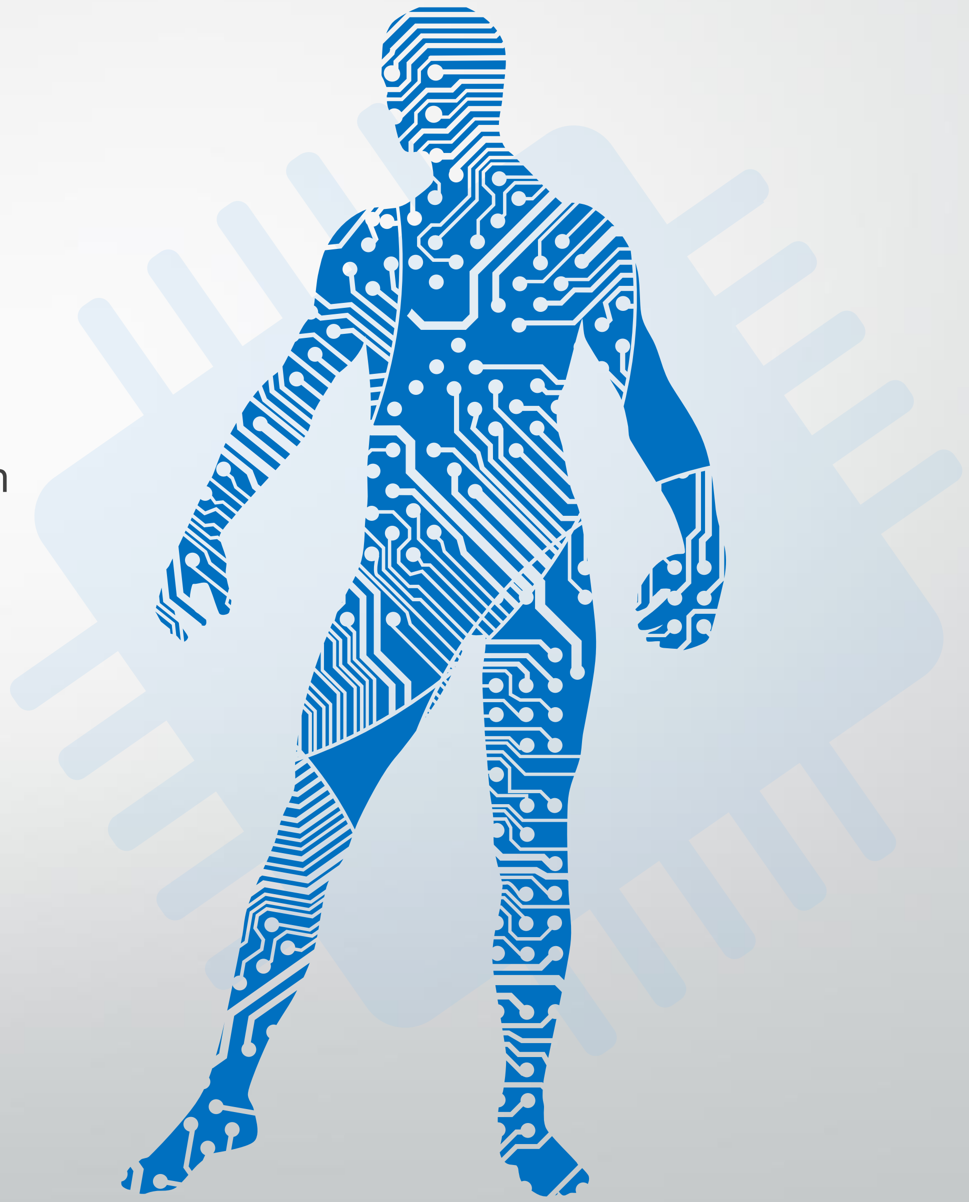
Machine Learning

PROBABILITY

AGENDA

01 Probability Review

Introduction, Random Variables, Bayes Theorem



Probability

DEFINITIONS

CONDITIONAL PROBABILITY

PROBABILITY
DEFINITIONS

Sample space Ω: set of all results $\boldsymbol{\omega}$ of a random experiment. Where each $\boldsymbol{\omega} \in \boldsymbol{\Omega}$ is defined as a complete description of the real state of the world after the experiment.

Event space $\mathcal{F}$: set whose elements $A \in \mathcal{F}$ (called events) are subsets of $\boldsymbol{\Omega}$.

Probabilistic measure: it is a function $P: \mathcal{F} \rightarrow \mathbb{R}$ that satisfies the **3 axioms of probability.**

1) $P(A) \geq 0, \forall A \in \mathcal{F}$
2) $P(\Omega) = 1$
3) If $A_{1,} A_{2,\ldots}$ are disjoint events, therefore:

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

PROBABILITY
DEFINITIONS

Example: we perform the experiment of tossing 2 coins.

- **Sample space** Ω:
- **Event** $\boldsymbol{E}$: the 1st coin results in heads
- **Probabilistic measure** $\boldsymbol{P}$: probability of event E (given the assumption that all outcomes are just as likely to happen).

PROBABILITY
DEFINITIONS

Answer: we perform the experiment of tossing 2 coins.

- **Sample space** Ω:

$$S = \{(H,H),(H,T),(T,H),(T,T)\}$$

- **Event** E: 1st coin results in heads

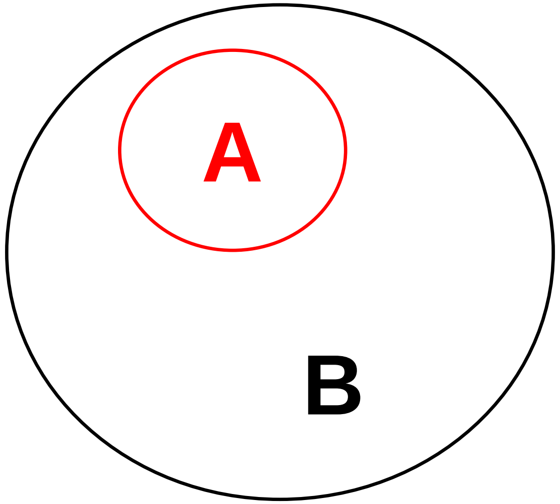
$$E = \{(H,H),(H,T)\}$$

- **Probabilistic measure** $\boldsymbol{P}$: the probability of event E would be

