Chapter 6. Probability and Distribution

FPT Education

FPT UNIVERSITY

Chapter 6. Probability and Distribution

FPT Education
FPT UNIVERSITY

FPT Education
FPT UNIVERSITY

Definition

- The sample space is the set of all possible outcomes of the experiment, sample space usually denoted by Ω.
- The event space is the space of potential results of the experiment. A subset A of the sample space Ω is in the event space $\mathcal{A}$ if at the end of the experiment we can observe whether a particular outcome $\omega \in \Omega$ is in A.
- Associate each event $A \in \mathcal{A}$ a number $P(A)$ that measures the probability or degree of belief that the event will occur. $P(A)$ is called the probability of A. We have $0 \leq P(A) \leq 1, \ \forall \ A \in \mathcal{A}, \ P(\Omega) = 1$.

Note. In machine learning, we refer to probabilities on quantities of interest as the target space $\mathcal{T}$ and refer to elements of $\mathcal{T}$ as states. Define a target space function $X : \Omega \to \mathcal{T}$ that takes an element of Ω (an outcome) and returns a particular quantity of interest x, a value in $\mathcal{T}$. This mapping is called a random variable.

Example

There are a red ball and blue ball (of the same size). You are going to draw two balls with replacement successively. Assume that the composition of the bag of coins is such that a draw returns at random a red with probability 0.3.

The state space $\Omega = \{RR, RB, BR, BB\}$, ($R$: red, B : blue).
Let us define a random variable X that maps the sample space Ω to $\mathcal{T}$, which denotes the number of times we draw the red ball out of the bag, then target space $\mathcal{T} = \{0, 1, 2\}$. The random variable X:

$$X(RR) = 2, X(RB) = X(BR) = 1, X(BB) = 0.$$

The probability mass function of X given by

$$P(X = 2) = P(R).P(R) = 0.3 * 0.3 = 0.9$$
$$P(X = 1) = P((RB)) + P((BR)) = 0.3 * 0.7 + 0.7 * 0.3 = 0.42$$
$$P(X = 0) = P((BB)) = P(B)P(B) = 0.7 * 0.7 = 0.49$$

Note. Consider the random variable $X : \Omega \to \mathcal{T}$ and a subset $S \subseteq T$.

1. $$P_X(S) = P(X \in S) = P(X^{-1}(S)) = P(\{\omega \in \Omega : X(\omega) \in S\}),$$
   where the function P_X or equivalently $P \circ X^{-1}$ is the law or distribution of random variable X.
2. When $\mathcal{T}$ is finite or countably infinite, X is called a discrete random variable. For continuous random variables, we only consider $\mathcal{T} = \mathbb{R}$ or $\mathcal{T} = \mathbb{R}^D$.

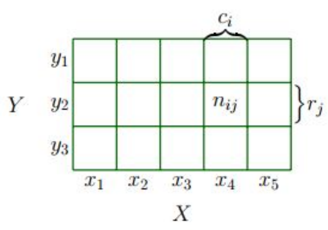
Definition

The expression $P(X = x)$ and $P(X \leq x)$ are called the probability mass function and the cumulative distribution function of X respectively.

Let X, Y be two discrete random variables.

Definition

- The joint probability is defined as the entry of both values jointly: $P(X = x_i, Y = y_j)$ $= P(X = x_i \cap Y = y_j) = \frac{n_{ij}}{N}$, where n_{ij} is the number of events with state x_i and y_j and N the total number of events.



- The probability that $X = x$ and $Y = y$ written as $p(x, y)$ is called the joint probability.
- The marginal probability that X takes the value x irrespective of the value of random variable Y is written as $p(x)$. Write $X \sim p(x)$ to denote that the random variable X is distributed according to $p(x)$.
- The conditional probability $p(y|x)$ is the probability for which $Y = y$ occurring in the presence of $X = x$.

Example

There are 3 red balls, 5 blue balls and 2 pink ball. Ann is going to draw two balls without replacement. Let X denote the number of red balls chosen and let Y denote the number of blue balls chosen. Find $p(x, y)$.

