Appendix B Elements of Probability Theory

B.1 Probability Theory

We begin by presenting a brief review of the probability theory. The very reason of doing so is that contemporary machine learning methodologies are developed in a probabilistic framework. In other words, most part of machine leaning makes sense only when you look at it from a statistical perspective.

B.1.1 Definition of Probability

A. Classical Probability

There are different kinds of probability: probability as intuition, probability as the ratio of favorable to total outcomes (also known as classical theory of probability), probability as a measure of frequency of occurrence, and probability based on axiomatic theory.

Definition B.1 *Random experiment* and *basic event*

An experiment *E* is said to be a *random experiment* if it can be performed repeatedly and its outcomes are *not* deterministic but probabilistic. An outcome of a random experiment, denoted by ω, is called a *basic event.* ■

Definition B.2 *Sample space*

The set of all basic events, denoted by Ω, is called *sample space.* ■

Example B.1

Let *E* be the experiment of picking a ball at random from a box containing *N* identical balls, where the balls have been marked as #1, #2, …, to #*N*. Obviously *E* is random experience because its outcome has *N* possibilities. If the number on the ball that has been picked is *i*, then the outcome is a basic event and is denoted by ω_i. Obviously, the sample space is $\Omega = \{\omega_1, \cdots, \omega_N\}$. ■

Example B.2

Let *E* be that of observing the average daytime temperature in a city in October. Obviously *E* is a random experiment. If we denote the outcome of experiment *E* by ω_a, then the sample space in this case may be described as $\Omega = \{\omega_a, -\infty < \omega_a < \infty\}$. Note that Ω contains infinite number of basic events. ■

Definition B.3 *General event*

A general event (or simply called it event), say *A*, consists of several basic events, hence a general event is a subset of sample space Ω. ■

Example B.3

Continue from Example B.1. Suppose one randomly picks a ball from the box, one can state that

• Event *A* of getting a ball whose number is no greater than 3 is a general event because it consists of three basic events: $A = \{\omega_1, \omega_2, \omega_3\}$ which is a proper subset of $\Omega = \{\omega_1, \cdots, \omega_N\}$.

• Event B of getting a ball with number even numbers is a general event because $B = \{\omega_2, \omega_4, \ldots, \omega_{100}\} \subset \Omega$. ■

The classical probability theory applies to the cases where the random experiment *E* obeys two basic assumptions: (i) the total number of outcomes, *N*, is finite; and (ii) all outcomes (i.e. the results of basic events) are equally likely. Consequently, the probability of a (general) *event A*, denoted by *P*(*A*), is obtained *a priori* (i.e., relating to reasoning that proceeds from theoretical deduction rather than from observation or experience) by counting the number of ways N_e that event *A* can occur, then computing the ratio N_e/N as the probability. This is,

$$P(A) = \frac{N_e}{N} \tag{B.1}$$

Example B.4

Continue from Example B.3.

Let *A* be the event of obtaining a ball whose number is no greater than 3. The probability *P*(*A*) = 3/100 = 0.03.

Let *B* be the event of obtaining a ball that is even-numbered. The probability *P*(*B*) = 50/100 = 0.5.

Let *C* be the event of obtaining a ball whose number is a multiple of 3. The probability *P*(*C*) = 33/100 = 0.33.

■

The major problems with the classical theory of probability are that it cannot handle outcomes that are not equally likely; and it cannot deal with infinite number of outcomes.

B. Probability Based On Axiomatic Theory

Developed by A. N. Kolmogorov in 1930's, the probability based on axiomatic theory is the one that is followed by most modern texts on the subject. Kolmogorov's probability theory is set-theoretic in that a random event is considered as a subset of a sample space in an abstract setting where the two basic assumptions made in the classical probability theory are no longer necessary.

In order to present Kolmogorov's axiomatic definition of probability, basic elements of *set algebra* and the sigma algebra ($\sigma-$ algebra) are sketched first.

Set Algebra

• A *set* is a collection of objects (or elements).

• A set *B* is said to be a *subset* of set *A*, and this relation is denoted by $B \subseteq A$, if *B* is contained within set *A*. Let for example set *A* denote all Victoria residents and set *B* be all Victoria residents whose height is between 5.5 and 6.5 feet. Obviously set *B* is a (proper) subset of set *A*.

• Let *A* and *B* be two sets in space Ω. The *union* (sum) of *A* and *B*, denoted by $A \bigcup B$ or *A* + *B*, is the set of elements that are in at least one of the sets *A* or *B*.

• The *intersection* (or set product) of *A* and *B*, denoted by $A \bigcap B$ or *AB*, is the set of elements that are in both *A* and *B*.

• The *empty* set, denoted by ϕ, is a set that contains no objects.

• Let *A* be a set in space Ω, the complement of *A*, denoted by A^c, is a set of all elements that are not in *A*. Obviously, $A \bigcup A^c = \Omega$ and $A \bigcap A^c = \phi$.

- Two sets *A* and *B* are said to be equal if both $B \subseteq A$ and $A \subseteq B$ hold.
- The difference of *A* and *B*, denoted by $A - B$, is the set of elements that are in *A* but not in *B*. It follows that

$$A - B = AB^c \quad \text{and} \quad B - A = BA^c$$

- The *exclusive-or* of sets *A* and *B*, denoted by $A \oplus B$, is the set of elements that are in *A* or *B*, but not in both. Obviously, we can write

$$A \oplus B = (A - B) \bigcup (B - A)$$

- Sets *A* and *B* are said to be disjoint if they have no elements in common, thus $AB = \phi$.

