Probability Theory and Bayes’ Rule

Elements of Probability

Sample space Ω: The set of all the outcomes of a random experiment. Here, each outcome $\omega \in \Omega$ can be thought of as a complete description of the state of the real world at the end of the experiment.

Set of events (or **event space**) F: A set whose elements $A \in F$ (called **events**) are subsets of Ω (i.e., $A \subseteq \Omega$ is a collection of possible outcomes of an experiment).

Probability measure: A function $P : F \rightarrow R$ that satisfies the following properties:

1. $P(A) \geq 0$, for all $A \in F$
2. $P(\Omega) = 1$
3. If $A_1, A_2, \ldots$ are disjoint events (i.e., $A_i \cap A_j = \emptyset$), $P(\cup A_i) = \sum_i P(A_i)$

Example: tossing a 6-sided dice

Sample space: Ω = *{*1*,* 2*,* 3*,* 4*,* 5*,* 6*}*.

One possible event space*: F* = *{∅,* Ω}

Probability measure*: P* (∅) = 0*,P* (Ω) = 1.

Another possible event space*:* all subsets of F

$$F = \{\phi, 1, 2, 3, 4, 5, 6, (1,2), (1,3), \ldots, \Omega\}$$

Probability measure*: P* (A)=i/6 where i is the number of elements in A.

$$P(\{1,2,3\}) = {}^{3}\!/_{6}$$

$$P(\{1,2,3,4\}) = {}^{4}\!/_{6}$$

Some properties

- If $A \subseteq B \Rightarrow P(A) \leq P(B)$.

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

- $P(A \cup B) \leq P(A) + P(B)$.

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

- If $A_1, \ldots, A_k$ are a set of disjoint events such that $\cup_{i=1} A_i = \Omega$, then

$$\sum_{i=1}^{k} P(A_i) = 1$$

Problem

Suppose that a die is biased (or loaded) so that 3 appears twice as often as each other number but that the other five outcomes are equally likely. What is the probability that an odd number appears when we roll this die?

P(odd | fair dice) = 3/6= ½

P(odd | loaded dice) = ?

LOADED DICE:

P(1) = a
P(2) = a
P(3) = 2a
P(4) = a
P(5) = a
P(6) = a

a +a + 2a + a + a + a = 1 ➔ 7a =1 ➔ a=1/7

P(odd|loaded dice) = P(1) + P(3) + P(5) = 1/7 + 2/7 + 1/7 = 4/7

Conditional probability and independence

Let B be an event with non-zero probability. The conditional probability of any event A given B is defined as:

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

Two events A and B are called independent if:

$$P\left(A|B\right)=P(A)$$

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

Bayes’ Theorem

Let A and B be two events. We have that:

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

$$P(A|B)P(B) = P(B|A)P(A)$$

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

Law of Total Probability

Let B_n be a partition of the sample space. Then:

$$\mathrm{P}(A) = \sum_n P(A|B_n)P(B_n)$$

Example:

$\mathrm{P}(\text{obtaining an A in the course }) = 0.6$
$\mathrm{P}(\text{not obtaining an A in the course}) = 0.4$

$\mathrm{P}(\text{working for OpenAI} \mid \text{obtained an A}) = 0.2$
$\mathrm{P}(\text{working for OpenAI} \mid \text{did not obtain an A}) = 0.1$

What is the probability of working for OpenAI?

$\mathrm{P}(\text{working for OpenAI}) = \mathrm{P}(\text{OpenAI} \mid \text{A})\ \mathrm{P}(\text{A}) + \mathrm{P}(\text{OpenAI} \mid \text{no A})\ \mathrm{P}(\text{no A}) = 0.2 \cdot 0.6 + 0.4 \cdot 0.1 = 0.16$

Problem

Suppose that a bag filled with die has an equal number of loaded and unloaded die. We pull out one dice, roll it, and observe a 3. What is the probability that we pulled out a loaded die?

Random variables

A random variable X is a mapping from the sample space to the real number line:

$$X(\omega):\Omega \to R$$

Example: Consider an experiment in which we flip 10 coins, and we want to know the number of coins that come up heads.