Answer: Both X, Y can take on values $0, 1, 2$ and $X + Y \leq 2$. We have

$$P(X = 0, Y = 0) = \frac{1}{C(10,2)} = \frac{1}{45}, P(X = 0, Y = 1) = \frac{5 * 2}{C(10,2)} = \frac{10}{45}$$
$$P(X = 0, Y = 2) = \frac{C(5,2)}{C(10,2)} = \frac{10}{45}, P(X = 1, Y = 0) = \frac{3 * 2}{C(10,2)} = \frac{6}{45}$$
$$P(X = 1, Y = 1) = \frac{3 * 5}{C(10,2)} = \frac{15}{45}, P(X = 2, Y = 0) = \frac{C(3,2)}{C(10,2)} = \frac{3}{45}.$$

Definition

A function $f : \mathbb{R} \to R$ is called a probability density function (pdf) if

1. $f(x) \geq 0,\ \forall\ x \in \mathbb{R}^D$.
2. $\int_{\mathbb{R}^D} f(x)dx = 1$.

Definition

A cumulative distribution function (cdf) of a multivariate real-valued random variable X with states $x \in \mathbb{R}^D$ is given by

$$\begin{aligned} F_X(x) &= P(X_1 \leq x_1, \ldots, X_D \leq x_D) \\ &= \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_D} f(z_1, \ldots, z_D)dz_1 \cdots dz_D, \end{aligned}$$

Let X, Y be two random variables. Let $p(x,y), p(x), p(y)$ be the joint distribution, the marginal distribution of X, the marginal distribution of Y respectively and let $p(y|x)$ is the conditional distribution of Y given $X = x$.

- **Sum rule**

$$p(x) = \begin{cases} \sum_{y \in \mathcal{Y}} p(x,y) & \text{if } Y \text{ is discrete} \\ \int_{\mathcal{Y}} p(x,y)dy & \text{if } Y \text{ is continuous} \end{cases},$$

  where $\mathcal{Y}$ are the states of the target space of random variable Y.

- **Product rule**

$$p(x,y) = p(y|x)p(x).$$

- **Bayes' theorem**

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

FPT Education
FPT UNIVERSITY

Definition

The expected value of a function $g : \mathbb{R} \to \mathbb{R}$ of a univariate variable $X \sim p(x)$ is given by

$$E_X[g(x)] = \begin{cases} \sum_{x \in \mathcal{X}} g(x)p(x) & \text{if } X \text{ is discrete} \\ \int_{\mathcal{X}} g(x)p(x)dx & \text{if } X \text{ is continuous} \end{cases}$$

where $\mathcal{X}$ is the target space of X.

For multivariate random variable $X = \left[X_1, \ldots, X_D\right]^T$, we define the expected value

$$E_X[g(x)] = \begin{bmatrix} E_{X_1}[g(x)] \\ \vdots \\ E_{X_D}[g(x)] \end{bmatrix} \in \mathbb{R}^D.$$

Definition

The mean of a random variable X with states $x \in \mathbb{R}^D$ is defined as

$$E_X[x] = \begin{bmatrix} E_{X_1}[x_1] \\ \vdots \\ E_{X_D}[x_D] \end{bmatrix} \in \mathbb{R}^D$$

where

$$E_{X_d}[x_d] = \begin{cases} \sum_{x_i \in \mathcal{X}} x_i p(X_d = x_i) & \text{if } X \text{ is discrete} \\ \int_{\mathcal{X}} x_d p(x_d) dx_d & \text{if } X \text{ is continuous.} \end{cases}$$

Note. The expected value is a linear operator. For example, given a real-valued function $f(x) = ag(x) + bh(x)$ where $a, b \in \mathbb{R}$ and $x \in \mathbb{R}^D$, we obtain

$$E_X[f(x)] = aE_X[g(x)] + bE_X[h(x)].$$

Definition

The covariance between two univariate random variables $X, Y \in \mathbb{R}$ is given by the expected product of their deviations from their respective means

$$\begin{aligned} \mathrm{Cov}_{X,Y}[x,y] &:= E_{X,Y}[(x - E_X(x))(y - E_Y(y))] \\ &= E[xy] - E[x]E[y]. \end{aligned}$$

Note. When the random variable associated with the expectation or covariance is clear by its arguments, the subscript is often suppressed (for example, $E_X[x]$ is often written as $E[x]$).

