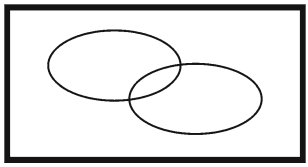
Chapter 2
Probability Theory



2.1 Introduction

Probability theory originated in games of chance and has a long and interesting history; it has developed into a mathematical language for quantifying uncertainty.

Consider a certain *experiment*, such as throwing a die; this experiment can have different results, we call each result an *outcome* or element. In the die example, the possible outcomes or elements are the following: {1, 2, 3, 4, 5, 6}. The set of all possible outcomes of an experiment is called the sample space, Ω. An *event* is a set of elements or subset of Ω. Continuing with the die example, one event could be that the die shows an even number, that is, {2, 4, 6}.

Before we mathematically define probability, it is worth discussing the meaning or interpretation of probability. Several definitions or interpretations of probability have been proposed, starting from the *classical* definition by Laplace, and including the *limiting frequency*, the *subjective*, the *logical*, and the *propensity* interpretations [1]:

Classical: probability has to do with equiprobable events; if a certain experiment has N possible outcomes, the probability of each outcome is $1/N$.

Logical: probability is a measure of rational belief; that is, according to the available evidence, a rational person will have a certain belief regarding an event, which will define its probability.

Subjective: probability is a measure of the personal degree of belief in a certain event; this could be measured in terms of a betting factor—the probability of a certain event for an individual is related to how much that person is willing to bet on that event.

Frequency: probability is a measure of the number of occurrences of an event given a certain experiment, when the number of repetitions of the experiment tends to infinity.

Propensity: probability is a measure of the number of occurrences of an event under repeatable conditions, even if the experiment only occurs once.

L.E. Sucar, *Probabilistic Graphical Models*,
Advances in Computer Vision and Pattern Recognition,
DOI 10.1007/978-1-4471-6699-3_2

These interpretations can be grouped into what are the two main approaches in probability and statistics:

- Objective (classical, frequency, propensity): probabilities exist in the *real* world and can be measured.
- Epistemological (logical, subjective): probabilities have to do with human knowledge, they are measures of belief.

Both approaches follow the same mathematical axioms defined below; however, there are differences in the manner in which probability is applied, in particular in statistical inference. These differences gave way to the main two schools for statistics: the frequentist and the Bayesian schools. In the field of artificial intelligence, in particular in expert systems, the preferred approach tends to be the epistemological or subjective one; however, the objective approach is also used [4].

We will consider the logical or normative approach and define probabilities in terms of the degree of plausibility of a certain proposition given the available evidence [2]. Based on Cox's work, Jaynes establishes some basic desiderata that this degree of plausibility must follow [2]:

- Representation by real numbers.
- Qualitative correspondence with common sense.
- Consistency.

Based on these intuitive principles, we can derive the three axioms of probability:

1. $P(A)$ is a continuous monotonic function in $[0, 1]$.
2. $P(A, B \mid C) = P(A \mid C)P(B \mid A, C)$ (product rule).
3. $P(A \mid B) + P(\neg A \mid B) = 1$ (sum rule).

Where A, B, C are propositions (binary variables) and $P(A)$ is the probability of proposition A. $P(A \mid C)$ is the probability of A given that C is known, which is called *conditional probability*. $P(A, B \mid C)$ is the probability of A AND B given C (logical conjunction) and $P(\neg A \mid C)$ is the probability of NOT A (logical negation) given C. These rules are equivalent to the most commonly used Kolmogorov axioms. From these axioms, all conventional probability theory can be derived.

2.2 Basic Rules

The probability of the disjunction (logical sum) of two propositions is given by the *sum rule*: $P(A + B \mid C) = P(A \mid C) + P(B \mid C) - P(A, B \mid C)$; if propositions A and B are mutually exclusive given C, we can simplify it to: $P(A + B \mid C) = P(A \mid C) + P(B \mid C)$. This can be generalized for N mutually exclusive propositions to:

$$P(A_1 + A_2 + \cdots A_N \mid C) = P(A_1 \mid C) + P(A_2 \mid C) + \cdots + P(A_N \mid C) \quad (2.1)$$

In the case that there are N mutually exclusive and exhaustive hypotheses, $H_1, H_2, \ldots, H_N$, and if the evidence B does not favor any of them, then according to the principle of indifference: $P(H_i \mid B) = 1/N$.

According to the logical interpretation there are no *absolute* probabilities, all are conditional on some background information.[1] $P(H \mid B)$ conditioned only on the background B is called a *prior* probability; once we incorporate some additional information D we call it a *posterior* probability, $P(H \mid D, B)$. From the product rule we obtain:

$$P(D, H \mid B) = P(D \mid H, B)P(H \mid B) = P(H \mid D, B)P(D \mid B) \tag{2.2}$$

From which we obtain:

$$P(H \mid D, B) = \frac{P(H \mid B)P(D \mid H, B)}{P(D \mid B)} \tag{2.3}$$

This last equation is known as the *Bayes rule* and the term $P(D \mid H, B)$ as the *likelihood*, $L(H)$.