- It can readily be verified that

$$(A \bigcup B)^c = A^c \bigcap B^c \quad \text{and} \quad (A \bigcap B)^c = A^c \bigcup B^c$$

which can be extended by induction to the general case of *n* sets as

$$\left(\bigcup_{i=1}^{n} A_i \right)^c = \bigcap_{i=1}^{n} A_i^c \quad \text{and} \quad \left(\bigcap_{i=1}^{n} A_i \right)^c = \bigcup_{i=1}^{n} A_i^c$$

Definition B.4 σ*-Algebra*

Let Ω be a set and we consider a collection of subsets of Ω, denoted by $\mathcal{F}$. $\mathcal{F}$ is said to be a σ*-Algebra* if

1. $\phi \in \mathcal{F}$ and $\Omega \in \mathcal{F}$,

2. If $A \in \mathcal{F}$ and $B \in \mathcal{F}$, then $A \bigcup B \in \mathcal{F}$ and $A \bigcap B \in \mathcal{F}$.

3. If $A \in \mathcal{F}$ then $A^c \in \mathcal{F}$.

4. $\mathcal{F}$ is closed under countable set of unions, intersections, and combinations. Hence if A_1, …, A_n, … belong to $\mathcal{F}$, then

$$\bigcup_{i=1}^{\infty} A_i \in \mathcal{F} \text{ and } \bigcap_{i=1}^{\infty} A_i \in \mathcal{F} \qquad \blacksquare$$

Why is σ ***-Algebra relevant?***

Recall in the classical theory of probability the number of basic events must be finite and all outcomes are equally likely, i.e. the probability of every basic event is the same. The axiomatic probability theory built on σ-algebra allows us to remove these two fundamental limitations. Consider a random experiment and the associated sample space Ω. In the axiomatic probability theory we only consider the collection $\mathcal{F}$ of those subsets of Ω that form a σ-algebra. These subsets are called events. It turns out that $\mathcal{F}$ usually includes all subsets of engineering and science interest.

Definition B.5 *Axiomatic probability*

Given a sample space Ω and a σ-algebra $\mathcal{F}$ of Ω, probability $P[\cdot]$ is a set function that assigns to every event $A \in \mathcal{F}$ that obeys the three axioms:

(1) $P[A] \geq 0$.

(2) $P[\Omega] = 1$.

(3) $P[A \bigcup B] = p[A] + P[B]$ if $A \bigcap B = \phi$.

From these axioms, one can establish the following important properties:

(4) $P[\phi] = 0$.

(5) $P[A \bigcap B^c] = P[A] - P[A \bigcap B]$.

(6) $P[A] = 1 - P[A^c]$.

(7) $P[A \bigcup B] = P[A] + P[B] - P[A \bigcap B]$.

(8) $P\left[\bigcup_{i=1}^{n} A_i\right] = \sum_{i=1}^{n} P[A_i]$ if $A_i \cap A_j = \phi$ for all $i \neq j$. ■

Definition B.6 *Probability space*

The triple of a sample space Ω, a collection of (general) events $\mathcal{F}$ that form a σ-algebra, and a probability measure *P*, namely $(\Omega, \mathcal{F}, P)$, is called a probability space. ■

B.1.2 Joint Probability, Conditional Probability, and Independence

A. Joint Probability

The *joint probability* of two events *A* and *B* is defined as the probability of "events *A* and *B* both occur", namely $P[A \bigcap B]$ or P[*AB*].

B. Conditional Probability

Often times people want to know the probability of "event *A* occurs" given that event *B* has occurred. This is called *conditional probability* and is denoted by *P*[*A*|*B*]. It is intuitively clear that because of the presence of the condition that "event *B* has occurred", *P*[*A*|*B*] is in general different from *P*[*A*].

Example B.5

Suppose there are four identical balls in a box, which are marked as #1, #2, #3, and #4. The random experiment is to pick up a ball from the box. Now consider two events: Event *A* is that of obtaining ball #4; and event *B* is that of obtaining an even-numbered ball. Obviously, we have *P*[*A*] = ¼ and *P*[*A*|*B*] = ½. ■

A question that naturally arises is how to define and compute conditional probability in general circumstances? Let *A* and *B* are two events in a random experiment. Each outcome of the random experiment must fall into one of the four cases: (1) *A* occurs and *B* does not; (2) *B* occurs and *A* does not; (3) *A* and *B* both occur; and (4) *A* and *B* both do not occur.

Suppose one applies above analysis to Example B.5, repeats the experiment *n* times, and denotes the times that each of the four cases occurs by n_1, n_2, n_3, and n_4, respectively. We can state that

- $n_1 + n_2 + n_3 + n_4 = n$
- the frequency of event $B = F_n(B) = \frac{n_2 + n_3}{n}$.
- the frequency of event $AB = F_n(AB) = \frac{n_3}{n}$.
- given that event *B* has occurred, the frequency of event $A = F_n(A|B) = \frac{n_3}{n_2 + n_3}$.

It follows that

$$F_n(A \mid B) = \frac{F_n(AB)}{F_n(B)} \quad \text{provided that} \quad F_n(B) > 0 \tag{B.2}$$

Based on (B.2), conditional probability is defined as follows.