$$\mathrm{P}(E) = \frac{2}{4} = \frac{1}{2}$$

PROBABILITY
PROPERTIES

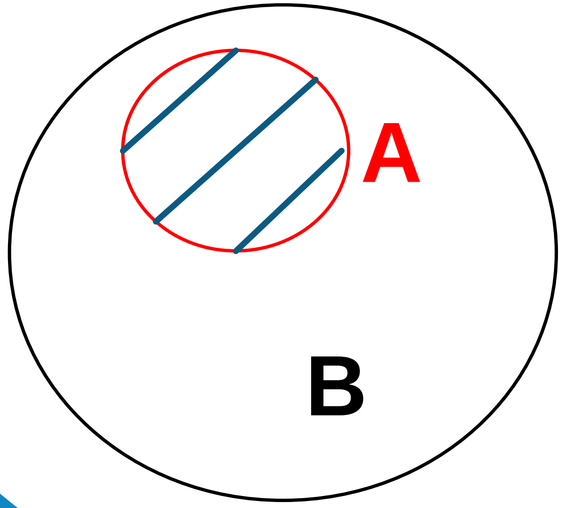
1 If $A \subseteq B \Longrightarrow P(A) \leq P(B)$.

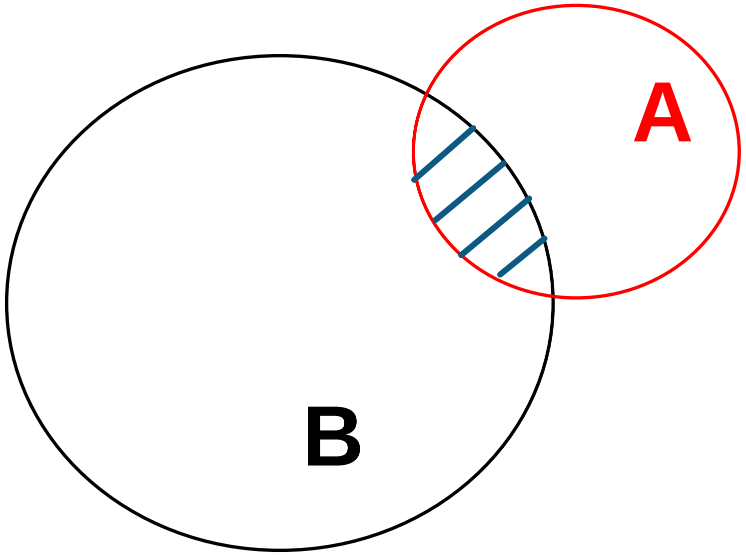


PROBABILITY
PROPERTIES

2 $P(A \cap B) \leq \min(P(A), P(B)).$

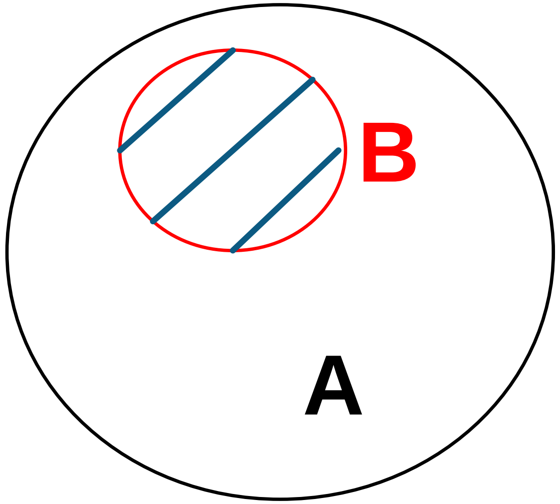
$P(A \cap B) = P(A)$





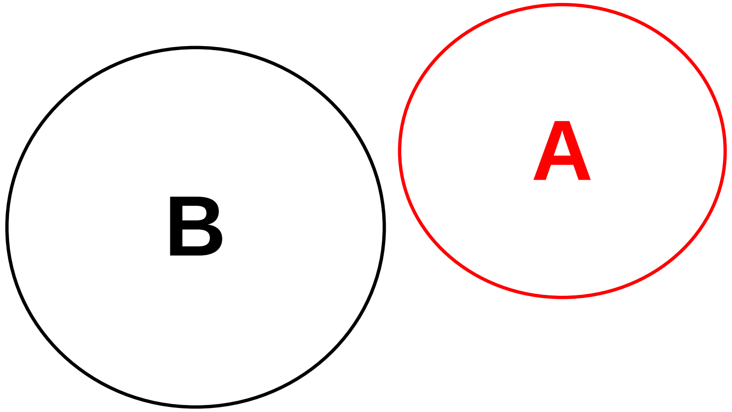
$P(A \cap B) < P(A)$
$P(A \cap B) < P(B)$

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

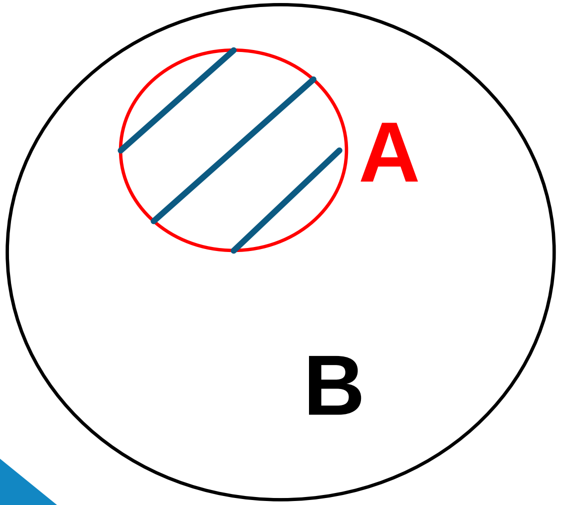


PROBABILITY
PROPERTIES

3 (Union Bound) $P(A \cup B) \leq P(A) + P(B)$.

