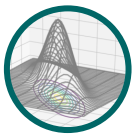
6

Probability and Distributions



Probability, loosely speaking, concerns the study of uncertainty. Probability can be thought of as the fraction of times an event occurs, or as a degree of belief about an event. We then would like to use this probability to measure the chance of something occurring in an experiment. As mentioned in Chapter 1, we often quantify uncertainty in the data, uncertainty in the machine learning model, and uncertainty in the predictions produced by the model. Quantifying uncertainty requires the idea of a *random variable*, which is a function that maps outcomes of random experiments to a set of properties that we are interested in. Associated with the random variable is a function that measures the probability that a particular outcome (or set of outcomes) will occur; this is called the *probability distribution*.

random variable

probability distribution

Probability distributions are used as a building block for other concepts, such as probabilistic modeling (Section 8.4), graphical models (Section 8.5), and model selection (Section 8.6). In the next section, we present the three concepts that define a probability space (the sample space, the events, and the probability of an event) and how they are related to a fourth concept called the random variable. The presentation is deliberately slightly hand wavy since a rigorous presentation may occlude the intuition behind the concepts. An outline of the concepts presented in this chapter are shown in Figure 6.1.

6.1 Construction of a Probability Space

The theory of probability aims at defining a mathematical structure to describe random outcomes of experiments. For example, when tossing a single coin, we cannot determine the outcome, but by doing a large number of coin tosses, we can observe a regularity in the average outcome. Using this mathematical structure of probability, the goal is to perform automated reasoning, and in this sense, probability generalizes logical reasoning (Jaynes, 2003).

6.1.1 Philosophical Issues

When constructing automated reasoning systems, classical Boolean logic does not allow us to express certain forms of plausible reasoning. Consider

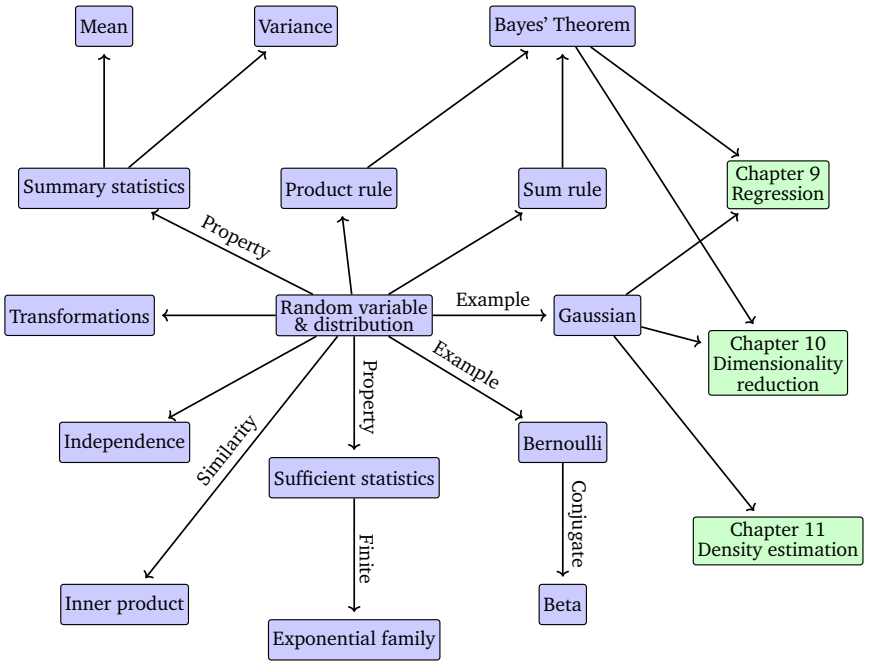


Figure 6.1 A mind map of the concepts related to random variables and probability distributions, as described in this chapter.

the following scenario: We observe that A is false. We find B becomes less plausible, although no conclusion can be drawn from classical logic. We observe that B is true. It seems A becomes more plausible. We use this form of reasoning daily. We are waiting for a friend, and consider three possibilities: H1, she is on time; H2, she has been delayed by traffic; and H3, she has been abducted by aliens. When we observe our friend is late, we must logically rule out H1. We also tend to consider H2 to be more likely, though we are not logically required to do so. Finally, we may consider H3 to be possible, but we continue to consider it quite unlikely. How do we conclude H2 is the most plausible answer? Seen in this way, probability theory can be considered a generalization of Boolean logic. In the context of machine learning, it is often applied in this way to formalize the design of automated reasoning systems. Further arguments about how probability theory is the foundation of reasoning systems can be found in Pearl (1988).

"For plausible reasoning it is necessary to extend the discrete true and false values of truth to continuous plausibilities" (Jaynes, 2003).

The philosophical basis of probability and how it should be somehow related to what we think should be true (in the logical sense) was studied by Cox (Jaynes, 2003). Another way to think about it is that if we are precise about our common sense we end up constructing probabilities. E. T. Jaynes (1922–1998) identified three mathematical criteria, which must apply to all plausibilities:

1. The degrees of plausibility are represented by real numbers.
2. These numbers must be based on the rules of common sense.

3. The resulting reasoning must be consistent, with the three following meanings of the word "consistent":
   (a) Consistency or non-contradiction: When the same result can be reached through different means, the same plausibility value must be found in all cases.
   (b) Honesty: All available data must be taken into account.
   (c) Reproducibility: If our state of knowledge about two problems are the same, then we must assign the same degree of plausibility to both of them.

The Cox–Jaynes theorem proves these plausibilities to be sufficient to define the universal mathematical rules that apply to plausibility p, up to transformation by an arbitrary monotonic function. Crucially, these rules *are* the rules of probability.

Remark. In machine learning and statistics, there are two major interpretations of probability: the Bayesian and frequentist interpretations (Bishop, 2006; Efron and Hastie, 2016). The Bayesian interpretation uses probability to specify the degree of uncertainty that the user has about an event. It is sometimes referred to as "subjective probability" or "degree of belief". The frequentist interpretation considers the relative frequencies of events of interest to the total number of events that occurred. The probability of an event is defined as the relative frequency of the event in the limit when one has infinite data. ♢

Some machine learning texts on probabilistic models use lazy notation and jargon, which is confusing. This text is no exception. Multiple distinct concepts are all referred to as "probability distribution", and the reader has to often disentangle the meaning from the context. One trick to help make sense of probability distributions is to check whether we are trying to model something categorical (a discrete random variable) or something continuous (a continuous random variable). The kinds of questions we tackle in machine learning are closely related to whether we are considering categorical or continuous models.

6.1.2 Probability and Random Variables

There are three distinct ideas that are often confused when discussing probabilities. First is the idea of a probability space, which allows us to quantify the idea of a probability. However, we mostly do not work directly with this basic probability space. Instead, we work with random variables (the second idea), which transfers the probability to a more convenient (often numerical) space. The third idea is the idea of a distribution or law associated with a random variable. We will introduce the first two ideas in this section and expand on the third idea in Section 6.2.