Definition B.7 *Conditional probability*

Let $(\Omega, \mathcal{F}, P)$ be a probability space, $A \in \mathcal{F}, B \in \mathcal{F}$ with $P[B] > 0$. The conditional probability of event *A* given that event *B* has occurred is defined by

$$P[A \mid B] = \frac{P[AB]}{P[B]} \tag{B.3}$$

Similarly, the conditional probability of event *B* given that event *A* has occurred is defined by

$$P[B \mid A] = \frac{P[AB]}{P[A]} \tag{B.4}$$

provided that $P[A] > 0$. From (B.3) and (B.4) it follows that

$$P[AB] = P[A]P[B \mid A] = P[B]P[A \mid B] \qquad \blacksquare \tag{B.5}$$

Example B.6

Consider a binary communication system with a two-symbol alphabet, i.e., 0 and 1. Let *X* and *Y* be the transmitted and received symbols, respectively. Here the sample space is given by $\Omega = \{(X,Y) : X = 0 \text{ or } 1, Y = 0 \text{ or } 1\} = \{(0,0), (0,1), (1,0), (1,1)\}$. Suppose that the communication system is slightly corrupted by noise such that

$$P[Y=1\,|\,X=1]=0.92,\quad P[Y=0\,|\,X=1]=0.08$$
$$P[Y=0\,|\,X=0]=0.92,\quad P[Y=1\,|\,X=0]=0.08$$

And by design, $P[X = 0] = P[X = 1] = 0.5$. Under these circumstances, we have

$$P[X=0,Y=0]=P[X=0]\cdot P[Y=0\,|\,X=0]=0.5\times 0.92=0.46$$
$$P[X=0,Y=1]=P[X=0]\cdot P[Y=1\,|\,X=0]=0.5\times 0.08=0.04$$
$$P[X=1,Y=0]=P[X=1]\cdot P[Y=0\,|\,X=1]=0.5\times 0.08=0.04$$
$$P[X=1,Y=1]=P[X=1]\cdot P[Y=1\,|\,X=1]=0.5\times 0.92=0.46$$

■

Properties of Conditional Probability

Property 1 (Multiplication formula)

Let A_1, A_2, …, A_n be n events with $n \ge 2$ and $P[A_1A_2\cdots A_{n-1}]>0$, then

$$P[A_1A_2\cdots A_n]=P[A_1]\cdot P[A_2\,|\,A_1]\cdot P[A_3\,|\,A_1A_2]\cdots P[A_n\,|\,A_1A_2\cdots A_{n-1}] \tag{B.6}$$

Proof:

Using (B.3), we see that the right-hand side of (B.6) is equal to

$$P[A_1]\cdot\frac{P[A_1A_2]}{P[A_1]}\cdot\frac{P[A_1A_2A_3]}{P[A_1A_2]}\cdots\frac{P[A_1A_2\cdots A_n]}{P[A_1A_2\cdots A_{n-1}]}=P[A_1A_2\cdots A_n] \qquad ■$$

Property 2 (Unconditional probability)

Let A_1, A_2, …, A_n be mutually exclusive events such that $\bigcup_{i=1}^n A_i=\Omega$ with $P[A_i]>0$ for all i. Let B be any event defined over the probability space of A_i's. Then

$$P[B]=\sum_{i=1}^{n}P[A_i]P[B\,|\,A_i] \tag{B.7}$$

Proof:

From

$$B=B\Omega=B\bigcup_{i=1}^n A_i=\bigcup_{i=1}^n BA_i$$

where $\{BA_i\}$ are mutually exclusive, it follows that

$$P[B]=P\left[\bigcup_{i=1}^n BA_i\right]=\sum_{i=1}^{n}P[BA_i]=\sum_{i=1}^{n}P[A_i]\cdot P[B\,|\,A_i] \qquad ■$$

Property 3 (Bayes formula)

Let A_1, A_2, …, A_n be mutually exclusive events such that $\bigcup_{i=1}^n A_i=\Omega$ with $P[A_i]>0$ for all i.

Let B be any event with $P[B]>0$. Then

$$P[A_j \mid B] = \frac{P[B \mid A_j] \cdot P[A_j]}{\sum_{i=1}^{n} P[B \mid A_i] \cdot P[A_i]} \tag{B.8}$$

Proof:

By following the definition of conditional probability and the formula of unconditional probability in (B.7), we obtain

$$P[A_j \mid B] = \frac{P[BA_j]}{P[B]} = \frac{P[B \mid A_j] \cdot P[A_j]}{\sum_{i=1}^{n} P[B \mid A_i] \cdot P[A_i]} \qquad \blacksquare$$

The Bayes formula finds many applications in science and engineering. The terms in (B.8) bear various names: $P[A_j|B]$ is often called *a posteriori* probability of A_j given B; $P[B|A_j]$ is called the *a priori* probability of B given A_j; and $P[A_j]$ is called the *causal* or *a priori* probability of A_j. Typically *a priori* probabilities are estimated from past measurements or presupposed by experience, while *a posteriori* probabilities are measured or computed from observations.

Example B.7

Suppose there are three boxes that look identical, each contains certain number of red and blue balls that are identical except the color. The number of red and blue balls in box i are r_i and b_i, respectively, for $i = 1, 2, 3$. The experiment in question is to randomly pick a box, then randomly pick a ball. The outcome was a red ball. Given that it is a red ball, compute the probability of the ball belongs to box 1.

Solution

Let A_i be the event of the ball in question belongs to box i for $i = 1, 2, 3$. Obviously, $\{A_i$, for $i = 1, 2, 3\}$ are mutually exclusive, $\bigcup_{i=1}^{3} A_i = \Omega$, and $P[A_1] = P[A_2] = P[A_3] = 1/3$. We also define event B as "it is a red ball". With these definitions, the problem we need to address is to compute the conditional probability $P[A_1|B]$.