Here, the elements of the sample space Ω are 10-length sequences of heads and tails.

$$\omega_o = \langle H,T,H,H,T,T,T,H,T,H \rangle$$

X is a discrete random variable:

$$P(X = k) := P(\{\omega : X(\omega)= k\})$$

Continuous random variables

Suppose that X(ω) is a random variable indicating the amount of time it takes a radioactive particle to decay.

In this case X takes on an infinite number of possible values, so it is called a **continuous random variable**.

We denote the probability that X takes on a value between two constants a and b (where a<b) as:

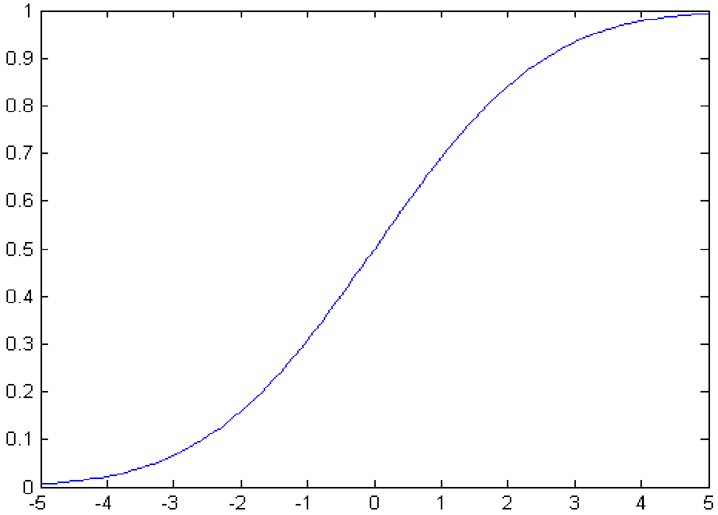
$$P(a \leq X \leq b) := P(\{\omega : a \leq X(\omega) \leq b\}).$$

Cumulative distribution function (CDF)

A **cumulative distribution function (CDF)** is a function $F_X : \mathbb{R} \rightarrow [0, 1]$ which specifies a probability measure as:

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

The CDF answers the question: what is the probability that the random variable X has a value less than x?



Properties of CDF

- $0 \leq F_X(x) \leq 1$
- $\lim_{x\to-\infty} F_X(x) = 0$
- $\lim_{x\to\infty} F_X(x) = 1$
- $x \leq y \Longrightarrow F_X(x) \leq F_X(y)$.

Probability mass function (PMF)

For discrete random variables, we can directly specify the probability of each possible value.

$$p_X(x) : \Omega \to R$$

$$p_X(x) = P(X = x)$$

Properties

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

$$\sum_{x \in val(x)} p_X(x) = 1$$

Val(x): all possible values X can take

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

Probability density function (PDF)

For continuous random variables, the derivative of the CDF produces what is known as the probability density function (PDF):

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

Important: $f_X(x)$ is NOT equal to P(X=x)

Rather

$$f_X(x)\Delta x = P(x \le X \le x + \Delta x)$$

Properties of PDF

$$f_X(x) \geq 0$$

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

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

Expectation

Assume that X is a discrete random variable with PMF $p\ (x)$.

Let *g(X)* be a function of *X* such that *g(X)*: $\mathbb{R} \rightarrow \mathbb{R}$

g(X) is then itself a random variable with expectation given by

$$E\{g(X)\} = \sum_{x} g(x)p\ (x)$$

If X is a continuous random variable with PDF $f(x)$, then

$$E\{g(X)\} = \int_{-\infty}^{\infty} g(x)f(x)dx$$

Properties of expectation

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

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

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

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

Variance

The variance of a random variable X is a measure of how much it deviates around its mean.

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

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

Properties:

$Var(a) = 0$ where a is a constant

$$Var[af(X)] = a^2 Var[f(X)]$$

Example

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

Common Discrete Random Variables

X ~ *Bernoulli*(*p*) (where 0 ≤ *p* ≤ 1): one if a coin with heads probability *p* comes up heads, zero otherwise.