Modern probability is based on a set of axioms proposed by Kolmogorov

(Grinstead and Snell, 1997; Jaynes, 2003) that introduce the three concepts of sample space, event space, and probability measure. The probability space models a real-world process (referred to as an experiment) with random outcomes.

The sample space Ω

The *sample space* is the set of all possible outcomes of the experiment, usually denoted by Ω. For example, two successive coin tosses have a sample space of $\{\text{hh}, \text{tt}, \text{ht}, \text{th}\}$, where "h" denotes "heads" and "t" denotes "tails". sample space

The event space $\mathcal{A}$

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. The event space $\mathcal{A}$ is obtained by considering the collection of subsets of Ω, and for discrete probability distributions (Section 6.2.1) $\mathcal{A}$ is often the power set of Ω. event space

The probability P

With each event $A \in \mathcal{A}$, we associate 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. probability

The probability of a single event must lie in the interval $[0, 1]$, and the total probability over all outcomes in the sample space Ω must be 1, i.e., $P(\Omega) = 1$. Given a probability space $(\Omega, \mathcal{A}, P)$, we want to use it to model some real-world phenomenon. In machine learning, we often avoid explicitly referring to the probability space, but instead refer to probabilities on quantities of interest, which we denote by $\mathcal{T}$. In this book, we refer to $\mathcal{T}$ as the *target space* and refer to elements of $\mathcal{T}$ as states. We introduce a 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 association/mapping from Ω to $\mathcal{T}$ is called a *random variable*. For example, in the case of tossing two coins and counting the number of heads, a random variable X maps to the three possible outcomes: $X(\text{hh}) = 2$, $X(\text{ht}) = 1$, $X(\text{th}) = 1$, and $X(\text{tt}) = 0$. In this particular case, $\mathcal{T} = \{0, 1, 2\}$, and it is the probabilities on elements of $\mathcal{T}$ that we are interested in. For a finite sample space Ω and finite $\mathcal{T}$, the function corresponding to a random variable is essentially a lookup table. For any subset $S \subseteq \mathcal{T}$, we associate $P_X(S) \in [0, 1]$ (the probability) to a particular event occurring corresponding to the random variable X. Example 6.1 provides a concrete illustration of the terminology. target space random variable The name "random variable" is a great source of misunderstanding as it is neither random nor is it a variable. It is a function.

Remark. The aforementioned sample space Ω unfortunately is referred to by different names in different books. Another common name for Ω is "state space" (Jacod and Protter, 2004), but state space is sometimes reserved for referring to states in a dynamical system (Hasselblatt and

$A_1 = \{hh, ht\}$

$A_2 = \{th, tt, ht\}$

$A \subset \Omega \qquad A = \Omega$

Probability of an event A

$P : S(A) \longrightarrow [0,1]$

$S(A)$ → set of all the possible events

$P(A) \in [0,1]$

Properties:

1) $P(A) \geq 0$
2) $P(\Omega) = 1$
3) If $A_1 \cap A_2 \cap A_3 \dots \cap A_n = \{\emptyset\}$ then

$$P(A_1 \cup A_2 \dots \cup A_n) = \sum_{i=1}^{n} P(A_i)$$

How compute $P(A)$?

Equally likely model (EQM)

$$P(A) = \frac{\#\, A}{\#\, \Omega} = \frac{\#\ \text{favourable results}}{\#\ \text{possible results}}$$

1) Example: toss of two coins

$\Omega = \{hh, ht, th, tt\}$

$\rightarrow A = \{hh, ht\}$ $\rightarrow P(A)? \frac{\#\, A}{\#\, \Omega} = \frac{2}{4} = \frac{1}{2}$

Conditional probability of an event A conditioned by an event B

Example

Roll two dies. Compute the probability that the sum of the results is 6 if you know that one of the two dies has result 2.

A → "the sum is 6"

B → "one of the result is 2"

$\Omega = \{(i,j) \mid i,j = 1 \dots 6\}$ 36 elements

$A = \{(1,5), (2,4) (3,3), (4,2), (5,1)\}$ 5 elem.

$B = \{(2,1), (2,2), (2,3), (2,4), (2,5), (2,6)$ ←
$(2,2) (3,2), (4,2), (5,2), (6,2)\}$ 11 elem.

$P(A|B)$ → probability of A given B or "A conditioned To B

favourable cases in B are:

$(2,4), (4,2)$

$$P(A|B) = \frac{2}{11} = \frac{\#\text{ favourable cases}}{\#\text{ possible cases}}$$

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

Probability of A given B

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

Probability of B given A

Independent events

Events A and B are independent if $P(A \cap B) = P(A) \cdot P(B)$

Otherwise they are dependent.

Events $A_1, A_2 \cdots A_n$ are independent if

$$P(A_1 \cap A_2 \cdots \cap A_n) = \prod_{i=1}^{n} P(A_i)$$

$$A \cap B = \{(2,4), (4,2)\}$$

$$P(A \cap B) = \frac{\#(A \cap B)}{\#\,\Omega} = \frac{2}{36}$$

$$P(B) = \frac{\#\,B}{\#\Omega} = \frac{11}{36}$$

$$P(A|B) = \frac{\frac{2}{36}}{\frac{11}{36}} = \frac{2}{11}$$

RANDOM VARIABLE

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

$$\omega \in \Omega, \quad X(\omega) = x$$

T Target space $= \{x \in \mathbb{R} \mid X(\omega) = x\} \subset \mathbb{R}$

Example -
Tossing a coin twice

$\Omega = \{hh, ht, th, tt\}$

$X(\omega) = \{$number of $h\}$

$X(hh) = 2$
$X(ht) = 1$
$X(th) = 1$
$X(tt) = \phi$

$T = \{\phi, 1, 2\}$

DISCRETE RANDOM VARIABLE if the Target space T is a finite or countable set.

Example.
Repeat tossing a die until you obtain 6

$Y = \{$number of die tossing$\}$

$Y = \{1, 2, 3, 4 \ldots\}$

CONTINOUS RANDOM VARIABLE

when the target space T is an interval or the union of intervals.