Clearly, formula (B.8) is applicable if we are able to compute conditional probabilities $P[B|A_i]$ for $i = 1, 2, 3$. These conditional probabilities are found to be

$$P[B \mid A_i] = \frac{r_i}{r_i + b_i} \quad \text{for} \quad i = 1, 2, 3.$$

Hence (B.8) yields

$$P[A_1 \mid B] = \frac{\frac{1}{3}\frac{r_1}{r_1 + b_1}}{\frac{1}{3}\frac{r_1}{r_1 + b_1} + \frac{1}{3}\frac{r_2}{r_2 + b_2} + \frac{1}{3}\frac{r_3}{r_3 + b_3}} = \frac{1}{1 + \frac{r_2(r_1 + b_1)}{r_1(r_2 + b_2)} + \frac{r_3(r_1 + b_1)}{r_1(r_3 + b_3)}} \qquad \blacksquare$$

C. Independence

Consider two events A and B. We have seen that in general probability $P[A]$ differs from the conditional probability $P[A|B]$ as long as the occurrence of event B has an impact on occurrence of event A. In other words, if $P[A] = P[A|B]$ happens, then the occurrence of B has no impact on the occurrence of A, and we will say events A and B are mutually independent. Note that $P[A] = P[A|B]$ in conjunction with (B.3) leads to

$$P[AB] = P[A] \cdot P[B] \tag{B.9}$$

This motivates the following definition.

Definition B.8 *Independence*

- Events A and B are said to be independent from each other if (B.9) holds.
- n events $A_1, A_2, \ldots, A_n$ are said to be mutually independent if

$$P[A_1 A_2 \cdots A_n] = P[A_1] \cdot P[A_2] \cdots P[A_n] \qquad \blacksquare \tag{B.10}$$

B.1.3 Random Variables, Distribution Function, and Probability Density

A. Random Variables

Let E be a random experiment associated with sample space Ω. Corresponding to each outcome ω, suppose there is a real-valued function $\xi(\omega)$. In probability theory, one is concerned not only with the value of $\xi(\omega)$, but also with the probability of $\xi(\omega)$ taking certain values.

Example B.8

In a junior-high school a student (that is an ω) is randomly selected, whose height is recorded as $\xi(\omega)$. One of the valid questions in this case would be "what is the probability of the height no greater than x cm?", namely, what is $P\left[\xi(\omega) \le x\right]$? ■

Evidently, to address the question in the associated probability space $(\Omega, \mathcal{F}, P)$, it is necessary to assure that $\xi(\omega) \le x$ belongs to $\sigma -$ algebra $\mathcal{F}$ so that $P\left[\xi(\omega) \le x\right]$ is well defined. In probability theory functions of this kind are called *random variables*.

Definition B.9 *Random variables*

Let $\xi(\omega)$ be a real-valued function defined over sample space Ω that is associated with a probability space $(\Omega, \mathcal{F}, P)$. If, for any real x, $(\xi(\omega) \le x)$ is an event, i.e., $(\xi(\omega) \le x) \in \mathcal{F}$, then $\xi(\omega)$ is called a random variable. ■

The introduction of random variables was a major event in the development of modern probability theory as the concept made it possible to extend the probability theory to include studies of random variables and a variety of related issues. One of the issues is *distribution function* of a random variable.

B. <u>Distribution Function and Probability Density</u>

Definition B.10 *Distribution function of a random variable*

Let $\xi(\omega)$ be a random variable, the distribution function of ξ, denoted by $\Phi_\xi(x)$, is defined as the probability of event $(\xi(\omega) \le x)$. Namely,

$$\Phi_\xi(x) = P[\xi(\omega) \le x] \qquad \blacksquare \qquad \text{(B.11)}$$

Several basic properties of distribution functions follow:

- $\Phi_\xi(x)$ is monotonically non-decreasing, i.e.,

$$x_2 \ge x_1 \text{ implies that } \Phi_\xi(x_2) \ge \Phi_\xi(x_1) \qquad \text{(B.12)}$$

- $$\Phi_\xi(-\infty) = 0,\ \Phi_\xi(\infty) = 1 \qquad \text{(B.13)}$$

- $$P[x_1 < \xi(\omega) \le x_2] = \Phi_\xi(x_2) - \Phi_\xi(x_1) \qquad \text{(B.14)}$$

- $$P[\xi(\omega) > x] = 1 - \Phi_\xi(x) \qquad \text{(B.15)}$$

- $$P[\xi(\omega) = x] = \Phi_\xi(x) - \Phi_\xi(x-0) \qquad \text{(B.16)}$$

where $\Phi_\xi(x-0) = \lim_{a \uparrow x} \Phi_\xi(a)$.

Types of distribution functions

There are several types of distribution functions, with the most important being *discrete type* and *continuous type*.

Discrete type distribution

A discrete random variable ξ takes on discrete values x_0, x_1, x_2, … (finite or countable infinite), with probabilities p_0, p_1, p_2, …, respectively. Thus a discrete distribution is characterized by a two-row matrix, known as *density matrix*, with finite or countable infinite columns as

$$\xi \Leftrightarrow \begin{bmatrix} x_0 & x_1 & x_2 & \cdots \\ p_0 & p_1 & p_2 & \cdots \end{bmatrix} \qquad \text{(B.17)}$$

Obviously we have $p_i \ge 0$ and $\sum_i p_i = 1$. The distribution function in this case is given by

$$F(x) = \sum_{i:\, x_i \le x} p_i \qquad \text{(B.18)}$$

which is a staircase function that is continuous from right, see Fig. B.1.