Definition

- $Cov[x,x]$ is called the variance and denoted by $V_X[x]$.
- The square root of the variance is called the standard deviation and denoted by $\sigma(x)$.

Definition

For two multivariate random variables X and Y with states $x \in \mathbb{R}^D$ and $y \in \mathbb{R}^E$ respectively, the covariance between X and Y is defined as

$$\mathrm{Cov}[x, y] = E[xy^T] - E[x]E[y]^T = \mathrm{Cov}[y, x]^T \in \mathbb{R}^{D \times E}.$$

The variance of a random variable X with states $x \in \mathbb{R}^D$ and a mean vector $\mu \in \mathbb{R}^D$ is defined as

$$\begin{aligned} V_X[x] &= \mathrm{Cov}_X[x, x] = E_X[(x - \mu)(x - \mu)^T] \\ &= E[xx^T] - E[x]E[x]^T \\ &= \begin{bmatrix} \mathrm{Cov}[x_1, x_1] & \mathrm{Cov}[x_1, x_2] & \cdots & \mathrm{Cov}[x_1, x_D] \\ \mathrm{Cov}[x_2, x_1] & \mathrm{Cov}[x_2, x_2] & \cdots & \mathrm{Cov}[x_2, x_D] \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}[x_D, x_1] & \mathrm{Cov}[x_D, x_2] & \cdots & \mathrm{Cov}[x_D, x_D] \end{bmatrix} \end{aligned}$$

the covariance matrix of X.

Example

Given the probability mass function with discrete random variables X, Y

$\frac{X}{Y}$	0	1	2
0	$\frac{1}{45}$	$\frac{6}{45}$	$\frac{3}{45}$
1	$\frac{10}{45}$	$\frac{15}{45}$	0
2	$\frac{10}{45}$	0	0

Find $E[x], E[y], \mathrm{Cov}[x, y]$.

$$E[x] = 0.\frac{21}{45} + 1 * \frac{21}{45} + 2 * \frac{3}{45} = \frac{27}{45} = \frac{3}{5}$$

$$E[y] = 0.\frac{10}{45} + 1.\frac{25}{45} + 2.\frac{10}{45} = 1$$

$$\mathrm{Cov}[x, y] = E[xy] - E[x]E[y] = 1.1.\frac{15}{45} - \frac{3}{5}.1 = -\frac{4}{15}.$$

Definition

The correlation between two random variables X, Y is given by

$$\text{corr}[x,y] = \frac{\text{Cov}[x,y]}{\sqrt{V[x]V[y]}} \in [-1,1].$$

Example

Consider above example, we have

$$V[x] = E[x^2] - (E[x])^2 = 0^2 * \frac{21}{45} + 1^2 * \frac{21}{45} + 2^2 * \frac{3}{45} - \left(\frac{3}{5}\right)^2 = \frac{28}{75}$$

$$V[y] = E[y^2] - (E[y])^2 = 0^2 * \frac{10}{45} + 1^2 * \frac{25}{45} + 2^2 * \frac{10}{45} - 1^2 = \frac{4}{9}$$

Hence

$$\text{corr}[x,y] = \frac{\text{Cov}[x,y]}{\sqrt{V[x]V[y]}} = \frac{-4/15}{\sqrt{\frac{28}{75} * \frac{4}{9}}} \approx -0.654$$

6.4.2 Empirical Means and Covariances

Given a particular dataset $x_1, \dots, x_N$ taken from a population, we can obtain an estimate of the mean, which is called the empirical mean or sample mean.

Definition

- The empirical mean (sample mean) vector is defined as

$$\bar{x} = \frac{1}{N}\sum_{n=1}^{N} x_n \in \mathbb{R}^D.$$

- The empirical covariance matrix is a $D \times D$ matrix

$$\Sigma = \frac{1}{N}\sum_{n=1}^{N} (x - \bar{x})(x - \bar{x})^T.$$

FPT Education FPT UNIVERSITY

Theorem

Consider two random variables X, Y with states $x, y \in \mathbb{R}^D$. Then

$$E[x+y] = E[x] + E[y]$$
$$E[x-y] = E[x] - E[y]$$
$$V[x+y] = V[x] + V[y] + Cov[x,y] + Cov[y,x]$$
$$V[x-y] = V[x] + V[y] - Cov[x,y] - Cov[y,x].$$