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

X ~ *Binomial*(*n, p*) (where 0 ≤ *p* ≤ 1): the number of heads in *n* independent flips of a coin with heads probability *p*.

$$p(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}$$

X~Geometric(p), (where 0 ≤ *p* ≤ 1), the number of flips of a coin with probability p until the first heads

$$p(X = k) = p^1 (1 - p)^{n-1}$$

$X \sim Poisson(\lambda)$ (where $\lambda > 0$): a probability distribution over the nonnegative integers used for modeling the frequency of rare events.

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

Common Continuous Random Variables

$X \sim$ *Uniform*(a, b) (where $a < b$): equal probability density to every value between a and b on the real line.

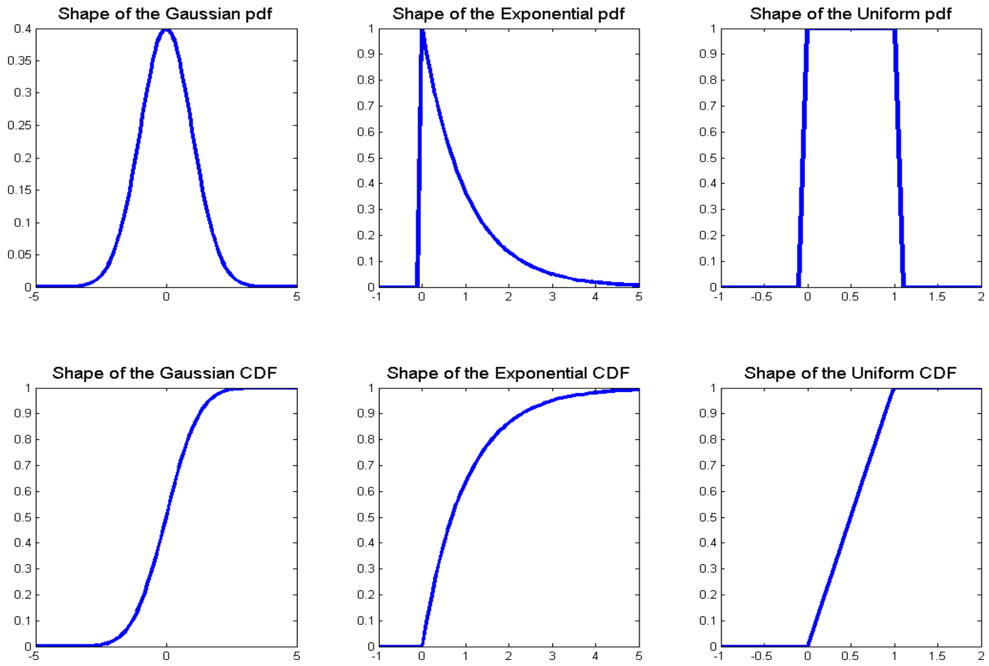
$$f_X(x) = \begin{cases} \frac{1}{b-a} & a \le x \le b \\ 0 & \text{otherwise} \end{cases}$$

$X \sim$ *Exponential*(λ) (where $\lambda > 0$): decaying probability density over the nonnegative reals.

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

$X \sim$ *Normal*(μ, σ^2): also known as the Gaussian distribution

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



Shape of the Gaussian pdf

Shape of the Exponential pdf

Shape of the Uniform pdf

Shape of the Gaussian CDF

Shape of the Exponential CDF

Shape of the Uniform CDF

Distribution	PDF or PMF	Mean	Variance
$Bernoulli(p)$	$p,$if $x = 1$ $1\ \ p,$ if $x = 0.$	p	$p(1\ p)$
$Binomial(n, p)$	k	np	npq
$Geometric(p)$	$p(1\ \ p)^{k1}$ for $k = 1, 2, \ldots$	$\frac{1}{p}$	$\frac{1\ p}{p2}$
$Poisson(\lambda)$	$e^{\lambda}\lambda^{x}/x!$ for $k = 1, 2, . . .$	λ	λ
$Uniform(a, b)$	$\frac{1}{b\ a}$ $\forall x \in (a, b)$	$\frac{a+b}{2}$	$\frac{(b\ a)^2}{12}$
$Gaussian(\mu,\sigma^2)$	$\frac{1}{o\ \sqrt{\ }\ 2\pi}\ e^{\frac{(x-\mu)^2}{2\sigma2}}$	μ	$\sigma2$
$Exponential(\lambda)$	$\lambda e^{\lambda x}\ \ x \geq 0, \lambda > 0$	$\frac{1}{\lambda}$	$\frac{1}{\lambda 2}$