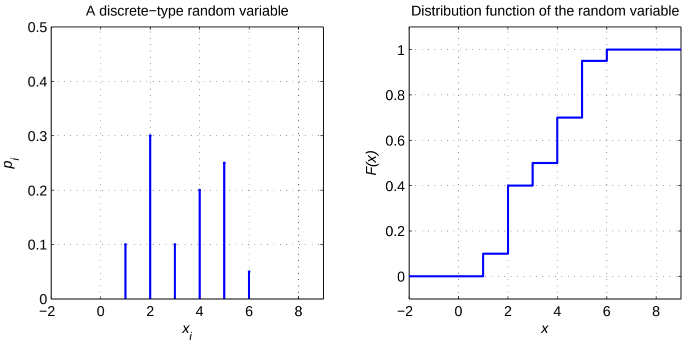


Fig. B.1 Left: A discrete-type random variable with $x_0 = 1$, $x_1 = 2$, …, $x_5 = 6$ and $p_0 = 0.1$, $p_1 = 0.3$, $p_2 = 0.1$, $p_3 = 0.2$, $p_4 = 0.25$, and $p_5 = 0.05$; Right: The distribution function of the random variable.

Example B.9 *Bernoulli distribution*

Also known as *binomial distribution*, the Bernoulli distribution is discrete and assumes the form of

$$\begin{bmatrix} 0 & 1 & \cdots & n \\ p_0 & p_1 & \cdots & p_n \end{bmatrix}$$

where

$$p_k = \binom{n}{k} p^k q^{n-k} \quad p \geq 0,\ q \geq 0,\ p + q = 1 \tag{B.19}$$

with

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

being the number of combinations of taking k elements out of n elements. Note that p_k in (B.19) are exactly the coefficient of polynomial $(p+q)^n$, hence the name of binomial distribution. The reason it also bears the name of Bernoulli distribution has to do with the well-known *Bernoulli trial* which is a random experiment with exactly two possible outcomes: "success" and "failure". Denote the probability of success and failure by p and q, respectively, and suppose the experiment is repeated n time independently. It is obvious that there are 2^n possible outcomes; the number of outcomes with k successes and $(n - k)$ failures is equal to $\binom{n}{k}$; and the probability of "one such event occurs" is $p^k q^{n-k}$. Therefore, the probability of having k successes and $(n - k)$ failures is precisely given by p_k as seen in (B.19). ■

Continuous type distribution

A distribution of random variable ξ is said to be of continuous type if its distribution function assumes the form

$$\Phi_{\xi}(x) = \int_{-\infty}^{x} \varphi_{\xi}(y)\, dy \qquad \text{(B.20)}$$

where $\varphi_{\xi}(x) \ge 0$ is called *density function* or *density*. Since $\Phi_{\xi}(\infty) = 1$, we have

$$\int_{-\infty}^{\infty} \varphi_{\xi}(y)\, dy = 1$$

Example B.10 *Uniform distribution*

Let *a* < *b*. Consider the random variable ξ that takes a value over the interval [*a*, *b*] uniformly randomly. The associated distribution, called uniform distribution, is of continuous type with the density function given by

$$\varphi_{\xi}(x) = \begin{cases} \dfrac{1}{b-a} & \text{if } x \in [a,b] \\ 0 & \text{if } x \notin [a,b] \end{cases} \qquad \text{(B.21)}$$

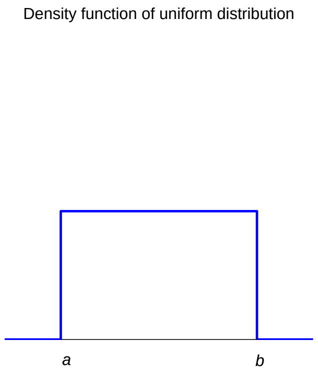
Given the density in (B.21), the distribution function is found to be

$$\Phi_{\xi}(x) = \int_{-\infty}^{x} \varphi_{\xi}(y)\, dy = \begin{cases} 0 & \text{if } x < a \\ \dfrac{x-a}{b-a} & \text{if } a \le x \le b \\ 1 & \text{if } x > b \end{cases} \qquad \text{(B.22)}$$

See Figs. B.2 and B.3 for plots of the density $\varphi_{\xi}(x)$ and distribution function $\Phi_{\xi}(x)$. ■

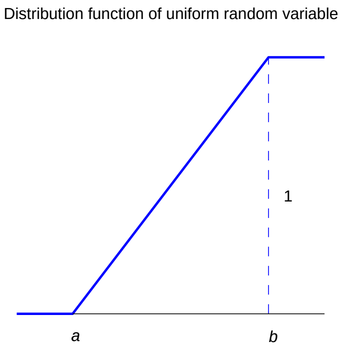
C. *Expectation, Variance and Covariance*

An important operation involving random variables is that of finding weighted average of a random variable ξ or a (measurable) function of ξ.



Density function of uniform distribution

Figure B.2



Distribution function of uniform random variable

Figure B.3

Definition B.11 *Expectation*

The average of a random variable ξ under its probability distribution is called the expectation of

ξ, denoted by $E[\xi]$.

