\vdots	
<i>k</i>			ϵ_k

□ **Regularization** – The regularization procedure aims at avoiding the model to overfit the data and thus deals with high variance issues. The following table sums up the different types of commonly used regularization techniques:

LASSO	Ridge	Elastic Net
<ul style="list-style-type: none"> - Shrinks coefficients to 0 - Good for variable selection 	<ul style="list-style-type: none"> Makes coefficients smaller 	<ul style="list-style-type: none"> Tradeoff between variable selection and small coefficients

- **Model selection** – Train model on training set, then evaluate on the development set, then pick best performance model on the development set, and retrain all of that model on the whole training set.
- **Overfitting** – The model prediction for given data points.
- **Bias** – The bias of a model is the difference between the expected prediction and the correct model that we try to predict for given data points.
- **Variance** – The variance of a model is the variability of the model prediction for given data points.

4.3 Diagnostics

- **Cross-validation** – Cross-validation, also noted CV, is a method that is used to select a model that does not rely too much on the initial training set. The different types are summed up in the table below:
- **Underfitting** – Training error close to test error
- **Just right** – Training error slightly lower than test error
- **Oversetting** – Low training error
- **High variance** – Training error much lower than test error

Underfitting	Just right	Oversetting
<ul style="list-style-type: none"> - High training error - Training error close to test error - High bias 	<ul style="list-style-type: none"> - Training error slightly lower than test error 	<ul style="list-style-type: none"> - Low training error - Training error much lower than test error

Classification			
Deep learning			
Remedies	<ul style="list-style-type: none"> - Complexify model - Add more features - Train longer 	<ul style="list-style-type: none"> - Regularize - Get more data 	

5 Refreshers

5.1 Probabilities and Statistics

5.1.1 Introduction to Probability and Combinatorics

□ **Sample space** – The set of all possible outcomes of an experiment is known as the sample space of the experiment and is denoted by S .

□ **Event** – Any subset E of the sample space is known as an event. That is, an event is a set consisting of possible outcomes of the experiment. If the outcome of the experiment is contained in E , then we say that E has occurred.

□ **Axioms of probability** – For each event E , we denote $P(E)$ as the probability of event E occurring. By noting E_1, \dots, E_n mutually exclusive events, we have the 3 following axioms:

$$(1) \quad 0 \leq P(E) \leq 1 \quad (2) \quad P(S) = 1 \quad (3) \quad P\left(\bigcup_{i=1}^n E_i\right) = \sum_{i=1}^n P(E_i)$$

□ **Permutation** – A permutation is an arrangement of r objects from a pool of n objects, in a given order. The number of such arrangements is given by $P(n, r)$, defined as:

$$P(n, r) = \frac{n!}{(n-r)!}$$

□ **Error analysis** – Error analysis is analyzing the root cause of the difference in performance between the current and the perfect models.

□ **Ablative analysis** – Ablative analysis is analyzing the root cause of the difference in performance between the current and the baseline models.

□ **Combination** – A combination is an arrangement of r objects from a pool of n objects, where the order does not matter. The number of such arrangements is given by $C(n, r)$, defined as:

$$C(n, r) = \frac{P(n, r)}{r!} = \frac{n!}{r!(n-r)!}$$

Remark: we note that for $0 \leq r \leq n$, we have $P(n, r) \geq C(n, r)$.

5.1.2 Conditional Probability

□ **Bayes' rule** – For events A and B such that $P(B) > 0$, we have:

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

Remark: we have $P(A \cap B) = P(A)P(B|A) = P(A|B)P(B)$.

□ **Partition** – Let $\{A_i, i \in [1, n]\}$ be such that for all i , $A_i \neq \emptyset$. We say that $\{A_i\}$ is a partition if we have:

$$\forall i \neq j, A_i \cap A_j = \emptyset \quad \text{and} \quad \bigcup_{i=1}^n A_i = S$$

Remark: for any event B in the sample space, we have $P(B) = \sum_{i=1}^n P(B|A_i)P(A_i)$.

□ Extended form of Bayes' rule – Let $\{A_i, i \in [1, n]\}$ be a partition of the sample space. We have:

$$P(A_k | B) = \frac{P(B | A_k) P(A_k)}{\sum_{i=1}^n P(B | A_i) P(A_i)}$$

□ Independence – Two events A and B are independent if and only if we have:

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

5.1.3 Random Variables

□ Random variable – A random variable, often noted X , is a function that maps every element in a sample space to a real line.