Example → clients enters a bank

$Z = \{$ time before the first client enter the bank$\}$ $T = (0, 480)$

DISCRETE RV → PROBABILITY MASS FUNCTION

Associated to each discrete RV X the Probability Mass Function (PMF):

$$f_X : T_X \longrightarrow [0,1]$$

$$f_X(x) = \mathbb{P}(X = x)$$

Example:

$T = \{\phi, 1, 2\}$ $\quad X(hh) = 2$

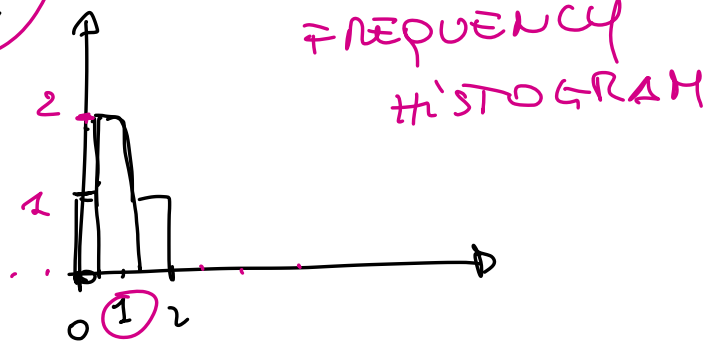
$X(th) = 1 \quad X(ht) = 1 \quad X(tt) = 0$

$$f_X : T \longrightarrow [0,1]$$

$$f_X(\phi) = \mathbb{P}(X = \phi) = \frac{1}{4}$$

$$f_X(1) = \frac{2}{4} = \frac{1}{2}$$

$$f_X(2) = \frac{1}{4}$$



FREQUENCY HISTOGRAM

Properties:

1. $f_X(x) \geq 0 \quad \forall x \in T$

2. $\sum_{x \in T} f_X(x) = 1$

POPULATION MEAN AND VARIANCE

MEAN

$$\mu = \mathbb{E}(X) = \sum_{x \in T} x \, f_X(x)$$

previous experiment $T = \{0, 1, 2\}$

$$\mu = 0 \cdot f_X(0) + 1 \cdot f_X(1) + 2 \, f_X(2) =$$

$$= 0 \cdot \frac{1}{4} + 1 \cdot \frac{1}{2} + 2 \cdot \frac{1}{4} = 1$$

VARIANCE

$$\sigma^2 = \sum_{x \in T} (x - \underbrace{\mu}_{1})^2 \cdot f_X(x)$$

$$\sigma^2 = (0-1)^2 \cdot f_X(0) + (1-1)^2 f_X(1) +$$

$$(2-1)^2 \cdot f_X(2) = 1 \cdot \frac{1}{4} + 0 \cdot \frac{1}{2} + 1 \cdot \frac{1}{4} =$$

$$= \frac{2}{4} = \frac{1}{2}$$

UNIFORM DISTRIBUTION

PMF: $f_X = \frac{1}{n}$

$n = \#T$ = number of elements of the target space

POISSON DISTRIBUTION

"rare events" $T = \{0, 1, 2, \ldots\}$

parameter λ → mean of the events in the unit of time

Example

Arrivals of clients in a bank

$X = \{$ number of clients in the time unit (1 hour) $\}$ discrete RV

$T = \{0, 1, 2, \ldots\}$

$$f_X(x) = e^{-\lambda} \frac{\lambda^x}{x!} \qquad \forall x = 0, 1, 2, \ldots$$

$f_X(x; \lambda)$

$$\mu = \sigma^2 = \lambda$$

Example: $\lambda = 20$

$$P(X = 16) = f_X(16) = e^{-20} \frac{20^{16}}{16!}$$

Probability of 16 clients in 2 hours: $\lambda = 20 \cdot 2 \rightarrow f_X(16) = e^{-40} \frac{40^{16}}{16!}$

CONTINUOUS RV → PROBABILITY DENSITY FUNCTION

X continuous RV, the probability density function (PDF)

$$f_X : T \to \mathbb{R}$$

(T interval = $[a,b]$)

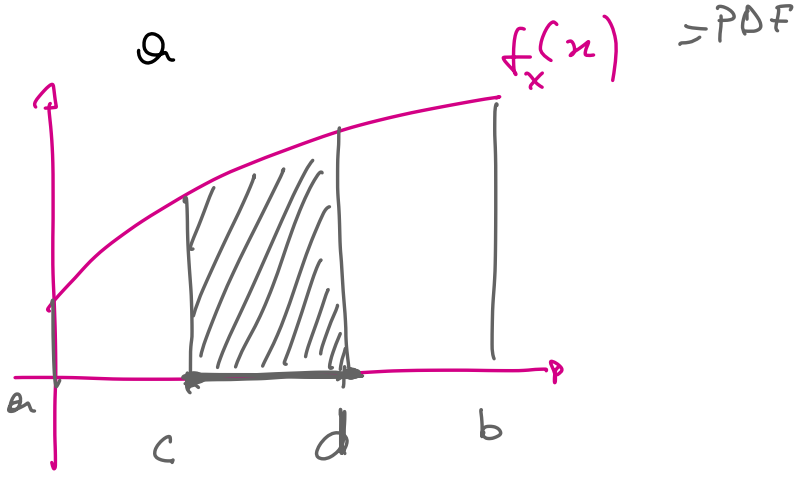
$$P(x \in [c,d]) = \int_c^d f_X(x)\, dx$$

~~$f_X(x)$?~~ $[c,d] \subset [a,b]$

Properties

1) $f_X(x) > 0 \quad \forall x \in T$

2) $$\int_{x \in T} f_X(x)\, dx = \int_a^b f_X(x)\, dx = 1$$



POPULATION MEAN & VARIANCE

MEAN $$\mu = \int_a^b x f_x(x) dx = \int_{x \in T} x f_x(x) dx$$

VARIANCE $$\sigma^2 = \int_a^b (x-\mu)^2 f_x(x) dx$$

STANDARD DEVIATION $$\sigma = \sqrt{\sigma^2}$$

CONTINUOUS UNIFORM distribution

$$f_x = \frac{1}{b-a} \qquad T = [a,b]$$

GAUSSIAN DISTRIBUTION

→ $$f_x(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right) x \in R \ (-\infty, \infty)$$

μ, σ → mean and standard deviation

$$f_x(x; \mu, \sigma)$$

UNIVARIATE DISTRIBUTION

ONE RANDOM VARIABLE

BIVARIATE DISTRIBUTION

TWO RANDOM VARIABLES

MULTIVARIATE DISTRIBUTIONS

MULTIVARIATE RV

Katok, 2003). Other names sometimes used to describe Ω are: "sample description space", "possibility space," and "event space". ♢

This toy example is essentially a biased coin flip example.

Example 6.1

We assume that the reader is already familiar with computing probabilities of intersections and unions of sets of events. A gentler introduction to probability with many examples can be found in chapter 2 of Walpole et al. (2011).

Consider a statistical experiment where we model a funfair game consisting of drawing two coins from a bag (with replacement). There are coins from USA (denoted as \$) and UK (denoted as £) in the bag, and since we draw two coins from the bag, there are four outcomes in total. The state space or sample space Ω of this experiment is then (\$, \$), (\$, £), (£, \$), (£, £). Let us assume that the composition of the bag of coins is such that a draw returns at random a \$ with probability 0.3.