For a discrete type random variable ξ that takes values x_0, x_1, x_2, … with probability p_0, p_1, p_2, … (recall the density matrix given by (B.17), namely

$$\begin{bmatrix} x_0 & x_1 & x_2 & \cdots \\ p_0 & p_1 & p_2 & \cdots \end{bmatrix}$$

), the expectation of ξ, also known as mean value of ξ, is defined as

$$E[\xi] = \sum_i x_i p_i$$

Since the random variable ξ here takes values x_i, for the sake of notation convenience we denote the random variable by x, and write the above expression as

$$E[x] = \sum_i x_i p_i \tag{B.23}$$

For a continuous type random variable ξ with density $\varphi_\xi(x)$, its expectation is defined as

$$E\left[x\right] = \int_{-\infty}^{\infty} x\varphi_\xi(x)dx \tag{B.24}$$

The concept of expectation can be extended to expectation of a (measurable) function, say $g(\xi)$, of random variable ξ as

$$E\left[g(x)\right] = \sum_i g(x_i)p_i \tag{B.25}$$

for discrete-type case, or

$$E\left[g(x)\right] = \int_{-\infty}^{\infty} g(x)\varphi_\xi(x)dx \tag{B.26}$$

for continuous-type case. ■

Definition B.12 *Variance*

The variance of a function g of random variable ξ provides a measure of how much there is in g around its mean value $E\left[g(x)\right]$:

$$\text{var}\left[g(x)\right] = E\left[\left(g(x) - E\left[g(x)\right]\right)^2\right] = E\left[g(x)^2\right] - \left(E\left[g(x)\right]\right)^2 \tag{B.27}$$

In particular, with $g(x) = x$ (B.27) yields the variance of a random variable ξ itself as

$$\text{var}\left[x\right] = E\left[x^2\right] - \left(E\left[x\right]\right)^2 \qquad ■ \tag{B.28}$$

Definition B.13 *Covariance*

Covariance is concerned with two random variables *x* and *y* and quantifies the extent to which *x* and *y* vary together, namely,

$$\operatorname{cov}\left[x, y\right] = E_{x,y}\left[\left(x - E[x]\right)\left(y - E[y]\right)\right] = E_{x,y}[xy] - E[x]E[y] \qquad \text{(B.29)}$$

where $E_{x,y}$ denotes expectation with respect to joint density $\varphi(x, y)$, namely,

$$E_{x,y}[xy] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xy\,\varphi(x, y)\,dx\,dy \qquad \blacksquare$$

D. Independence of Random Variables and Conditional Distribution

Definition B.14 *Independence of random variables*

Random variables $\xi_1(\omega),\ldots,\xi_n(\omega)$ with distribution functions $\Phi_1(x_1),\ldots,$ $\Phi_n(x_n)$ are said to be *mutually independent* if

$$\Phi(x_1,\ldots,x_n) = \Phi_1(x_1)\cdots\Phi_n(x_n) \qquad \text{(B.30)}$$

holds, where $\Phi(x_1,\ldots,x_n)$ is the joint distribution of $\xi_1(\omega),\cdots,\xi_n(\omega)$, namely,

$$\Phi(x_1,\ldots,x_n) = P[\xi_1 \le x_1,\cdots,\xi_n \le x_n] \qquad \blacksquare$$

From (B.30), it follows that

$$P[\xi_1 \le x_1,\cdots,\xi_n \le x_n] = P[\xi_1 \le x_1]\cdots P[\xi_n \le x_n] = \prod_{i=1}^{n} P[\xi_i \le x_i] \qquad \text{(B.31)}$$

In words, if random variables $\xi_1,\ldots,\xi_n$ are mutually independent, the joint probability of the event $(\xi_1 \le x_1,\cdots,\xi_n \le x_n)$ is equal to product of the probability of individual event $(\xi_i \le x_i)$.

For continuous type of random variables, (B.30) implies that $\xi_1(\omega),\cdots,\xi_n(\omega)$ are mutually independent if the joint probability density equals the product of the individual probability density almost everywhere (a.e.), i.e.,

$$\varphi(x_1,\ldots,x_n) = \varphi_1(x_1)\cdots\varphi_n(x_n) \qquad \text{(a.e.)} \qquad \text{(B.32)}$$

Definition B.15 *Conditional distribution*

For two discrete-type random variables ξ and η with

$$\xi \Leftrightarrow \begin{bmatrix} x_0 & x_1 & x_2 & \cdots \\ p_0 & p_1 & p_2 & \cdots \end{bmatrix} \text{ and } \eta \Leftrightarrow \begin{bmatrix} y_0 & y_1 & y_2 & \cdots \\ q_0 & q_1 & q_2 & \cdots \end{bmatrix}$$

where $p_i > 0$ and $q_i > 0$ are assumed, denote the joint probability

$$P[\xi = x_i, \eta = y_j] = p_{i,j}$$

hence we can write

$$p_i = \sum_j p_{i,j} \text{ and } q_j = \sum_i p_{i,j}$$

which in turn implies that

$$P[\xi = x_i \mid \eta = y_j] = \frac{P[\xi = x_i, \eta = y_j]}{P[\eta = y_j]} = \frac{p_{i,j}}{q_j} = \frac{p_{i,j}}{\sum_i p_{i,j}} \tag{B.33}$$

and

$$P[\eta = y_j \mid \xi = x_i] = \frac{p_{i,j}}{p_i} = \frac{p_{i,j}}{\sum_j p_{i,j}} \tag{B.34}$$

The conditional distribution refers to the probability of event $(\xi \le x \text{ given } \eta = y_j)$ or that of $(\eta \le y \text{ given } \xi = x_i)$. By using (B.33) and (B.34), we obtain

$$P[(\xi \le x \mid \eta = y_j)] = \frac{\sum_{i: x_i \le x} p_{i,j}}{\sum_i p_{i,j}} \tag{B.35}$$

and

$$P[\eta \le y \mid \xi = x_i] = \frac{\sum_{j: y_j \le y} p_{i,j}}{\sum_j p_{i,j}} \tag{B.36}$$

From (B.35) and (B.36), we see that conditional distributions involving two discrete-type random variables can be expressed (and evaluated) using their joint probability distribution $p_{i,j}$.