□ Cumulative distribution function (CDF) – The cumulative distribution function F , which is monotonically non-decreasing and is such that $\lim_{x \rightarrow -\infty} F(x) = 0$ and $\lim_{x \rightarrow +\infty} F(x) = 1$, is defined as:

$$F(x) = P(X \leq x)$$

Remark: we have $P(a < X \leq B) = F(b) - F(a)$.

□ Probability density function (PDF) – The probability density function f is the probability that X takes on values between two adjacent realizations of the random variable.

□ Relationships involving the PDF and CDF – Here are the important properties to know in the discrete (D) and the continuous (C) cases.

Case	CDF F	PDF f	Properties of PDF
(D)	$F(x) = \sum_{x_i \leq x} P(X = x_i)$	$f(x_j) = P(X = x_j)$	$0 \leq f(x_j) \leq 1$ and $\sum_j f(x_j) = 1$
(C)	$F(x) = \int_{-\infty}^x f(y)dy$	$f(x) = \frac{df}{dx}$	$f(x) \geq 0$ and $\int_{-\infty}^{+\infty} f(x)dx = 1$

□ Variance – The variance of a random variable, often noted $\text{Var}(X)$ or σ^2 , is a measure of the spread of its distribution function. It is determined as follows:

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

□ Standard deviation – The standard deviation of a random variable, often noted σ , is a measure of the spread of its distribution function which is compatible with the units of the actual random variable. It is determined as follows:

$$\sigma = \sqrt{\text{Var}(X)}$$

□ Expectation and Moments of the Distribution – Here are the expressions of the expected value $E[X]$, generalized expected value $E[g(X)]$, k^{th} moment $E[X^k]$ and characteristic function $\psi(\omega)$ for the discrete and continuous cases:

Case	$E[X]$	$E[g(X)]$	$E[X^k]$	$\psi(\omega)$
(D)	$\sum_{i=1}^n x_i f(x_i)$	$\sum_{i=1}^n g(x_i) f(x_i)$	$\sum_{i=1}^n x_i^k f(x_i)$	$\sum_{i=1}^n f(x_i) e^{i\omega x_i}$
(C)	$\int_{-\infty}^{+\infty} x f(x) dx$	$\int_{-\infty}^{+\infty} g(x) f(x) dx$	$\int_{-\infty}^{+\infty} x^k f(x) dx$	$\int_{-\infty}^{+\infty} f(x) e^{i\omega x} dx$

Remark: we have $e^{i\omega x} = \cos(\omega x) + i \sin(\omega x)$.

□ Revisiting the k^{th} moment – The k^{th} moment can also be computed with the characteristic function as follows:

$$E[X^k] = \frac{1}{i^k} \left[\frac{\partial^k \psi}{\partial \omega^k} \right]_{\omega=0}$$

□ Transformation of random variables – Let the variables X and Y be linked by some function. By noting f_X and f_Y the distribution function of X and Y respectively, we have:

$$f_Y(y) = f_X(x) \left| \frac{dx}{dy} \right|$$

□ Leibniz integral rule – Let g be a function of x and potentially c , and a, b boundaries that may depend on c . We have:

$$\frac{\partial}{\partial c} \left(\int_a^b g(x) dx \right) = \frac{\partial b}{\partial c} \cdot g(b) - \frac{\partial a}{\partial c} \cdot g(a) + \int_a^b \frac{\partial g}{\partial c}(x) dx$$

□ Chebyshev's inequality – Let X be a random variable with expected value μ and standard deviation σ . For $k, \sigma > 0$, we have the following inequality:

$$P(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}$$

5.1.4 Jointly Distributed Random Variables

□ Conditional density – The conditional density of X with respect to Y , often noted $f_{X|Y}$, is defined as follows:

$$f_{X|Y}(x|y) = \frac{f_{XY}(x,y)}{f_Y(y)}$$

□ Independence – Two random variables X and Y are said to be independent if we have:

$$f_{XY}(x,y) = f_X(x)f_Y(y)$$

□ Marginal density and cumulative distribution – From the joint density probability function f_{XY} , we have:

Case	Marginal density	Cumulative function
(D)	$f_X(x_i) = \sum_j f_{XY}(x_i, y_j)$	$F_{XY}(x, y) = \sum_{x_i \leq x} \sum_{y_j \leq y} f_{XY}(x_i, y_j)$
(C)	$f_X(x) = \int_{-\infty}^{+\infty} f_{XY}(x, y) dy$	$F_{XY}(x, y) = \int_{-\infty}^x \int_{-\infty}^y f_{XY}(x', y') dx' dy'$