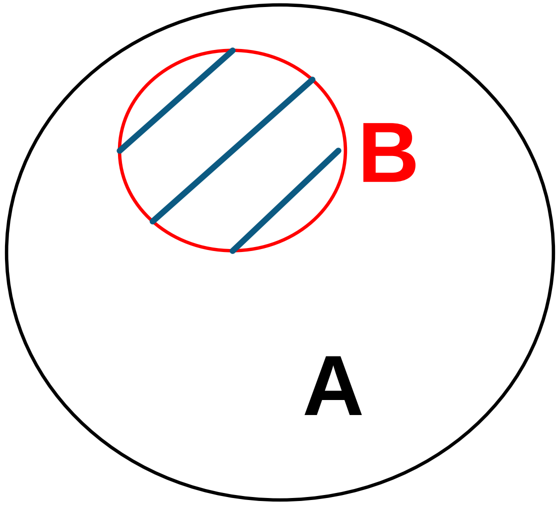


$P(A \cup B) = P(A)$



$P(A \cup B) = P(A) + P(B)$

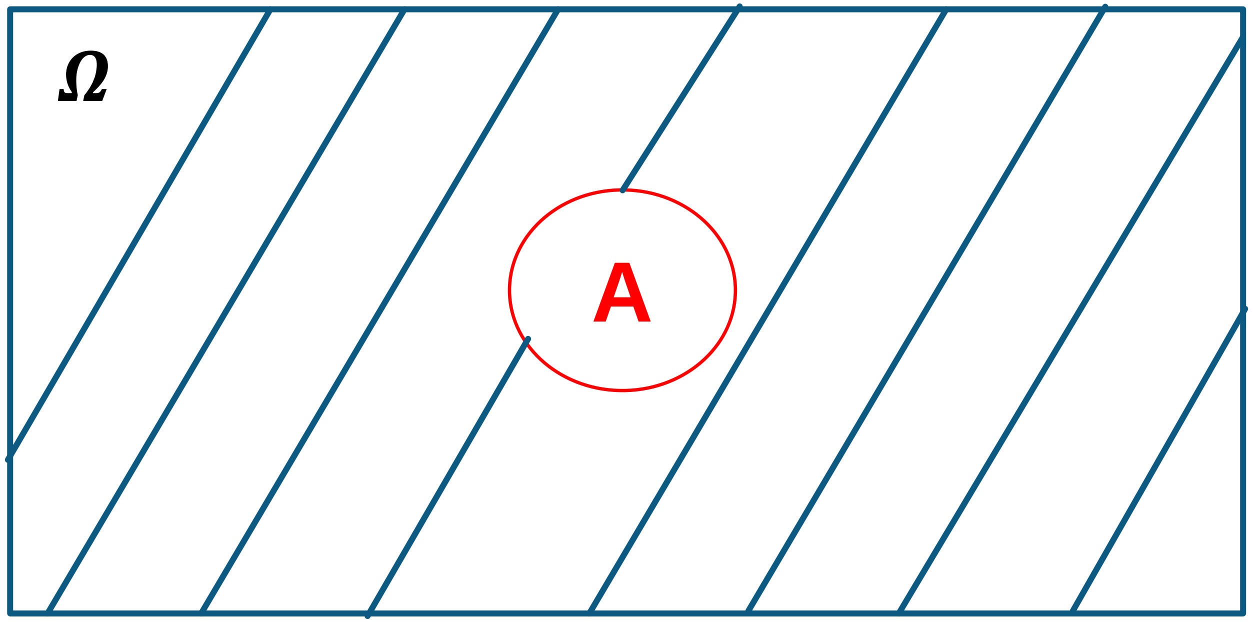
$P(A \cup B) = P(B)$



PROBABILIDAD

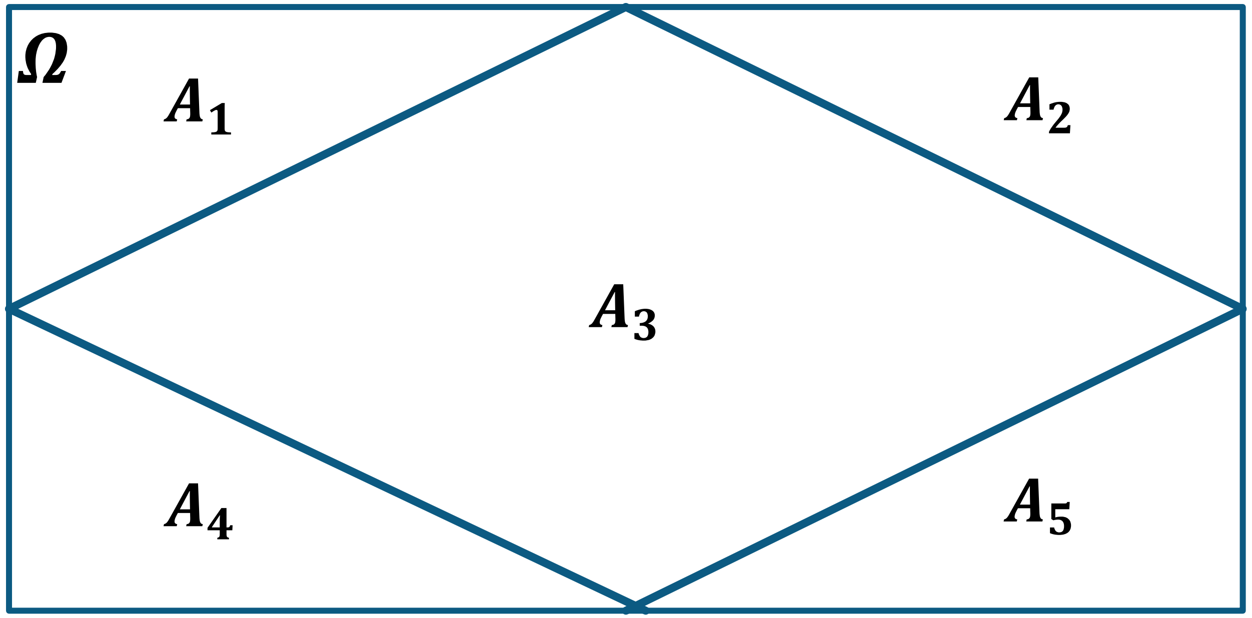
P R O P E R T I E S

4 $P(\Omega \setminus A) = 1 - P(A).$



PROBABILITY
PROPERTIES

5 (Law of Total Probability) If $A_1, \ldots, A_k$ are a set of disjoint events such that $\cup_{i=1}^{k} A_i = \Omega$, then $\sum_{i=1}^{k} P(A_k) = 1$.