For two continuous-type random variables ξ and η with joint probability density $\varphi(x, y)$, the conditional distributions are given by

$$P[\xi \le x \mid \eta = y] = \frac{\int_{-\infty}^{x} \varphi(z, y)\, dz}{\int_{-\infty}^{\infty} \varphi(z, y)\, dz} \tag{B.37}$$

and

$$P[\eta \le y \mid \xi = x] = \frac{\int_{-\infty}^{y} \varphi(x, z)\, dz}{\int_{-\infty}^{\infty} \varphi(x, z)\, dz} \tag{B.38}$$

By writing (B.37) as

$$P[\xi \le x \mid \eta = y] = \int_{-\infty}^{x} \left(\frac{\varphi(z, y)}{\int_{-\infty}^{\infty} \varphi(z, y)\, dz} \right) dz$$

one may interpolate the integrand in the above expression as the *conditional distribution density* and denote it as $\varphi(x \mid y)$:

$$\varphi(x \mid y) = \frac{\varphi(x, y)}{\int_{-\infty}^{\infty} \varphi(z, y)\, dz} \tag{B.39}$$

In this way, (B.37) becomes

$$P[\xi \le x \mid \eta = y] = \int_{-\infty}^{x} \varphi(z \mid y)\, dz$$

Similarly, by defining the conditional distribution density

$$\varphi(y \mid x) = \frac{\varphi(x, y)}{\int_{-\infty}^{\infty} \varphi(x, z)\, dz} \tag{B.40}$$

(B.38) can be expressed as

$$P[\eta \le y \mid \xi = x] = \int_{-\infty}^{y} \varphi(z \mid x)\, dz \qquad \blacksquare$$

B.1.4 Gaussian Distribution

Also known as the *normal distribution*, the Gaussian distribution played an extremely important role in the development of probability theory, and is arguably the most useful among all probabilistic distributions. The life span of bulbs produced under practically same manufacturing conditions, for example, obeys a Gaussian distribution. This is also true for several other measures of products that are manufactured in quantity under the same conditions. Gaussian distribution is also encountered in many natural, biological, and social events/phenomena: velocities of gas molecules; errors in measuring a physical object; heights/weights of biological species, yearly precipitations of a certain city, etc. A common thread of these random variables is that they are cumulative synthesis of many small (i.e. of minor importance), independent random components.

A. One-Dimensional Gaussian Distribution

Definition B.16 *Gaussian distribution*

A one-dimensional Gaussian distribution is a continuous-type distribution whose density is given by

$$\varphi_{\mu,\sigma}(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2} \tag{B.41}$$

where μ and σ are two real-valued parameters. Hence the Gaussian distribution function is given by

$$\Phi_{\mu,\sigma}(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{-(y-\mu)^2/2\sigma^2} dy \tag{B.42}$$

A common notation for the one-dimensional normal distribution is $\mathcal{N}(x, \mu, \sigma^2)$. ■

By (B.24) and (B.28), the expectation and variance of a Gaussian random variable are found to be

$$E[x] = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} x e^{-(x-\mu)^2/2\sigma^2} dx = \mu \tag{B.43}$$

and

$$\text{var}[x] = E[x^2] - E[x]^2 = \sigma^2 \tag{B.44}$$

respectively. In other words, we see that the Gaussian distribution is characterized by its expectation and variance. Fig. B.4 illustrates the meaning of μ as the mean value of *x* using two Gaussian density functions with different μ's, while Fig. B.5 illustrates the meaning of σ^2 as the variance of *x* by several Gaussian density functions with different σ's.

Example B.11

Let *x* be a Gaussian random variable with $\mu = 2$ and $\sigma = 1$. Compute the probability of *x* being in between $\mu - 3$ and $\mu + 3$ (i.e. *x* falls into the interval [–1, 5]).

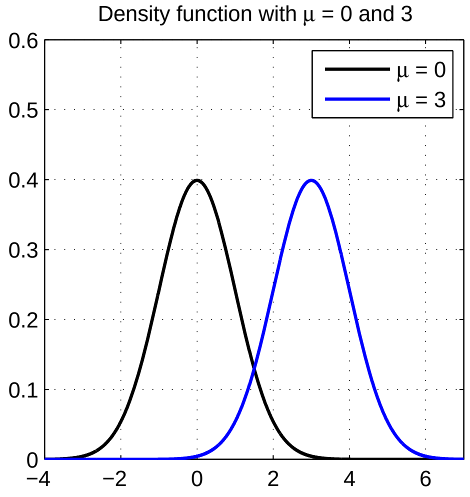


Figure B.4

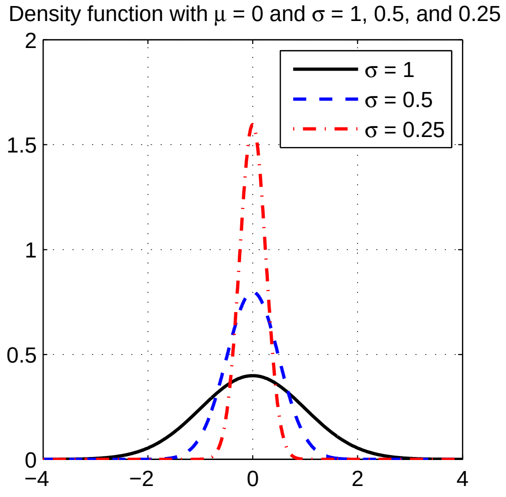


Figure B.5

Solution

From (B.14) and (B.42) it follows that

$$P[-1 \le \xi \le 5] = \frac{1}{\sqrt{2\pi}} \int_{-1}^{5} e^{-(y-2)^2/2} dy = \frac{1}{\sqrt{2\pi}} \int_{-3}^{3} e^{-z^2/2} dz \approx 99.7\% \quad ■$$