Two Random Variables

The joint CDF between two variables X and Y describes how X and Y behave as a pair

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

From this joint CDF we can calculate the *marginal* CDFs of X and Y

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

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

Properties of joint CDF

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

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

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

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

Joint and marginal PMFs

If X and Y are discrete random variables, then their joint PMF is given by:

$$p_{XY}(x,y) = P\left(X = x, Y = y\right)$$

Integrating across the possible values of Y produces the marginal PMF of random variable X:

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

Joint and marginal PDFs

Let X and Y be two continuous random variables with joint CDF F_{XY}. Then their joint PDF is found by differentiating F_{XY}:

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

The marginal PDF of X is again found by integrating across the possible values of random variable Y:

$$f_X(x) = \int_y f_{XY}(x,y)dy$$

Conditional distributions

Conditional distributions answer questions such as: what is the probability of an event *given that* another event has occurred. For **discrete** random variables, we define the PMF of X conditioned on Y according to:

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

For **continuous** random variables, we define the PDF of X conditioned on Y according to:

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

Bayes’ Rule

An important formula for relating the conditional distribution of X|Y to that of Y|X is given by Bayes’ Rule. For discrete random variables X and Y, we have

$$p_{Y|X}(y\,|\,x) = \frac{p_{XY}\left(x,y\right)}{p_X(x)} = \frac{p_{X|Y}\left(x\,|\,y\right)p_Y(y)}{\sum\limits_{y'} p_{X|Y}\left(x\,|\,y'\right)p_Y(y')}$$

If X and Y are continuous random variables,

$$f_{Y|X}(y\,|\,x) = \frac{f_{XY}\left(x,y\right)}{f_X(x)} = \frac{f_{X|Y}\left(x\,|\,y\right)f_Y(y)}{\int\limits_{-\infty}^{\infty} f_{X|Y}\left(x\,|\,y'\right)f_Y(y')\,dy'}$$

Independence of Random Variables

Two random variables *X* and *Y* are **independent** if $F_{XY}(x, y) = F_X(x)F_Y(y)$ for all values of *x* and *y*.

Discrete random variables:

$$p_{XY}\left(x, y\right) = p_X(x)p_Y(y)$$

$$p_{X|Y}(x \mid y) = p_X(x)$$

Continuous random variables:

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

$$f_{X|Y}(x \mid y) = f_X(x)$$

Expectation of a function of two random variables

Suppose that X and Y are two discrete random variables. Let g(X,Y) be a function of X and Y. G is then itself a random variable with expectation given by:

$$E\left[g(X,Y)\right]=\sum_{x}\sum_{y}g(x,y)p_{XY}(x,y)$$

If X and Y are now continuous, then the expectation of g(X,Y) is given by:

$$E\left[g(X,Y)\right]=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}g(x,y)f_{XY}(x,y)dxdy$$

Covariance

The covariance of random variables X and Y is given by:

$$cov(X,Y) = E\{(X - E\{X\})(Y - E\{Y\})\}$$

By expanding this expression, one can show that :

$$cov(X,Y) = E\{XY\}\text{-}\ E\{Y\}\,E\{Y\}$$

Properties of Expectation and Covariance

$E[f(X, Y) + g(X, Y)] = E[f(X, Y)] + E[g(X, Y)].$

$Var[X + Y] = Var[X] + Var[Y] + 2Cov[X, Y].$

If X and Y are independent, then $Cov[X, Y] = 0.$

If X and Y are independent, then $E[f(X)g(Y)] = E[f(X)]E[g(Y)].$

Uncorrelated random variables

If cov(X,Y)=0, then we say that X and Y are uncorrelated

X and Y can be uncorrelated but not independent

Multiple random variables

Let X_1, X_2, …, X_N denote *N* random variables whose collective behavior we are interested in.