Theorem

Consider a random variable X with mean μ and covariance matrix Σ and a (deterministic) affine transformation $y = Ax + b$ of x. Then y is itself a random variable whose mean vector and covariance matrix are given by

$$E[y] = E_X[Ax+b] = A\mu + b$$
$$V_Y[y] = V_X[Ax+b] = A\Sigma A^T.$$

FPT Education
FPT UNIVERSITY

Definition

Two random variables X, Y are statistically independent if and only if one of the followings holds

$$
\begin{aligned}
&(1) \quad p(x,y) = p(x)p(y)\\
&(2) \quad p(y|x) = p(y)\\
&(3) \quad p(x|y) = p(x)\\
&(4) \quad V[x+y] = V[x] + V[y].
\end{aligned}
$$

Note. If X, Y are independent then $\mathrm{Cov}[x,y] = 0$, but the inverse statement is not true.

Definition

Two random variables X and Y are conditionally independent given Z, denoted by $X \perp\!\!\!\perp Y|Z$ if and only if $\forall\ z \in \mathcal{Z}$ one of the following holds:

(1) $p(x, y|z) = p(x|z)p(y|z)$

(2) $p(x|y, z) = p(x|z)$

(3) $p(y|x, z) = p(y|z)$.

6.4.6 Inner Products of Random Variables

Let X, Y be two random variables. Since random variables can be considered vectors in a vector space. We define the inner product:

$$\langle X, Y\rangle = X \cdot Y := \mathrm{Cov}[x, y].$$

Hence

$$\langle X, X\rangle = \|X\|^2 = V(X).$$

The angle θ between X, Y:

$$\cos\theta = \frac{\langle X, Y\rangle}{\|X\|\,\|Y\|} = \frac{\mathrm{Cov}[x, y]}{\sqrt{V[x]V[y]}}.$$

This means that X and Y are orthogonal if and only if $\mathrm{Cov}[x, y] = 0$, i.e., they are uncorrelated.

FPT Education
FPT UNIVERSITY

Definition

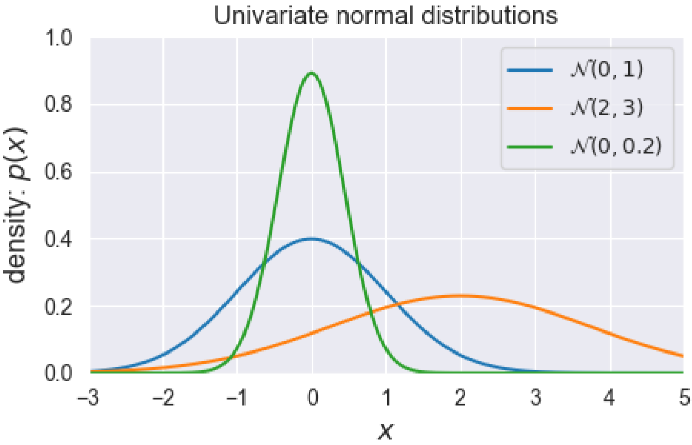
For a univariate random variable, the Gaussian distribution (or normal distribution) has a density that is given by

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

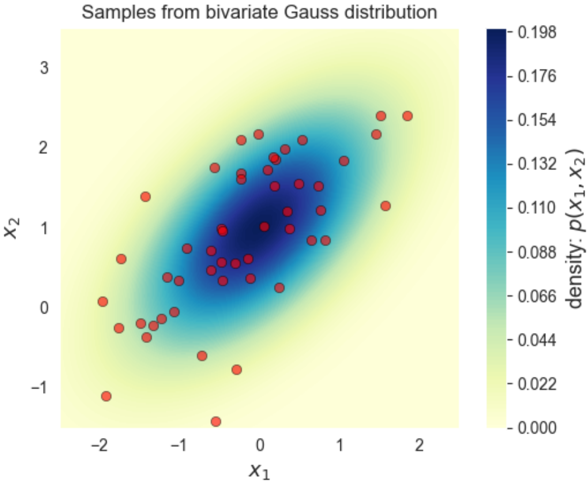
The multivariate Gaussian distribution is fully characterized by a mean vector μ and a covariance matrix Σ and defined as

$$p(x|\mu,\Sigma) = \frac{1}{\sqrt{(2\pi)^D \det\Sigma}} \exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right),$$

where $x\mathbb{R}^D$. We write $p(x) = \mathcal{N}(x|\mu,\Sigma)$ or $X \approx \mathcal{N}(\mu,\Sigma)$.
The special case of the Gaussian with $\mu = 0$ and $\sigma = I_D$, is referred to as the standard normal distribution.