B. Multidimensional Gaussian Distribution

Let us consider *n* random variables $\xi_1(\omega),\cdots,\xi_n(\omega)$. The joint distribution of these random variables is defined by

$$\Phi(x_1,x_2,\cdots,x_n)=P[\xi_1(\omega)\le x_1,\xi_2(\omega)\le x_2,\cdots,\xi_n(\omega)\le x_n] \tag{B.45}$$

If $\xi_1(\omega),\cdots,\xi_n(\omega)$ are of continuous type, then $\Phi(x_1,x_2,\cdots,x_n)$ can be expressed as

$$\Phi(\boldsymbol{x})=\int_{-\infty}^{x_1}\cdots\int_{-\infty}^{x_n}\varphi(y_1,\cdots,y_n)\,dy_n\cdots dy_1 \tag{B.46}$$

where $\boldsymbol{x}=\begin{bmatrix}x_1 & x_2 & \cdots & x_n\end{bmatrix}^T$ and $\varphi(x_1,\cdots,x_n)$ is the density function.

As expected, the most important continuous-type multidimensional distribution is Gaussian distribution whose density in matrix notation is given by

$$\varphi(\boldsymbol{x},\boldsymbol{\mu},\Sigma)=\frac{1}{(2\pi)^{n/2}}\frac{1}{|\Sigma|^{1/2}}\exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right\} \tag{B.47}$$

where Σ is a symmetric and positive definite matrix of size *n* by *n*, called *covariance matrix*, and $|\Sigma|$ denotes the determinant of Σ. Oftentimes the density of normal distribution is denoted by $\mathcal{N}(\boldsymbol{x},\boldsymbol{\mu},\Sigma)$. It can be verified that the expectation of Gaussian $\boldsymbol{\xi}$ is $\boldsymbol{\mu}$, and the covariance of between the individual components (as random variables themselves) of $\boldsymbol{\xi}$ is given by Σ.

An important special case of multidimensional Gaussian distribution is when the individual random variables are mutually independent, the density function in this case becomes separable as

$$\varphi(x_1,\cdots,x_n)=\varphi_1(x_1)\cdots\varphi(x_n)$$

with each $\varphi_i(x_i)$ being a one-dimensional Gaussian density, i.e.,

$$\varphi_i(x_i)=\frac{1}{\sigma_i\sqrt{2\pi}}e^{-(x_i-\mu_i)^2/2\sigma_i^2}$$

Therefore the probability density of *n* mutually independent Gaussian distribution is given by

$$\varphi(x_1,\cdots,x_n)=\prod_{i=1}^{n}\frac{1}{\sigma_i\sqrt{2\pi}}e^{-(x_i-\mu_i)^2/2\sigma_i^2}=\frac{1}{(2\pi)^{n/2}}\frac{1}{\sigma_1\cdots\sigma_n}\exp\left(-\frac{1}{2}\sum_{i=1}^{n}\frac{(x_i-\mu_i)^2}{\sigma_i^2}\right) \tag{B.48}$$

It can be readily verified that (B.47) coincides with (B.48) by assigning $\boldsymbol{\mu}=\begin{bmatrix}\mu_1 & \cdots & \mu_n\end{bmatrix}$ and

$$\Sigma = \begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_n^2 \end{bmatrix}$$

In other words, (B.47) becomes joint density of *n* independent Gaussian distributions if and only if Σ is *diagonal*.

B.1.5 Likelihood Function and Log-Likelihood

Consider a continuous-type distribution with density $\varphi(x,\boldsymbol{w})$ where vector $\boldsymbol{w}$ collects the parameters involved. Let $\mathcal{D}$ be a data set $\mathcal{D} = \{x_1, x_2, \ldots, x_N\}$ that are drawn independently from the same distribution. In literature, a data set of this type is said to be *independently and identically distributed* (i.i.d.). Because of its probabilistic independence, given parameter $\boldsymbol{w}$ the probability density of an i.i.d. data set $\mathcal{D}$ assumes the form

$$p(\mathcal{D} \mid \boldsymbol{w}) = \prod_{i=1}^{N} \varphi(x_i, \boldsymbol{w}) \tag{B.49}$$

which is called *likelihood function* of the probability distribution. A popular approach for determining the parameters of a probability distribution using an observed data is to maximize the likelihood function with respect to the parameters. Because logarithm is a monotonically increasing function and logarithm of the likelihood function simplifies subsequent mathematical analysis, one maximizes the log-likelihood instead:

$$\log p(\mathcal{D} \mid \boldsymbol{w}) = \sum_{i=1}^{n} \log \varphi(x_i, \boldsymbol{w}) \tag{B.50}$$

Example B.12

Consider the 1-D Gaussian distribution in (B.41), i.e.

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

Following (B.50), its log-likelihood function for an observed i.i.d. data set, denoted by $L(\mu,\sigma^2,\mathcal{D})$, is given by

$$L(\mu,\sigma^2,\mathcal{D}) = -\frac{1}{2\sigma^2}\sum_{i=1}^{N}(x_i-\mu)^2 - \frac{N}{2}\log\sigma^2 - \frac{N}{2}\log(2\pi)$$

Maximizing $L(\mu,\sigma^2,\mathcal{D})$ with respect to $\{\mu,\sigma^2\}$ yields *maximum likelihood* (ML) estimates

$$\mu_{ML} = \frac{1}{N}\sum_{i=1}^{N} x_i \quad \text{and} \quad \sigma_{ML}^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i-\mu_{ML})^2 \tag{B.51}$$