□ Distribution of a sum of independent random variables – Let $Y = X_1 + \dots + X_n$ with X_1, \dots, X_n independent. We have:

$$\psi_Y(\omega) = \prod_{k=1}^n \psi_{X_k}(\omega)$$

□ Covariance – We define the covariance of two random variables X and Y , that we note σ_{XY}^2 or more commonly $\text{Cov}(X, Y)$, as follows:

$$\text{Cov}(X, Y) \triangleq \sigma_{XY}^2 = E[(X - \mu_X)(Y - \mu_Y)] = E[XY] - \mu_X \mu_Y$$

□ Correlation – By noting σ_X, σ_Y the standard deviations of X and Y , we define the correlation between the random variables X and Y , noted ρ_{XY} , as follows:

$$\rho_{XY} = \frac{\sigma_{XY}^2}{\sigma_X \sigma_Y}$$

Remarks: For any X, Y , we have $\rho_{XY} \in [-1, 1]$. If X and Y are independent, then $\rho_{XY} = 0$.

□ Main distributions – Here are the main distributions to have in mind:

Type	Distribution	PDF	$\psi(\omega)$	$E[X]$	$\text{Var}(X)$
(D)	$X \sim \mathcal{B}(n, p)$ Binomial	$P(X = x) = \binom{n}{x} p^x q^{n-x}$ $x \in [0, n]$	$(pe^{i\omega} + q)^n$	np	npq
	$X \sim \text{Po}(\mu)$ Poisson	$P(X = x) = \frac{\mu^x}{x!} e^{-\mu}$ $x \in \mathbb{N}$	$e^{\mu(e^{i\omega}-1)}$	μ	μ
(C)	$X \sim \mathcal{U}(a, b)$ Uniform	$f(x) = \frac{1}{b-a}$ $x \in [a, b]$	$\frac{e^{i\omega b} - e^{i\omega a}}{(b-a)i\omega}$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
	$X \sim \mathcal{N}(\mu, \sigma)$ Gaussian	$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2} \left(\frac{x-\mu}{\sigma} \right)^2}$ $x \in \mathbb{R}$	$e^{i\omega\mu - \frac{1}{2}\omega^2\sigma^2}$	μ	σ^2
	$X \sim \text{Exp}(\lambda)$ Exponential	$f(x) = \lambda e^{-\lambda x}$ $x \in \mathbb{R}_+$	$\frac{1}{1 - \frac{i\omega}{\lambda}}$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$

5.1.5 Parameter estimation

□ Random sample – A random sample is a collection of n random variables X_1, \dots, X_n that are independent and identically distributed with X .

□ Estimator – An estimator $\hat{\theta}$ is a function of the data that is used to infer the value of an unknown parameter θ in a statistical model.

□ Bias – The bias of an estimator $\hat{\theta}$ is defined as being the difference between the expected value of the distribution of $\hat{\theta}$ and the true value, i.e.:

$$\text{Bias}(\hat{\theta}) = E[\hat{\theta}] - \theta$$

Remark: an estimator is said to be unbiased when we have $E[\hat{\theta}] = \theta$.

□ Sample mean and variance – The sample mean and the sample variance of a random sample are used to estimate the true mean μ and the true variance σ^2 of a distribution, are noted \bar{X} and s^2 respectively, and are such that:

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i \quad \text{and} \quad s^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$$

□ Central Limit Theorem – Let us have a random sample X_1, \dots, X_n following a given distribution with mean μ and variance σ^2 , then we have:

$$\bar{X} \underset{n \rightarrow +\infty}{\sim} \mathcal{N} \left(\mu, \frac{\sigma}{\sqrt{n}} \right)$$

5.2 Linear Algebra and Calculus

5.2.1 General notations

□ Vector – We note $x \in \mathbb{R}^n$ a vector with n entries, where $x_i \in \mathbb{R}$ is the i^{th} entry:

$$x = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \in \mathbb{R}^n$$

□ Matrix – We note $A \in \mathbb{R}^{m \times n}$ a matrix with m rows and n columns, where $A_{i,j} \in \mathbb{R}$ is the entry located in the i^{th} row and j^{th} column:

$$A = \begin{pmatrix} A_{1,1} & \dots & A_{1,n} \\ \vdots & & \vdots \\ A_{m,1} & \dots & A_{m,n} \end{pmatrix} \in \mathbb{R}^{m \times n}$$

Remark: the vector x defined above can be viewed as a $n \times 1$ matrix and is more particularly called a column-vector.

□ Identity matrix – The identity matrix $I \in \mathbb{R}^{n \times n}$ is a square matrix with ones in its diagonal and zero everywhere else:

$$I = \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \dots & 0 & 1 \end{pmatrix}$$

Remark: for all matrices $A \in \mathbb{R}^{n \times n}$, we have $A \times I = I \times A = A$.