The event we are interested in is the total number of times the repeated draw returns \$. Let us define a random variable X that maps the sample space Ω to $\mathcal{T}$, which denotes the number of times we draw \$ out of the bag. We can see from the preceding sample space we can get zero \$, one \$, or two \$s, and therefore $\mathcal{T} = \{0, 1, 2\}$. The random variable X (a function or lookup table) can be represented as a table like the following:

$$X((\$, \$)) = 2 \tag{6.1}$$
$$X((\$, £)) = 1 \tag{6.2}$$
$$X((£, \$)) = 1 \tag{6.3}$$
$$X((£, £)) = 0\,. \tag{6.4}$$

Since we return the first coin we draw before drawing the second, this implies that the two draws are independent of each other, which we will discuss in Section 6.4.5. Note that there are two experimental outcomes, which map to the same event, where only one of the draws returns \$. Therefore, the probability mass function (Section 6.2.1) of X is given by

$$\begin{aligned} P(X = 2) &= P((\$, \$)) \\ &= P(\$) \cdot P(\$) \\ &= 0.3 \cdot 0.3 = 0.09 \end{aligned} \tag{6.5}$$
$$\begin{aligned} P(X = 1) &= P((\$, £) \cup (£, \$)) \\ &= P((\$, £)) + P((£, \$)) \\ &= 0.3 \cdot (1 - 0.3) + (1 - 0.3) \cdot 0.3 = 0.42 \end{aligned} \tag{6.6}$$
$$\begin{aligned} P(X = 0) &= P((£, £)) \\ &= P(£) \cdot P(£) \\ &= (1 - 0.3) \cdot (1 - 0.3) = 0.49\,. \end{aligned} \tag{6.7}$$

In the calculation, we equated two different concepts, the probability of the output of X and the probability of the samples in Ω. For example, in (6.7) we say $P(X = 0) = P((\pounds, \pounds))$. Consider the random variable $X : \Omega \to \mathcal{T}$ and a subset $S \subseteq \mathcal{T}$ (for example, a single element of $\mathcal{T}$, such as the outcome that one head is obtained when tossing two coins). Let $X^{-1}(S)$ be the pre-image of S by X, i.e., the set of elements of Ω that map to S under X; $\{\omega \in \Omega : X(\omega) \in S\}$. One way to understand the transformation of probability from events in Ω via the random variable X is to associate it with the probability of the pre-image of S (Jacod and Protter, 2004). For $S \subseteq \mathcal{T}$, we have the notation

$$P_X(S) = P(X \in S) = P(X^{-1}(S)) = P(\{\omega \in \Omega : X(\omega) \in S\}) . \quad (6.8)$$

The left-hand side of (6.8) is the probability of the set of possible outcomes (e.g., number of \$ = 1) that we are interested in. Via the random variable X, which maps states to outcomes, we see in the right-hand side of (6.8) that this is the probability of the set of states (in Ω) that have the property (e.g., $\$\pounds$, $\pounds\$$). We say that a random variable X is distributed according to a particular probability distribution P_X, which defines the probability mapping between the event and the probability of the outcome of the random variable. In other words, the function P_X or equivalently $P \circ X^{-1}$ is the *law* or *distribution* of random variable X.

law
distribution

Remark. The target space, that is, the range $\mathcal{T}$ of the random variable X, is used to indicate the kind of probability space, i.e., a $\mathcal{T}$ random variable. When $\mathcal{T}$ is finite or countably infinite, this is called a discrete random variable (Section 6.2.1). For continuous random variables (Section 6.2.2), we only consider $\mathcal{T} = \mathbb{R}$ or $\mathcal{T} = \mathbb{R}^D$. ♢

6.1.3 Statistics

Probability theory and statistics are often presented together, but they concern different aspects of uncertainty. One way of contrasting them is by the kinds of problems that are considered. Using probability, we can consider a model of some process, where the underlying uncertainty is captured by random variables, and we use the rules of probability to derive what happens. In statistics, we observe that something has happened and try to figure out the underlying process that explains the observations. In this sense, machine learning is close to statistics in its goals to construct a model that adequately represents the process that generated the data. We can use the rules of probability to obtain a "best-fitting" model for some data.

Another aspect of machine learning systems is that we are interested in generalization error (see Chapter 8). This means that we are actually interested in the performance of our system on instances that we will observe in future, which are not identical to the instances that we have

seen so far. This analysis of future performance relies on probability and statistics, most of which is beyond what will be presented in this chapter. The interested reader is encouraged to look at the books by Boucheron et al. (2013) and Shalev-Shwartz and Ben-David (2014). We will see more about statistics in Chapter 8.

6.2 Discrete and Continuous Probabilities

Let us focus our attention on ways to describe the probability of an event as introduced in Section 6.1. Depending on whether the target space is discrete or continuous, the natural way to refer to distributions is different. When the target space $\mathcal{T}$ is discrete, we can specify the probability that a random variable X takes a particular value $x \in \mathcal{T}$, denoted as $P(X = x)$. The expression $P(X = x)$ for a discrete random variable X is known as the *probability mass function*. When the target space $\mathcal{T}$ is continuous, e.g., the real line $\mathbb{R}$, it is more natural to specify the probability that a random variable X is in an interval, denoted by $P(a \leqslant X \leqslant b)$ for $a < b$. By convention, we specify the probability that a random variable X is less than a particular value x, denoted by $P(X \leqslant x)$. The expression $P(X \leqslant x)$ for a continuous random variable X is known as the *cumulative distribution function*. We will discuss continuous random variables in Section 6.2.2. We will revisit the nomenclature and contrast discrete and continuous random variables in Section 6.2.3.

probability mass function

cumulative distribution function

univariate

Remark. We will use the phrase *univariate* distribution to refer to distributions of a single random variable (whose states are denoted by non-bold x). We will refer to distributions of more than one random variable as *multivariate* distributions, and will usually consider a vector of random variables (whose states are denoted by bold $\boldsymbol{x}$). ♢

multivariate

6.2.1 Discrete Probabilities

When the target space is discrete, we can imagine the probability distribution of multiple random variables as filling out a (multidimensional) array of numbers. Figure 6.2 shows an example. The target space of the joint probability is the Cartesian product of the target spaces of each of the random variables. We define the *joint probability* as the entry of both values jointly

joint probability

$$P(X = x_i, Y = y_j) = \frac{n_{ij}}{N}, \tag{6.9}$$

where n_{ij} is the number of events with state x_i and y_j and N the total number of events. The joint probability is the probability of the intersection of both events, that is, $P(X = x_i, Y = y_j) = P(X = x_i \cap Y = y_j)$. Figure 6.2 illustrates the *probability mass function* (pmf) of a discrete probability distribution. For two random variables X and Y, the probability

probability mass function

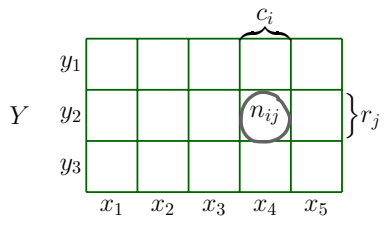


Figure 6.2 Visualization of a discrete bivariate probability mass function, with random variables X and Y. This diagram is adapted from Bishop (2006).