PROBABILITY

CONDITIONAL PROBABILITY

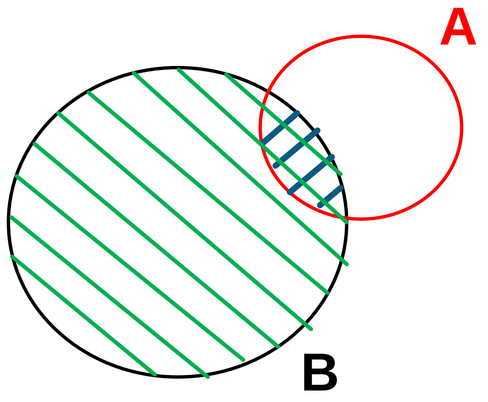
Conditional probability: the probability measure $P\ (A\ /\ B)$ defines the probability of an event A after observing the occurrence of event B with a probability $P\ (B) \neq 0$.

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

Independence: 2 events A and B are independent if the observation of event B doesn't affect probability of event A.

$$\boldsymbol{P(A/B) = P(A)}$$

$$\therefore \boldsymbol{P(A \cap B) = P(A)P(B)}$$



PROBABILITY

CONDITIONAL PROBABILITY

Example: a coin is tossed in the air 2 times. Assuming that all outcomes are equiprobable, what is the probability that both tosses were heads (H) given that:

a) 1st coin resulted in H?

b) At least one toss resulted in H?

PROBABILITY

CONDITIONAL PROBABILITY

Answers:

a) $P(both\ H\ /\ 1°\ H) = \frac{1}{2}$

b) $P(both\ H\ /\ at\ least\ 1\ H) = \frac{1}{3}$

PROBABILITY

RANDOM VARIABLES

PROBABILITY
RANDOM VARIABLES

Many times, we **DO NOT** care about the results $\boldsymbol{\omega \in \Omega}$ of an experiment. What we care about are **functions** of these results $\boldsymbol{X}$ $(\boldsymbol{\varpi})$.

Formally a **random variable** is a function:

$$X: \Omega \to \mathbb{R}$$

Discrete random variables	Continuous random variables
$\boldsymbol{X(\omega)}$ can take a **finite** amount of **values.**	$\boldsymbol{X(\omega)}$ can take an **infinite** amount of **values.**
$P(X = k) \coloneqq P(\{\varpi: X(\omega) = k\})$	$P(a \le X \le b) \coloneqq P(\{\varpi: (a \le X(\omega) \le b)\})$
Example: $\boldsymbol{X}$ $(\boldsymbol{\omega})$ is the number of heads that occur in a sequence of tosses ω.	**Example:** $\boldsymbol{X}$ $(\boldsymbol{\omega})$ is the radioactive decay time of a particle.

PROBABILITY

PROBABILITY FUNCTIONS

Different **measures of probability** are needed when considering **random variables.** For this, **probability functions** are defined:

1. Cumulative Distribution Function (discrete and continuous) → CDF.

2. Probability Mass Function(discrete)→ PMF.

3. Probability Density Function (continuous) → PDF.

NOTE: when the random variable X takes on a specific value, it is denoted by lowercase x.

PROBABILITY

CUMULATIVE DISTRIBUTION FUNCTION

The Cumulative Distribution Function is a function $F_X: \mathbb{R} \to [0,1]$ that specifies the probability measure as:

$$F_X(x) \triangleq P(X \leq x).$$

Intuitively it can be said that this function defines the probabilities of all events $A_i \in \Omega$ when $x \to \infty$

Properties:

$$0 \leq F_X(x) \leq 1.$$

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

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

$$x \leq y \Longrightarrow F_X(x) \leq F_X(y).$$

PROBABILITY

PROBABILITY MASS FUNCTION

The **Probability Mass Function** assigns a **probability measure** to each **value** that the **random variable** $\boldsymbol{X}$ can take.

$$p_X(x) \triangleq P(X = x).$$

Properties:

$$0 \le p_X(x) \le 1.$$

$$\sum_{x \in Val(X)} p_X(x) = 1.$$

$$\sum_{x \in A} p_X(x) = P(X \in A).$$

Where $Val\ (X)$ represents all the possible values that it can take

PROBABILITY

PROBABILITY DENSITY FUNCTION

The **Probability Density Function assigns a probability measure** to each **value** that the **random variable** X can take in a **continuous interval.**

It is formally defined as the **derivative** of the **Cumulative Distribution Function**:

$$f_X(x) \triangleq \frac{dF_X(x)}{dx}$$

Such function **may not exist.**

Properties:

$$f_X(x) \geq 0\,.$$

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

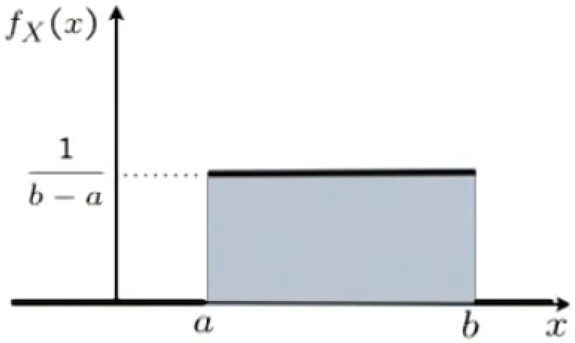
$$\int_{x \in A} f_X(x)dx = P(X \in A).$$

PROBABILITY
PRACTICAL EXAMPLES

Example:

Assuming X is a continuous random variable and its PDF is described by:

$$f_X(t) = \frac{1}{b-a}, \forall t \in [a,b]$$
$$f_X(t) = 0 \text{ de otra forma}$$



Find the graph of its CDF.