Joint CDF

$$F_{X_1X_2\ldots X_N}(x_1, x_2, \ldots, x_N) = P\left(X_1 \le x_1, X_2 \le x_2, \ldots, X_n \le x_n\right)$$

Joint PDF

$$f_{X_1X_2\ldots X_N}(x_1, x_2, \ldots, x_N) = \frac{\partial^N F_{X_1X_2\ldots X_N}(x_1, x_2, \ldots, x_N)}{\partial x_1 \partial x_2 \ldots \partial x_n}$$

Marginal PDF of X_1

$$f_{X_1}(x_1) = \int \ldots \int f_{X_1X_2\ldots X_N}(x_1, x_2, \ldots, x_N) dx_2 \ldots dx_N$$

Conditional PDF of X_1

$$f_{X_1|X_2\ldots X_N}(x_1 \mid x_2, \ldots, x_N) = \frac{f_{X_1X_2\ldots X_N}(x_1, x_2, \ldots, x_N)}{f_{X_2\ldots X_N}(x_2, \ldots, x_N)}$$

Chain Rule

$$f\left(x_1, x_2, \ldots, x_n\right) = f\left(x_n \mid x_1, x_2, \ldots, x_{n-1}\right) f\left(x_1, x_2, \ldots, x_{n-1}\right)$$

$$= f\left(x_n \mid x_1, x_2, \ldots, x_{n-1}\right) f\left(x_{n-1} \mid x_1, .x_2, .., x_{n-2}\right) f\left(x_1, x_2, \ldots, x_{n-2}\right)$$

$$= f(x_1) \prod_{i=2}^{n} f\left(x_i \mid x_1, x_2, \ldots, x_{i-1}\right)$$

Independence of multiple random variables

We say that random variables X_1, X_2, … X_n are mutually independent if their joint PDF factorizes into a product:

$$f\left(x_1, x_2, \ldots, x_n\right) = f\left(x_1\right) f\left(x_2\right) \ldots f\left(x_n\right)$$

Random vectors

When working with multiple random variables, it is convenient to group them into a random vector according to:

$$\boldsymbol{X} = \begin{bmatrix} X_1 & X_2 & \dots & X_n \end{bmatrix}^T$$

A function of this random vector g has expectation given by:

$$E\left[g(\boldsymbol{X})\right] = \int \dots \int g(x_1, x_2, \dots, x_n) f_{X_1 X_2 \dots X_N}\left(x_1, x_2, \dots, x_n\right) dx_1 \dots dx_n$$

Covariance matrix

For a random vector X, the covariance matrix Σ conveys how pairs of elements behave (i.e., correlate):

$$\Sigma = \begin{bmatrix} \operatorname{cov}(X_1, X_1) & \cdots & \operatorname{cov}(X_1, X_n) \\ \operatorname{cov}(X_n, X_1) & \cdots & \operatorname{cov}(X_n, X_n) \end{bmatrix}$$

$$= \begin{bmatrix} E\left[X_1^2\right] - E^2\left[X_1\right] & \cdots & E\left[X_1 X_n\right] - E\left[X_1\right]E\left[X_n\right] \\ E\left[X_n X_1\right] - E\left[X_n\right]E\left[X_1\right] & \cdots & E\left[X_n^2\right] - E^2\left[X_n\right] \end{bmatrix}$$

The covariance matrix is both **positive semidefinite** and **symmetric**.

Multivariate Gaussian Distribution

An important example of a probability distribution over random vector is given by the multivariate normal distribution.

Random vector **X** is said to be jointly Gaussian with a mean vector **μ** and a covariance matrix **Σ** if its probability distribution follows:

$$f_{X_1 X_2 \ldots X_n}(x_1, x_2, \ldots, x_n) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp \begin{matrix}\#\\ \%\\ \therefore\end{matrix} -\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu) \begin{matrix}\&\\ (\\ \end{matrix}$$

This is more compactly written as: **X**~*N*(**μ**,Σ)

Gaussian random vectors are often used in machine learning to model the noise affecting the data.