that $X = x$ and $Y = y$ is (lazily) written as $p(x, y)$ and is called the joint probability. One can think of a probability as a function that takes state x and y and returns a real number, which is the reason we write $p(x, y)$. The *marginal probability* that X takes the value x irrespective of the value of random variable Y is (lazily) written as $p(x)$. We write $X \sim p(x)$ to denote that the random variable X is distributed according to $p(x)$. If we consider only the instances where $X = x$, then the fraction of instances (the *conditional probability*) for which $Y = y$ is written (lazily) as $p(y \mid x)$.

marginal probability

conditional probability

Example 6.2

Consider two random variables X and Y, where X has five possible states and Y has three possible states, as shown in Figure 6.2. We denote by n_{ij} the number of events with state $X = x_i$ and $Y = y_j$, and denote by N the total number of events. The value c_i is the sum of the individual frequencies for the ith column, that is, $c_i = \sum_{j=1}^{3} n_{ij}$. Similarly, the value r_j is the row sum, that is, $r_j = \sum_{i=1}^{5} n_{ij}$. Using these definitions, we can compactly express the distribution of X and Y.

The probability distribution of each random variable, the marginal probability, can be seen as the sum over a row or column

$$P(X = x_i) = \frac{c_i}{N} = \frac{\sum_{j=1}^{3} n_{ij}}{N} \tag{6.10}$$

and

$$P(Y = y_j) = \frac{r_j}{N} = \frac{\sum_{i=1}^{5} n_{ij}}{N}, \tag{6.11}$$

where c_i and r_j are the ith column and jth row of the probability table, respectively. By convention, for discrete random variables with a finite number of events, we assume that probabilties sum up to one, that is,

$$\sum_{i=1}^{5} P(X = x_i) = 1 \quad \text{and} \quad \sum_{j=1}^{3} P(Y = y_j) = 1\,. \tag{6.12}$$

The conditional probability is the fraction of a row or column in a par-

joint probability of X and Y

$$f_{XY} : T_X \times T_Y \longrightarrow [0,1]$$

$$f_{XY}(x_i, y_j) = P(X = x_i, Y = y_j) = n_{ij}$$

marginal probability of X

$$p(x) = P(X = x_i) = \frac{c_i}{N}$$

marginal probability of Y

$$p(y) = P(Y = y_j) = \frac{r_j}{N}$$

conditional probabilities

$$P(X = x_i \mid Y = y_i) \;\&\; P(Y = y_j \mid X = x_i)$$

$$P(Y = y_j \mid X = x_i) = \frac{n_{ij}}{c_i}$$

$$P(X = x_i \mid Y = y_j) = \frac{n_{ij}}{r_j}$$

ticular cell. For example, the conditional probability of Y given X is

$$P(Y = y_j \,|\, X = x_i) = \frac{n_{ij}}{c_i}\,, \tag{6.13}$$

and the conditional probability of X given Y is

$$P(X = x_i \,|\, Y = y_j) = \frac{n_{ij}}{r_j}\,. \tag{6.14}$$

categorical variable

In machine learning, we use discrete probability distributions to model *categorical variables*, i.e., variables that take a finite set of unordered values. They could be categorical features, such as the degree taken at university when used for predicting the salary of a person, or categorical labels, such as letters of the alphabet when doing handwriting recognition. Discrete distributions are also often used to construct probabilistic models that combine a finite number of continuous distributions (Chapter 11).

6.2.2 Continuous Probabilities

We consider real-valued random variables in this section, i.e., we consider target spaces that are intervals of the real line $\mathbb{R}$. In this book, we pretend that we can perform operations on real random variables as if we have discrete probability spaces with finite states. However, this simplification is not precise for two situations: when we repeat something infinitely often, and when we want to draw a point from an interval. The first situation arises when we discuss generalization errors in machine learning (Chapter 8). The second situation arises when we want to discuss continuous distributions, such as the Gaussian (Section 6.5). For our purposes, the lack of precision allows for a briefer introduction to probability.

Remark. In continuous spaces, there are two additional technicalities, which are counterintuitive. First, the set of all subsets (used to define the event space $\mathcal{A}$ in Section 6.1) is not well behaved enough. $\mathcal{A}$ needs to be restricted to behave well under set complements, set intersections, and set unions. Second, the size of a set (which in discrete spaces can be obtained by counting the elements) turns out to be tricky. The size of a set is called its *measure*. For example, the cardinality of discrete sets, the length of an interval in $\mathbb{R}$, and the volume of a region in $\mathbb{R}^d$ are all measures. Sets that behave well under set operations and additionally have a topology are called a *Borel σ-algebra*. Betancourt details a careful construction of probability spaces from set theory without being bogged down in technicalities; see `https://tinyurl.com/yb3t6mfd`. For a more precise construction, we refer to Billingsley (1995) and Jacod and Protter (2004).

measure

Borel σ-algebra

In this book, we consider real-valued random variables with their cor-

responding Borel σ-algebra. We consider random variables with values in $\mathbb{R}^D$ to be a vector of real-valued random variables. ♢

Definition 6.1 (Probability Density Function)**.** A function $f : \mathbb{R}^D \to \mathbb{R}$ is called a *probability density function* (*pdf*) if

probability density function
pdf

1. $\forall \boldsymbol{x} \in \mathbb{R}^D : f(\boldsymbol{x}) \geqslant 0$
2. Its integral exists and

$$\int_{\mathbb{R}^D} f(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} = 1\,. \tag{6.15}$$

For probability mass functions (pmf) of discrete random variables, the integral in (6.15) is replaced with a sum (6.12).

Observe that the probability density function is any function f that is non-negative and integrates to one. We associate a random variable X with this function f by

$$P(a \leqslant X \leqslant b) = \int_a^b f(x) \mathrm{d}x\,, \tag{6.16}$$

where $a, b \in \mathbb{R}$ and $x \in \mathbb{R}$ are outcomes of the continuous random variable X. States $\boldsymbol{x} \in \mathbb{R}^D$ are defined analogously by considering a vector of $x \in \mathbb{R}$. This association (6.16) is called the *law* or *distribution* of the random variable X.

law

$P(X = x)$ is a set of measure zero.