PROBABILITY

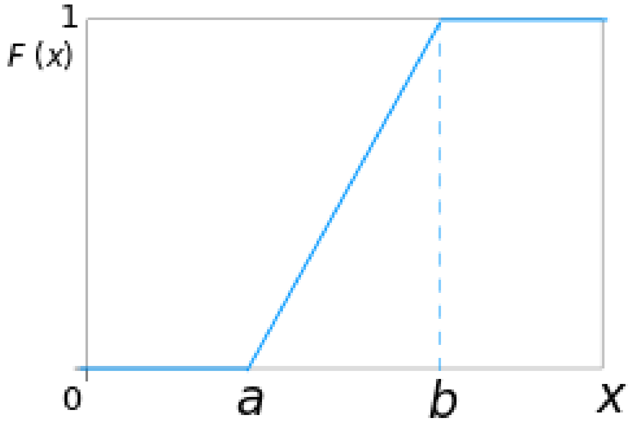
PRACTICAL EXAMPLES

Answer:

$$F_X(x) = P(X \leq x) = \int_{-\infty}^{x} f_X(t)dx$$

$$F_X(x) = \int_{-\infty}^{a} \frac{1}{b-a}dx + \int_{a}^{x} \frac{1}{b-a}dx + \int_{x}^{b} \frac{1}{b-a}dx$$

$$F_X(x) = 0 + \frac{x}{b-a} + 0$$



PROBABILITY

EXPECTED VALUE

We assume that $\boldsymbol{X}$ is a **discrete random variable** with a **PMF** $\boldsymbol{p_X(x)}$ and $\boldsymbol{g}: \mathbb{R} \to \mathbb{R}$ an arbitrary function. In this case, $\boldsymbol{g}\ (\boldsymbol{X})$ is considered a **random variable**, so the **expected value** of $\boldsymbol{g}\ (\boldsymbol{X})$ is defined as:

$$E[g(X)] \triangleq \sum_{x \in Val(X)} g(x) p_X(x).$$

PROBABILITY

EXPECTED VALUE

We assume that X is a **continuous random variable** with a **PDF $f_X(x)$,** therefore the **expected value** of $g\,(X)$ would be:

$$E[g(X)] \triangleq \int_{-\infty}^{\infty} g(x) f_X(x)\,dx.$$

PROBABILITY
EXPECTED VALUE

Intuitively what is calculated in the **expected value** is a **“weighted average”** of the **values** of $\boldsymbol{g}\ (\boldsymbol{x})$ where the **weights** are given either by $\boldsymbol{f_X}(\boldsymbol{x})$ **or** $\boldsymbol{p_X}(\boldsymbol{x})$**.**

NOTE: when $\boldsymbol{g}\ (\boldsymbol{x}) = \boldsymbol{x}$ the expected value of the random variable $\boldsymbol{X}$ would be the **arithmetic mean.**

Properties:

$E[a] = a$ for any constant $a \in \mathbb{R}$.

$E[af(X)] = aE[f(X)]$ for any constant $a \in \mathbb{R}$.

(Linearity of Expectation) $E[f(X) + g(X)] = E[f(X)] + E[g(X)]$.

For a discrete random variable X, $E[1\{X = k\}] = P(X = k)$.

PROBABILITY

VARIANCE

The **variance** of a **random variable** X is the **measure** of how **concentrated** the **distribution** of the **random variable** X is **around** the **mean.**

$$Var[X] \triangleq E[(X - E(X))^2]$$

PROBABILITY

V A R I A N C E

Homework:

Demonstrate the following equality

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

PROBABILITY

V A R I A N C E

Properties:

$$Var[a] = 0 \text{ for any constant } a \in \mathbb{R}.$$
$$Var[af(X)] = a^2 Var[f(X)] \text{ for any constant } a \in \mathbb{R}.$$

PROBABILITY

EXPECTED VALUE

Example:

Calculate the mean and variance of a random variable X with PDF $f_X(x) = 1, \forall\, x \in [0,1]$, 0 otherwise.

PROBABILITY

EXPECTED VALUE

Answer:

$$\boldsymbol{E[X] = \frac{1}{2}}$$

$$\boldsymbol{Var[X] = \frac{1}{12}}$$

PROBABILITY

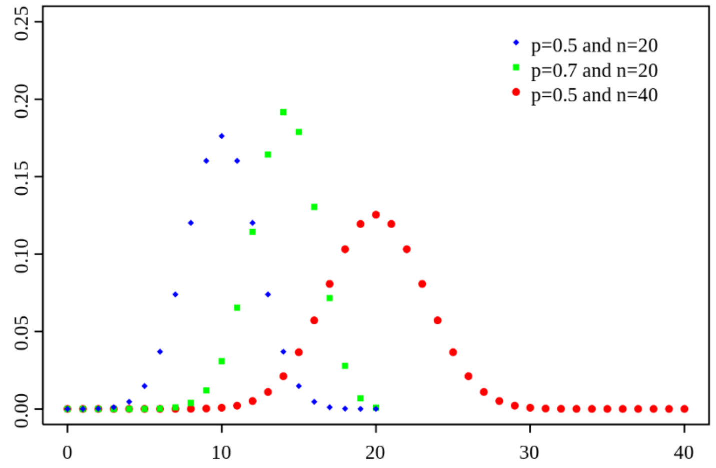
EXAMPLES OF RANDOM VARIABLES

PROBABILITY
EXAMPLES OF RANDOM VARIABLES

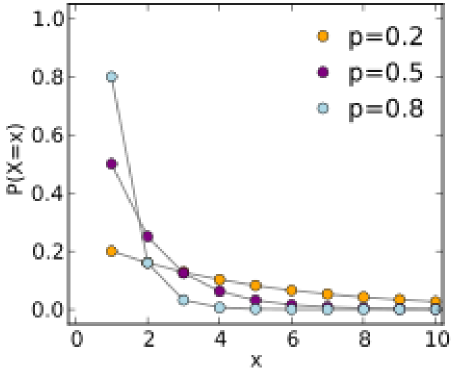
Discrete:

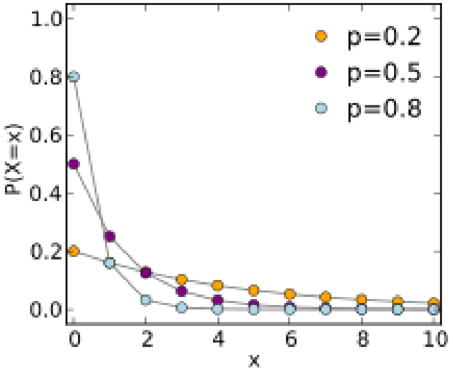
BERNOULLI $$p(x) = \begin{cases} p & \text{if } p = 1 \\ 1-p & \text{if } p = 0 \end{cases}$$

BINOMIAL $$p(x) = \binom{n}{x} p^x (1-p)^{n-x}$$



GEOMETRIC $$p(x) = p(1-p)^{x-1}$$





POISSON $$p(x) = e^{-\lambda} \frac{\lambda^x}{x!}$$

PROBABILITY

EXAMPLES OF RANDOM VARIABLES

Continuous:

UNIFORM

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{if } a \leq x \leq b \\ 0 & \text{otherwise} \end{cases}$$

EXPONENTIAL

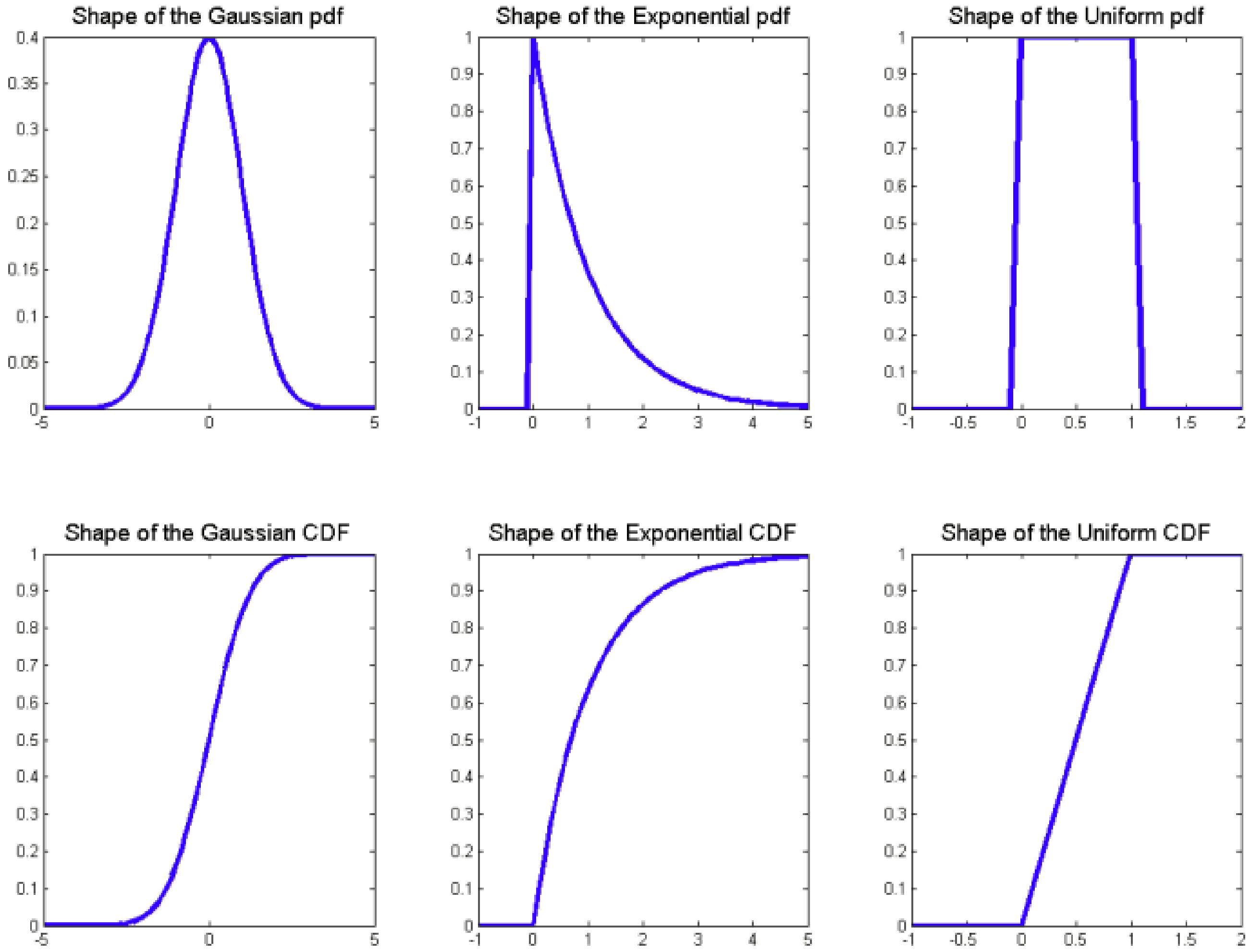
$$f(x) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

NORMAL

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

PROBABILITY

EXAMPLES OF RANDOM VARIABLES



PROBABILITY

EXAMPLES OF RANDOM VARIABLES

Homework:
Obtain the mean and variance of the 4 discrete distributions.