□ Diagonal matrix – A diagonal matrix $D \in \mathbb{R}^{n \times n}$ is a square matrix with nonzero values in its diagonal and zero everywhere else:

$$D = \begin{pmatrix} d_1 & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & d_n \end{pmatrix}$$

Remark: we also note D as $\text{diag}(d_1, \dots, d_n)$.

5.2.2 Matrix operations

□ Vector-vector multiplication – There are two types of vector-vector products:

- inner product: for $x, y \in \mathbb{R}^n$, we have:

$$x^T y = \sum_{i=1}^n x_i y_i \in \mathbb{R}$$

- outer product: for $x \in \mathbb{R}^m, y \in \mathbb{R}^n$, we have:

$$xy^T = \begin{pmatrix} x_1 y_1 & \cdots & x_1 y_n \\ \vdots & \ddots & \vdots \\ x_m y_1 & \cdots & x_m y_n \end{pmatrix} \in \mathbb{R}^{m \times n}$$

□ Matrix-vector multiplication – The product of matrix $A \in \mathbb{R}^{m \times n}$ and vector $x \in \mathbb{R}^n$ is a vector of size \mathbb{R}^m , such that:

$$Ax = \begin{pmatrix} a_{r,1}^T x \\ \vdots \\ a_{r,m}^T x \end{pmatrix} = \sum_{i=1}^n a_{c,i} x_i \in \mathbb{R}^m$$

where $a_{r,i}^T$ are the vector rows and $a_{c,j}$ are the vector columns of A , and x_i are the entries of x .

□ Matrix-matrix multiplication – The product of matrices $A \in \mathbb{R}^{m \times n}$ and $B \in \mathbb{R}^{n \times p}$ is a matrix of size $\mathbb{R}^{m \times p}$, such that:

$$AB = \begin{pmatrix} a_{r,1}^T b_{c,1} & \cdots & a_{r,1}^T b_{c,p} \\ \vdots & \ddots & \vdots \\ a_{r,m}^T b_{c,1} & \cdots & a_{r,m}^T b_{c,p} \end{pmatrix} = \sum_{i=1}^n a_{c,i} b_{r,i}^T \in \mathbb{R}^{m \times p}$$

where $a_{r,i}^T, b_{r,i}^T$ are the vector rows and $a_{c,j}, b_{c,j}$ are the vector columns of A and B respectively.

□ Transpose – The transpose of a matrix $A \in \mathbb{R}^{m \times n}$, noted A^T , is such that its entries are flipped:

For $x \in V$, the most commonly used norms are summed up in the table below:

$$\boxed{\forall i, j, \quad A_{i,j}^T = A_{j,i}}$$

□ **Inverse** – The inverse of an invertible square matrix A is noted A^{-1} and is the only matrix such that:

$$\boxed{AA^{-1} = A^{-1}A = I}$$

Remark: for matrices A, B , we have $(AB)^T = B^T A^T$.
 $B^{-1} A^{-1}$

□ Trace – The trace of a square matrix A , noted $\text{tr}(A)$, is the sum of its diagonal entries:

$$\boxed{\text{tr}(A) = \sum_{i=1}^n A_{i,i}}$$

Remark: for matrices A, B , we have $\text{tr}(A^T) = \text{tr}(A)$ and $\text{tr}(AB) = \text{tr}(BA)$

□ Determinant – The determinant of a square matrix $A \in \mathbb{R}^{n \times n}$, noted $|A|$ or $\det(A)$ is expressed recursively in terms of $A_{\setminus i, \setminus j}$, which is the matrix A without its i^{th} row and j^{th} column, as follows:

$$\boxed{\det(A) = |A| = \sum_{j=1}^n (-1)^{i+j} A_{i,j} |A_{\setminus i, \setminus j}|}$$

Remark: A is invertible if and only if $|A| \neq 0$. Also, $|AB| = |A||B|$ and $|A^T| = |A|$.

5.2.3 Matrix properties

□ Symmetric decomposition – A given matrix A can be expressed in terms of its symmetric and antisymmetric parts as follows:

$$\boxed{A = \underbrace{\frac{A + A^T}{2}}_{\text{Symmetric}} + \underbrace{\frac{A - A^T}{2}}_{\text{Antisymmetric}}}$$

□ Norm – A norm is a function $N : V \rightarrow [0, +\infty[$ where V is a vector space, and such that for all $x, y \in V$, we have:

- $N(x+y) \leq N(x) + N(y)$
- $N(ax) = |a|N(x)$ for a scalar
- if $N(x) = 0$, then $x = 0$