Remark. In contrast to discrete random variables, the probability of a continuous random variable X taking a particular value $P(X = x)$ is zero. This is like trying to specify an interval in (6.16) where $a = b$. ♢

Definition 6.2 (Cumulative Distribution Function)**.** A *cumulative distribution function* (cdf) of a multivariate real-valued random variable X with states $\boldsymbol{x} \in \mathbb{R}^D$ is given by

cumulative distribution function

$$F_X(\boldsymbol{x}) = P(X_1 \leqslant x_1, \ldots, X_D \leqslant x_D)\,, \tag{6.17}$$

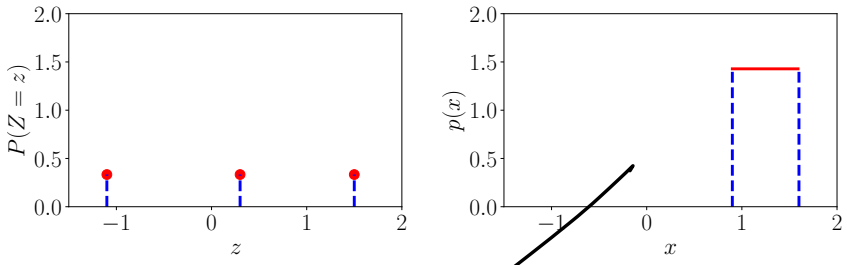
where $X = [X_1, \ldots, X_D]^\top$, $\boldsymbol{x} = [x_1, \ldots, x_D]^\top$, and the right-hand side represents the probability that random variable X_i takes the value smaller than or equal to x_i.

There are cdfs, which do not have corresponding pdfs.

The cdf can be expressed also as the integral of the probability density function $f(\boldsymbol{x})$ so that

$$F_X(\boldsymbol{x}) = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_D} f(z_1, \ldots, z_D) \mathrm{d}z_1 \cdots \mathrm{d}z_D\,. \tag{6.18}$$

Remark. We reiterate that there are in fact two distinct concepts when talking about distributions. First is the idea of a pdf (denoted by $f(x)$), which is a nonnegative function that sums to one. Second is the law of a random variable X, that is, the association of a random variable X with the pdf $f(x)$. ♢



(a) Discrete distribution (b) Continuous distribution

Figure 6.3 Examples of (a) discrete and (b) continuous uniform distributions. See Example 6.3 for details of the distributions.

For most of this book, we will not use the notation $f(x)$ and $F_X(x)$ as we mostly do not need to distinguish between the pdf and cdf. However, we will need to be careful about pdfs and cdfs in Section 6.7.

6.2.3 Contrasting Discrete and Continuous Distributions

Recall from Section 6.1.2 that probabilities are positive and the total probability sums up to one. For discrete random variables (see (6.12)), this implies that the probability of each state must lie in the interval $[0, 1]$. However, for continuous random variables the normalization (see (6.15)) does not imply that the value of the density is less than or equal to 1 for all values. We illustrate this in Figure 6.3 using the *uniform distribution* for both discrete and continuous random variables.

uniform distribution

Example 6.3

We consider two examples of the uniform distribution, where each state is equally likely to occur. This example illustrates some differences between discrete and continuous probability distributions.

Let Z be a discrete uniform random variable with three states $\{z = -1.1, z = 0.3, z = 1.5\}$. The probability mass function can be represented as a table of probability values:

z	-1.1	0.3	1.5
$P(Z = z)$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$

The actual values of these states are not meaningful here, and we deliberately chose numbers to drive home the point that we do not want to use (and should ignore) the ordering of the states.

Alternatively, we can think of this as a graph (Figure 6.3(a)), where we use the fact that the states can be located on the x-axis, and the y-axis represents the probability of a particular state. The y-axis in Figure 6.3(a) is deliberately extended so that is it the same as in Figure 6.3(b).

Let X be a continuous random variable taking values in the range $0.9 \leqslant X \leqslant 1.6$, as represented by Figure 6.3(b). Observe that the height of the

Type	"Point probability"	"Interval probability"
Discrete	$P(X = x)$ Probability mass function	Not applicable
Continuous	$p(x)$ Probability density function	$P(X \leqslant x)$ Cumulative distribution function

Table 6.1 Nomenclature for probability distributions.

density can be greater than 1. However, it needs to hold that

$$\int_{0.9}^{1.6} p(x)\mathrm{d}x = 1\,. \tag{6.19}$$

Remark. There is an additional subtlety with regards to discrete probability distributions. The states $z_1, \ldots, z_d$ do not in principle have any structure, i.e., there is usually no way to compare them, for example $z_1 = \text{red}, z_2 = \text{green}, z_3 = \text{blue}$. However, in many machine learning applications discrete states take numerical values, e.g., $z_1 = -1.1, z_2 = 0.3, z_3 = 1.5$, where we could say $z_1 < z_2 < z_3$. Discrete states that assume numerical values are particularly useful because we often consider expected values (Section 6.4.1) of random variables. ♢

Unfortunately, machine learning literature uses notation and nomenclature that hides the distinction between the sample space Ω, the target space $\mathcal{T}$, and the random variable X. For a value x of the set of possible outcomes of the random variable X, i.e., $x \in \mathcal{T}$, $p(x)$ denotes the probability that random variable X has the outcome x. For discrete random variables, this is written as $P(X = x)$, which is known as the probability mass function. The pmf is often referred to as the "distribution". For continuous variables, $p(x)$ is called the probability density function (often referred to as a density). To muddy things even further, the cumulative distribution function $P(X \leqslant x)$ is often also referred to as the "distribution". In this chapter, we will use the notation X to refer to both univariate and multivariate random variables, and denote the states by x and $\boldsymbol{x}$ respectively. We summarize the nomenclature in Table 6.1.

We think of the outcome x as the argument that results in the probability $p(x)$.

Remark. We will be using the expression "probability distribution" not only for discrete probability mass functions but also for continuous probability density functions, although this is technically incorrect. In line with most machine learning literature, we also rely on context to distinguish the different uses of the phrase probability distribution. ♢

6.3 Sum Rule, Product Rule, and Bayes' Theorem

We think of probability theory as an extension to logical reasoning. As we discussed in Section 6.1.1, the rules of probability presented here follow

naturally from fulfilling the desiderata (Jaynes, 2003, chapter 2). Probabilistic modeling (Section 8.4) provides a principled foundation for designing machine learning methods. Once we have defined probability distributions (Section 6.2) corresponding to the uncertainties of the data and our problem, it turns out that there are only two fundamental rules, the sum rule and the product rule.

Recall from (6.9) that $p(\boldsymbol{x}, \boldsymbol{y})$ is the joint distribution of the two random variables $\boldsymbol{x}, \boldsymbol{y}$. The distributions $p(\boldsymbol{x})$ and $p(\boldsymbol{y})$ are the corresponding marginal distributions, and $p(\boldsymbol{y} \mid \boldsymbol{x})$ is the conditional distribution of $\boldsymbol{y}$ given $\boldsymbol{x}$. Given the definitions of the marginal and conditional probability for discrete and continuous random variables in Section 6.2, we can now present the two fundamental rules in probability theory.

These two rules arise naturally (Jaynes, 2003) from the requirements we discussed in Section 6.1.1.