Distribution	PDF or PMF	Mean	Variance
$Bernoulli(p)$	$\begin{cases} p, & \text{if } x = 1 \\ 1-p, & \text{if } x = 0. \end{cases}$	p	$p(1-p)$
$Binomial(n, p)$	$\binom{n}{k} p^k (1-p)^{n-k}$ for $0 \leq k \leq n$	np	npq
$Geometric(p)$	$p(1-p)^{k-1}$ for $k = 1, 2, \ldots$	$\frac{1}{p}$	$\frac{1-p}{p^2}$
$Poisson(\lambda)$	$e^{-\lambda}\lambda^x/x!$ for $k = 1, 2, \ldots$	λ	λ

PROBABILITY

TWO RANDOM VARIABLES

PROBABILITY

JOINT AND MARGINAL CUMULATIVE DISTRIBUTIONS

If you want to know the **values** of **two random variables** $\boldsymbol{X}$ and $\boldsymbol{Y}$ **simultaneously**, you need the **cumulative joint distribution** of $\boldsymbol{X}$ and $\boldsymbol{Y}$:

$$F_{XY}(x,y) = P(X \leq x, Y \leq y)$$

The **distribution functions** $\boldsymbol{F_X(x)}$ and $\boldsymbol{F_Y(y)}$ are called **marginal cumulative distribution** functions

$$\begin{aligned} F_X(x) &= \lim_{y \to \infty} F_{XY}(x,y)dy \\ F_Y(y) &= \lim_{x \to \infty} F_{XY}(x,y)dx. \end{aligned}$$

PROBABILITY

JOINT AND MARGINAL CUMULATIVE DISTRIBUTIONS

Properties:

$$0 \leq F_{XY}(x,y) \leq 1.$$

$$\lim_{x,y\to\infty} F_{XY}(x,y) = 1.$$

$$\lim_{x,y\to-\infty} F_{XY}(x,y) = 0.$$

$$F_X(x) = \lim_{y\to\infty} F_{XY}(x,y).$$

PROBABILITY

JOINT AND MARGINAL MASS FUNCTIONS

If $\boldsymbol{X}$ and $\boldsymbol{Y}$ are **discrete random variables**, the **joint probability mass function** $p_{XY}: \mathbb{R}\text{x}\mathbb{R} \rightarrow [0,1]$ is defined by:

$$p_{XY}(x, y) = P(X = x, Y = y).$$

Properties:

$$0 \leq P_{XY}(x, y) \leq 1 \text{ for all } x, y.$$

$$\sum_{x \in Val(X)} \sum_{y \in Val(Y)} P_{XY}(x, y) = 1.$$

PROBABILITY

JOINT AND MARGINAL MASS FUNCTIONS

The **marginal probability mass functions** of ***X*** and ***Y*** are defined by:

$$p_X(x) = \sum_y p_{XY}(x, y).$$

$$p_Y(y) = \sum_x p_{XY}(x, y)$$

PROBABILITY

JOINT AND MARGINAL DENSITY FUNCTIONS

If $\boldsymbol{X}$ and $\boldsymbol{Y}$ are **continuous random variables**, with a **joint distribution** F_{XY} **differentiable throughout** the space, the **probability density function** is defined as:

$$f_{XY}(x, y) = \frac{\partial^2 F_{XY}(x, y)}{\partial x \partial y}.$$

$$\iint_{x \in A} f_{XY}(x, y) dx dy = P((X, Y) \in A).$$

PROBABILITY

JOINT AND MARGINAL DENSITY FUNCTIONS

The **marginal probability density functions** of ***X*** and ***Y*** would be:

$$f_X(x) = \int_{-\infty}^{\infty} f_{XY}(x, y)dy$$

$$f_Y(y) = \int_{-\infty}^{\infty} f_{XY}(x, y)dx$$

PROBABILITY

JOINT AND MARGINAL DENSITY FUNCTIONS

Example:

The **joint density function** of $\boldsymbol{X}$ and $\boldsymbol{Y}$ is given by:

$$f(x,y)=\begin{cases}2e^{-x}e^{-2y} & 0<x<\infty, \quad 0<y<\infty \\ & 0 \text{ otherwise}\end{cases}$$

Calculate $\boldsymbol{P(X>1, Y<1)}$

PROBABILITY

JOINT AND MARGINAL DENSITY FUNCTIONS

Answer:

$$P(X > 1, Y < 1) = e^{-1}(1 - e^{-2})$$

PROBABILITY

CONDITIONED DISTRIBUTIONS

The **conditional probability mass function** in the discrete case:

$$p_{Y|X}(y|x) = \frac{p_{XY}(x,y)}{p_X(x)}$$

The **conditional probability density function** in the continuous case:

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

PROBABILITY

EXPECTED VALUE

Discrete variables:

$$E[g(X,Y)] \triangleq \sum_{x \in Val(X)} \sum_{y \in Val(Y)} g(x,y) p_{XY}(x,y).$$

Continuous variables:

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

PROBABILITY

C O V A R I A N C E

Covariance is used to study the relationship between 2 random variables.

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

When $\boldsymbol{Cov}\,[\boldsymbol{X}, \boldsymbol{Y}] = \mathbf{0}$ it is said that $\mathbf{X}$ and $\mathbf{Y}$ are not correlated.

Homework:

Demonstrate that:

$$Cov[X,Y] = E[XY] - E[X]E[Y]$$

PROBABILITY

MULTIPLE RANDOM VARIABLES

PROBABILITY DISTRIBUTIONS

Assuming that we have n **continuous random variables** $X_1, X_2, \ldots, X_n$ we obtain the following **distributions**:

Cumulative joint probability function

$$F_{X_1,X_2,\ldots,X_n}(x_1, x_2, \ldots x_n) = P(X_1 \leq x_1, X_2 \leq x_2, \ldots, X_n \leq x_n)$$

Joint probability density function

$$f_{X_1,X_2,\ldots,X_n}(x_1, x_2, \ldots x_n) = \frac{\partial^n F_{X_1,X_2,\ldots,X_n}(x_1, x_2, \ldots x_n)}{\partial x_1 \ldots \partial x_n}$$

PROBABILITY DISTRIBUTIONS

Assuming that we have n **continuous random variables** $X_1, X_2, \ldots, X_n$ we obtain the following **distributions**:

Marginal probability density function of X_1

$$f_{X_1}(X_1) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_{X_1, X_2, \ldots, X_n}(x_1, x_2, \ldots x_n) dx_2 \ldots dx_n$$

Conditional probability density function

$$f_{X_1|X_2,\ldots,X_n}(x_1|x_2, \ldots x_n) = \frac{f_{X_1,X_2,\ldots,X_n}(x_1, x_2, \ldots x_n)}{f_{X_2,\ldots,X_n}(x_1, x_2, \ldots x_n)}$$