Univariate normal distributions
density: $p(x)$
x
$\mathcal{N}(0, 1)$
$\mathcal{N}(2, 3)$
$\mathcal{N}(0, 0.2)$
1.0
0.8
0.6
0.4
0.2
0.0
−3
−2
−1
0
1
2
3
4
5


Samples from bivariate Gauss distribution
x_2
3
2
1
0
−1
−2
−1
0
1
2
x_1
0.198
0.176
0.154
0.132
0.110
0.088
0.066
0.044
0.022
0.000
density: $p(x_1, x_2)$

Let X and Y be two multivariate Gauss variables, that may have different dimensions. Write the Gaussian distribution in terms of the concatenated states $[x^T, y^T]$:

$$p(x,y) = \mathcal{N}\left(\begin{bmatrix}\mu_x \\ \mu_y\end{bmatrix}, \begin{bmatrix}\Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_{yy}\end{bmatrix}\right)$$

where $\Sigma_{xx} = \mathrm{Cov}[x,x]$ and $\Sigma_{yy} = \mathrm{Cov}[y,y]$ are the marginal covariance matrices of x and y, respectively, and $\Sigma_{xy} = \mathrm{Cov}[x,y]$ is the cross covariance matrix between x and y.

Theorem

The marginal distribution $p(x)$ and the conditional distribution $p(x|y)$ are also Gaussian:

$$p(x) = \int p(x,y)dy = \mathcal{N}(x|\mu_x, \Sigma_{xx})$$

$$p(x|y) = \mathcal{N}\left(\mu_{x|y}, \Sigma_{x|y}\right),$$

$$\mu_{x|y} = \mu_x + \Sigma_{xy}\Sigma_{yy}^{-1}(y - \mu_y),\ \Sigma_{x|y} = \Sigma_{xx} - \Sigma_{xy}\Sigma_{yy}^{-1}\Sigma_{yx}.$$

Example

Consider the bivariate Gaussian distribution

$$p(x_1, x_2) = \mathcal{N}\left(\begin{bmatrix} -1 \\ 2 \end{bmatrix}, \begin{bmatrix} 0.2 & -0.5 \\ -0.5 & 5 \end{bmatrix}\right)$$

Find $p(x_1)$ and $p(x_1|x_2 = 1)$.
Answer:

$$p(x_1) = \mathcal{N}(-1, 0.2)$$

$$\mu_{x_1|x_2=1} = -1 + (-0.5) * 5^{-1} * (1 - 2) = -0.9$$

$$\sigma^2_{x_1|x_2=-1} = 0.2 - (-0.5) * 5^{-1} * (-0.5) = 0.15$$

$$\Rightarrow p(x_1|x_2 = 1) = \mathcal{N}(-0.9, 0.15).$$

FPT Education
FPT UNIVERSITY

Theorem

- *If X, Y are independent Gaussian random variables of the same dimension with*

$$p(x) = \mathcal{N}(\mu_x, \Sigma_x)$$
$$p(y) = \mathcal{N}(\mu_y, \Sigma_y)$$

then for $a, b \in \mathbb{R}$, $aX + bY$ is also Gaussian distributed

$$p(ax + by) = \mathcal{N}(a\mu_x + b\mu_y, a^2\Sigma_x + b^2\Sigma_y).$$

- *Suppose that $X \sim \mathcal{N}(\mu, \Sigma)$ of dimension D. For a matrix $A \in \mathbb{R}^{N \times D}$, let Y be a random variable such that $y = Ax$. Then*

$$Y \sim \mathcal{N}(A\mu, A\Sigma A^T).$$

Summary

We have studied:

- the notation of probability space;
- discrete and continuous probabilities;
- some important rules and theorems of probabilities;
- independence notation;
- Gaussian distribution.

Exercises for practice: 6.1-6.10 (pages 222, 223).