The first rule, the *sum rule*, states that

$$p(\boldsymbol{x}) = \begin{cases} \sum_{\boldsymbol{y} \in \mathcal{Y}} p(\boldsymbol{x}, \boldsymbol{y}) & \text{if } \boldsymbol{y} \text{ is discrete} \\ \int_{\mathcal{Y}} p(\boldsymbol{x}, \boldsymbol{y}) \mathrm{d}\boldsymbol{y} & \text{if } \boldsymbol{y} \text{ is continuous} \end{cases}, \tag{6.20}$$

where $\mathcal{Y}$ are the states of the target space of random variable Y. This means that we sum out (or integrate out) the set of states $\boldsymbol{y}$ of the random variable Y. The sum rule is also known as the *marginalization property*. The sum rule relates the joint distribution to a marginal distribution. In general, when the joint distribution contains more than two random variables, the sum rule can be applied to any subset of the random variables, resulting in a marginal distribution of potentially more than one random variable. More concretely, if $\boldsymbol{x} = [x_1, \ldots, x_D]^\top$, we obtain the marginal

marginalization property

$$p(x_i) = \int p(x_1, \ldots, x_D) \mathrm{d}\boldsymbol{x}_{\setminus i} \tag{6.21}$$

by repeated application of the sum rule where we integrate/sum out all random variables except x_i, which is indicated by $\setminus i$, which reads "all except i."

Remark. Many of the computational challenges of probabilistic modeling are due to the application of the sum rule. When there are many variables or discrete variables with many states, the sum rule boils down to performing a high-dimensional sum or integral. Performing high-dimensional sums or integrals is generally computationally hard, in the sense that there is no known polynomial-time algorithm to calculate them exactly. ♢

The second rule, known as the *product rule*, relates the joint distribution to the conditional distribution via

product rule

$$p(\boldsymbol{x}, \boldsymbol{y}) = p(\boldsymbol{y} \mid \boldsymbol{x}) p(\boldsymbol{x}) . \tag{6.22}$$

The product rule can be interpreted as the fact that every joint distribution of two random variables can be factorized (written as a product)

of two other distributions. The two factors are the marginal distribution of the first random variable $p(\boldsymbol{x})$, and the conditional distribution of the second random variable given the first $p(\boldsymbol{y} \mid \boldsymbol{x})$. Since the ordering of random variables is arbitrary in $p(\boldsymbol{x}, \boldsymbol{y})$, the product rule also implies $p(\boldsymbol{x}, \boldsymbol{y}) = p(\boldsymbol{x} \mid \boldsymbol{y})p(\boldsymbol{y})$. To be precise, (6.22) is expressed in terms of the probability mass functions for discrete random variables. For continuous random variables, the product rule is expressed in terms of the probability density functions (Section 6.2.3).

In machine learning and Bayesian statistics, we are often interested in making inferences of unobserved (latent) random variables given that we have observed other random variables. Let us assume we have some prior knowledge $p(\boldsymbol{x})$ about an unobserved random variable $\boldsymbol{x}$ and some relationship $p(\boldsymbol{y} \mid \boldsymbol{x})$ between $\boldsymbol{x}$ and a second random variable $\boldsymbol{y}$, which we can observe. If we observe $\boldsymbol{y}$, we can use Bayes' theorem to draw some conclusions about $\boldsymbol{x}$ given the observed values of $\boldsymbol{y}$. *Bayes' theorem* (also *Bayes' rule* or *Bayes' law*)

Bayes' theorem
Bayes' rule
Bayes' law

$$\underbrace{p(\boldsymbol{x} \mid \boldsymbol{y})}_{\text{posterior}} = \frac{\overbrace{p(\boldsymbol{y} \mid \boldsymbol{x})}^{\text{likelihood}} \overbrace{p(\boldsymbol{x})}^{\text{prior}}}{\underbrace{p(\boldsymbol{y})}_{\text{evidence}}} \tag{6.23}$$

is a direct consequence of the product rule in (6.22) since

$$p(\boldsymbol{x}, \boldsymbol{y}) = p(\boldsymbol{x} \mid \boldsymbol{y})p(\boldsymbol{y}) \tag{6.24}$$

and

$$p(\boldsymbol{x}, \boldsymbol{y}) = p(\boldsymbol{y} \mid \boldsymbol{x})p(\boldsymbol{x}) \tag{6.25}$$

so that

$$p(\boldsymbol{x} \mid \boldsymbol{y})p(\boldsymbol{y}) = p(\boldsymbol{y} \mid \boldsymbol{x})p(\boldsymbol{x}) \iff p(\boldsymbol{x} \mid \boldsymbol{y}) = \frac{p(\boldsymbol{y} \mid \boldsymbol{x})p(\boldsymbol{x})}{p(\boldsymbol{y})} . \tag{6.26}$$

In (6.23), $p(\boldsymbol{x})$ is the *prior*, which encapsulates our subjective prior knowledge of the unobserved (latent) variable $\boldsymbol{x}$ before observing any data. We can choose any prior that makes sense to us, but it is critical to ensure that the prior has a nonzero pdf (or pmf) on all plausible $\boldsymbol{x}$, even if they are very rare.

prior

The *likelihood* $p(\boldsymbol{y} \mid \boldsymbol{x})$ describes how $\boldsymbol{x}$ and $\boldsymbol{y}$ are related, and in the case of discrete probability distributions, it is the probability of the data $\boldsymbol{y}$ if we were to know the latent variable $\boldsymbol{x}$. Note that the likelihood is not a distribution in $\boldsymbol{x}$, but only in $\boldsymbol{y}$. We call $p(\boldsymbol{y} \mid \boldsymbol{x})$ either the "likelihood of $\boldsymbol{x}$ (given $\boldsymbol{y}$)" or the "probability of $\boldsymbol{y}$ given $\boldsymbol{x}$" but never the likelihood of $\boldsymbol{y}$ (MacKay, 2003).

likelihood
The likelihood is sometimes also called the "measurement model".

The *posterior* $p(\boldsymbol{x} \mid \boldsymbol{y})$ is the quantity of interest in Bayesian statistics because it expresses exactly what we are interested in, i.e., what we know about $\boldsymbol{x}$ after having observed $\boldsymbol{y}$.

posterior

The quantity

$$p(\boldsymbol{y}) := \int p(\boldsymbol{y} \mid \boldsymbol{x}) p(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} = \mathbb{E}_X[p(\boldsymbol{y} \mid \boldsymbol{x})] \tag{6.27}$$

marginal likelihood
evidence

is the *marginal likelihood/evidence*. The right-hand side of (6.27) uses the expectation operator which we define in Section 6.4.1. By definition, the marginal likelihood integrates the numerator of (6.23) with respect to the latent variable $\boldsymbol{x}$. Therefore, the marginal likelihood is independent of $\boldsymbol{x}$, and it ensures that the posterior $p(\boldsymbol{x} \mid \boldsymbol{y})$ is normalized. The marginal likelihood can also be interpreted as the expected likelihood where we take the expectation with respect to the prior $p(\boldsymbol{x})$. Beyond normalization of the posterior, the marginal likelihood also plays an important role in Bayesian model selection, as we will discuss in Section 8.6. Due to the integration in (8.44), the evidence is often hard to compute.