PROBABILITY

PRODUCT RULE

The **joint probability density function** can be expressed as the **product** of the **conditional probabilities:**

$$
\begin{aligned}
f(x_1, x_2, \ldots, x_n) &= f(x_n|x_1, x_2 \ldots, x_{n-1}) f(x_1, x_2 \ldots, x_{n-1}) \\
&= f(x_n|x_1, x_2 \ldots, x_{n-1}) f(x_{n-1}|x_1, x_2 \ldots, x_{n-2}) f(x_1, x_2 \ldots, x_{n-2}) \\
&= \ldots = f(x_1) \prod_{i=2}^{n} f(x_i|x_1, \ldots, x_{i-1}).
\end{aligned}
$$

PROBABILITY

INDEPENDENCE

The **property** of **independence** is **generalized** for $\boldsymbol{n}$ random variables $X_1, X_2, \ldots, X_n$:

$$f(x_1, \ldots, x_n) = f(x_1)f(x_2)\cdots f(x_n).$$

It is said that $\boldsymbol{k}$ **events** $A_1, A_2, \ldots, A_k$ are **mutually independent** if for **any subset** $S \subseteq \{1, 2, \ldots, k\}$ we have:

$$P(\cap_{i \in S} A_i) = \prod_{i \in S} P(A_i).$$

PROBABILITY

RANDOM VECTORS

PROBABILITY

RANDOM VECTORS

When working with $\boldsymbol{n}$ **random variables**, it is convenient to **represent** them using a **vector**, denominated **random vector**, which performs a **mapping** of $\boldsymbol{\Omega} \rightarrow \mathbb{R}^{\boldsymbol{n}}$**:**

$$X = [X_1 \; X_2 \; \ldots \; X_n]^T$$

PROBABILITY

EXPECTED VALUE

The **calculation** of the **expected value** for $\boldsymbol{n}$ **continuous random variables** is presented where there is a **weighting function** $\boldsymbol{g(x_1, x_2, \ldots, x_n)}$ and a **probability density function $\boldsymbol{f_{X_1X_2,\ldots,X_n}(x_1, x_2, \ldots, x_n)}$.**

$$E[g(X)] = \int_{\mathbb{R}^n} g(x_1, x_2, \ldots, x_n) f_{X_1, X_2, \ldots, X_n}(x_1, x_2, \ldots x_n) dx_1 dx_2 \ldots dx_n,$$

PROBABILITY

BAYES THEOREM

PROBABILITY

DERIVATION OF THE THEOREM

Assuming that we have **2 discrete random variables** $\boldsymbol{X}$ and $\boldsymbol{Y}$ we can write the conditional probability as:

$$p(Y/X) = \frac{p(X,Y)}{p(X)}$$

Applying the **symmetric property** $\boldsymbol{p}\ (\boldsymbol{Y},\ \boldsymbol{X}) = \boldsymbol{p}\ (\boldsymbol{X},\ \boldsymbol{Y})$ and the **rule of the probability product** $\boldsymbol{p}(\boldsymbol{X},\boldsymbol{Y}) = \boldsymbol{p}(\boldsymbol{X}/\boldsymbol{Y})\boldsymbol{p}(\boldsymbol{Y})$ we have:

$$p(Y/X) = \frac{p(X/Y)p(Y)}{p(X)}$$

PROBABILITY

DERIVATION OF THE THEOREM

In addition, it is known by the **rule of the sum of probabilities**, that the **probability** of an **event** is equal to the sum of the intersections of that **event** with all other **events:**

$$p(X) = \sum_{Y} p(X,Y) = \sum_{Y} p(X/Y)\,p(Y)$$

$$p(Y/X) = \frac{p(X/Y)p(Y)}{\sum_{Y} p(X/Y)\,p(Y)}$$

PROBABILITY

APPLICATION OF THE THEOREM

Example:

2. *Question:* A diagnostic test has a probability 0.95 of giving a positive result when applied to a person suffering from a certain disease, and a probability 0.10 of giving a (false) positive when applied to a non-sufferer. It is estimated that 0.5 % of the population are sufferers. Suppose that the test is now administered to a person about whom we have no relevant information relating to the disease (apart from the fact that he/she comes from this population). Calculate the following probabilities:

(a) that the test result will be positive;

(b) that, given a positive result, the person is a sufferer;

PROBABILITY

DERIVATION OF THE THEOREM

Answer:

(a) $$\mathbf{P}(T) = \mathbf{P}(T|S)\mathbf{P}(S) + \mathbf{P}(T|S')\mathbf{P}(S') = (0.95 \times 0.005) + (0.1 \times 0.995) = 0.10425.$$

(b) $$\mathbf{P}(S|T) = \frac{\mathbf{P}(T|S)\mathbf{P}(S)}{\mathbf{P}(T|S)\mathbf{P}(S) + \mathbf{P}(T|S')\mathbf{P}(S')} = \frac{0.95 \times 0.005}{(0.95 \times 0.005) + (0.1 \times 0.995)} = 0.0455.$$

PROBABILITY

EXPLANATION OF THE THEOREM

Interpreting the **Bayes Theorem**:

$$\overbrace{P(A|B)}^{\text{Posterior}} = \frac{\overbrace{P(B|A)}^{\text{Likelihood}} * \overbrace{P(A)}^{\text{Prior}}}{\underbrace{P(B)}_{\text{Evidence}}}$$

1. **Prior**: **“beliefs”** that we have about how the **random variable** $\boldsymbol{A}$ is **distributed before** obtaining some **evidence** $\boldsymbol{B}$.

2. **Posterior:** captures the **distribution** of the **random variable A after** having collected the **evidence** $\boldsymbol{B}$.

3. **Likelihood**: expresses the **probability** that our **beliefs** (distribution of $\boldsymbol{A}$) are true according to evidence **B**.

PROBABILITY

EXPLANATION OF THE THEOREM

$$\text{posterior} \propto \text{likelihood} \times \text{prior}$$

Read pages **21 – 23** from the book "***Pattern Recognition and Machine Learning***" by Christopher Bishop, 1st edition.