Bayes' theorem is also called the "probabilistic inverse."
probabilistic inverse

Bayes' theorem (6.23) allows us to invert the relationship between $\boldsymbol{x}$ and $\boldsymbol{y}$ given by the likelihood. Therefore, Bayes' theorem is sometimes called the *probabilistic inverse*. We will discuss Bayes' theorem further in Section 8.4.

Remark. In Bayesian statistics, the posterior distribution is the quantity of interest as it encapsulates all available information from the prior and the data. Instead of carrying the posterior around, it is possible to focus on some statistic of the posterior, such as the maximum of the posterior, which we will discuss in Section 8.3. However, focusing on some statistic of the posterior leads to loss of information. If we think in a bigger context, then the posterior can be used within a decision-making system, and having the full posterior can be extremely useful and lead to decisions that are robust to disturbances. For example, in the context of model-based reinforcement learning, Deisenroth et al. (2015) show that using the full posterior distribution of plausible transition functions leads to very fast (data/sample efficient) learning, whereas focusing on the maximum of the posterior leads to consistent failures. Therefore, having the full posterior can be very useful for a downstream task. In Chapter 9, we will continue this discussion in the context of linear regression. ♢

6.4 Summary Statistics and Independence

We are often interested in summarizing sets of random variables and comparing pairs of random variables. A statistic of a random variable is a deterministic function of that random variable. The summary statistics of a distribution provide one useful view of how a random variable behaves, and as the name suggests, provide numbers that summarize and characterize the distribution. We describe the mean and the variance, two well-known summary statistics. Then we discuss two ways to compare a pair of random variables: first, how to say that two random variables are independent; and second, how to compute an inner product between them.

6.4.1 Means and Covariances

Mean and (co)variance are often useful to describe properties of probability distributions (expected values and spread). We will see in Section 6.6 that there is a useful family of distributions (called the exponential family), where the statistics of the random variable capture all possible information.

The concept of the expected value is central to machine learning, and the foundational concepts of probability itself can be derived from the expected value (Whittle, 2000).

Definition 6.3 (Expected Value)**.** The *expected value* of a function $g : \mathbb{R} \to \mathbb{R}$ of a univariate continuous random variable $X \sim p(x)$ is given by

expected value

$$\mathbb{E}_X[g(x)] = \int_{\mathcal{X}} g(x)p(x)\mathrm{d}x\,. \tag{6.28}$$

Correspondingly, the expected value of a function g of a discrete random variable $X \sim p(x)$ is given by

$$\mathbb{E}_X[g(x)] = \sum_{x\in\mathcal{X}} g(x)p(x)\,, \tag{6.29}$$

where $\mathcal{X}$ is the set of possible outcomes (the target space) of the random variable X.

In this section, we consider discrete random variables to have numerical outcomes. This can be seen by observing that the function g takes real numbers as inputs.

The expected value of a function of a random variable is sometimes referred to as the law of the unconscious statistician (Casella and Berger, 2002, Section 2.2).

Remark. We consider multivariate random variables X as a finite vector of univariate random variables $[X_1, \ldots, X_D]^\top$. For multivariate random variables, we define the expected value element wise

$$\mathbb{E}_X[g(\boldsymbol{x})] = \begin{bmatrix} \mathbb{E}_{X_1}[g(x_1)] \\ \vdots \\ \mathbb{E}_{X_D}[g(x_D)] \end{bmatrix} \in \mathbb{R}^D\,, \tag{6.30}$$

where the subscript $\mathbb{E}_{X_d}$ indicates that we are taking the expected value with respect to the dth element of the vector $\boldsymbol{x}$. ♢

Definition 6.3 defines the meaning of the notation $\mathbb{E}_X$ as the operator indicating that we should take the integral with respect to the probability density (for continuous distributions) or the sum over all states (for discrete distributions). The definition of the mean (Definition 6.4), is a special case of the expected value, obtained by choosing g to be the identity function.

Definition 6.4 (Mean)**.** The *mean* of a random variable X with states

mean

$\boldsymbol{x} \in \mathbb{R}^D$ is an average and is defined as

$$\mathbb{E}_X[\boldsymbol{x}] = \begin{bmatrix} \mathbb{E}_{X_1}[x_1] \\ \vdots \\ \mathbb{E}_{X_D}[x_D] \end{bmatrix} \in \mathbb{R}^D \,, \tag{6.31}$$

where

$$\mathbb{E}_{X_d}[x_d] := \begin{cases} \int_{\mathcal{X}} x_d p(x_d) \mathrm{d}x_d & \text{if } X \text{ is a continuous random variable} \\ \sum_{x_i \in \mathcal{X}} x_i p(x_d = x_i) & \text{if } X \text{ is a discrete random variable} \end{cases} \tag{6.32}$$

for $d = 1, \ldots, D$, where the subscript d indicates the corresponding dimension of $\boldsymbol{x}$. The integral and sum are over the states $\mathcal{X}$ of the target space of the random variable X.

median

In one dimension, there are two other intuitive notions of "average", which are the *median* and the *mode*. The *median* is the "middle" value if we sort the values, i.e., 50% of the values are greater than the median and 50% are smaller than the median. This idea can be generalized to continuous values by considering the value where the cdf (Definition 6.2) is 0.5. For distributions, which are asymmetric or have long tails, the median provides an estimate of a typical value that is closer to human intuition than the mean value. Furthermore, the median is more robust to outliers than the mean. The generalization of the median to higher dimensions is non-trivial as there is no obvious way to "sort" in more than one dimension (Hallin et al., 2010; Kong and Mizera, 2012).

mode

The *mode* is the most frequently occurring value. For a discrete random variable, the mode is defined as the value of x having the highest frequency of occurrence. For a continuous random variable, the mode is defined as a peak in the density $p(\boldsymbol{x})$. A particular density $p(\boldsymbol{x})$ may have more than one mode, and furthermore there may be a very large number of modes in high-dimensional distributions. Therefore, finding all the modes of a distribution can be computationally challenging.

Example 6.4

Consider the two-dimensional distribution illustrated in Figure 6.4:

$$p(x) = 0.4\mathcal{N}\left(\boldsymbol{x} \,\middle|\, \begin{bmatrix} 10 \\ 2 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\right) + 0.6\mathcal{N}\left(\boldsymbol{x} \,\middle|\, \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 8.4 & 2.0 \\ 2.0 & 1.7 \end{bmatrix}\right). \tag{6.33}$$

We will define the Gaussian distribution $\mathcal{N}(\mu, \sigma^2)$ in Section 6.5. Also shown is its corresponding marginal distribution in each dimension. Observe that the distribution is bimodal (has two modes), but one of the