In some cases the probability of H is not influenced by the knowledge of D, so it is said that H and D are *independent* given some background B, therefore, $P(H, D \mid B) = P(H \mid B)$. In the case in which A and B are independent, the product rule can be simplified to: $P(A, B \mid C) = P(A \mid C)P(B \mid C)$, and this can be generalized to N mutually independent propositions:

$$P(A_1, A_2, \ldots, A_N \mid B) = P(A_1 \mid B)P(A_2 \mid B) \cdots P(A_N \mid B) \tag{2.4}$$

If two propositions are independent given only the background information they are *marginally* independent; however, if they are independent given some additional evidence, E, then they are *conditionally* independent: $P(H, D \mid B, E) = P(H \mid B, E)$. For example, consider that A represents the proposition *watering the garden*, B the *weather forecast* and C *raining*. Initially, watering the garden is not independent of the weather forecast; however, once we observe rain, they become independent. That is (omitting the background term), $P(A, B \mid C) = P(A \mid C)$.

Probabilistic graphical models are based on these conditions of marginal and conditional independence.

The probability of a conjunction of N propositions, that is, $P(A_1, A_2, \ldots, A_N \mid B)$, is usually called the *joint* probability. If we generalize the product rule to N propositions we obtain what is known as the *chain* rule:

$$P(A_1, A_2, \ldots, A_N \mid B) = P(A_1 \mid A_2, A_3, \ldots, A_N, B)P(A_2 \mid A_3, A_4, \ldots, A_N, B) \cdots P(A_N \mid B) \tag{2.5}$$

[1] It is commonly written $P(H)$ without explicit mention of the conditioning information. In this case we assume that there is still some context under which probabilities are considered even if it is not written explicitly.

Thus the joint probability of N propositions can be obtained by this rule. Conditional independence relations between the propositions can be used to simplify this product; that is, for instance if A_1 and A_2 are independent given $A_3, \ldots, A_N, B$, then the first term in Eq. 2.5 can be simplified to $P(A_1 \mid A_3, \ldots, A_N, B)$.

Another important relation is the rule of *total probability*. Consider a partition, $B = \{B_1, B_2, \ldots, B_n\}$, on the sample space Ω, such that $\Omega = B_1 \cup B_2 \cup \cdots \cup B_n$ and $B_i \cap B_j = \emptyset$. That is, B is a set of mutually exclusive events that cover the entire sample space. Consider another event A; A is equal to the union of its intersections with each event $A = (B_1 \cap A) \cup (B_2 \cap A) \cup \cdots \cup (B_n \cap A)$. Then, based on the axioms of probability and the definition of conditional probability we can derive the rule of total probability:

$$P(A) = \sum_i P(A \mid B_i) P(B_i) \tag{2.6}$$

Given the total probability rule, we can rewrite Bayes rule as (omitting the background term):

$$P(B \mid A) = \frac{P(B)P(A \mid B)}{\sum_i P(A \mid B_i)P(B_i)} \tag{2.7}$$

This last expression is commonly known as Bayes Theorem.

2.3 Random Variables

If we consider a finite set of exhaustive and mutually exclusive propositions,[2] then a discrete variable X can represent this set of propositions, such that each value x_i of X corresponds to one proposition. If we assign a numerical value to each proposition x_i, then X is a *discrete random variable*. For example, the outcome of the toss of a die is a discrete random variable with six possible values $1, 2, \ldots, 6$. The probabilities for all possible values of X, $P(X)$ is the probability distribution of X. Considering the die example, for a fair die the probability distribution will be:

x	1	2	3	4	5	6
$P(x)$	1/6	1/6	1/6	1/6	1/6	1/6

This is an example of a *uniform* probability distribution. There are several probability distributions that have been defined. Another common distribution is the *binomial* distribution. Assume we have an urn with N colored balls, red and black, of which M are red, so the fraction of red balls is $\pi = M/N$. We draw a ball at

[2] This means that one and only one of the propositions has a value of TRUE.

random, record its color, and return it to the urn, mixing the balls again (so that, in principle, each draw is independent of the previous one). The probability of getting r red balls in n draws is:

$$P(r \mid n, \pi) = \binom{n}{r} \pi^r (1-\pi)^{n-r}, \tag{2.8}$$

where $\binom{n}{r} = \frac{n!}{r!(n-r)!}$.

This is an example of a binomial distribution which is applied when there are n independent trials, each with two possible outcomes (success or failure), and the probability of success is constant over all trials. There are many other distributions; we refer the interested reader to the additional reading section at the end of the chapter.

There are two important quantities that in general help to characterize a probability distribution. The expected value or *expectation* of a discrete random variable is the average of the possible values, weighted according to their probabilities:

$$E(X \mid B) = \sum_{i=1}^{N} P(x_i \mid B) x_i \tag{2.9}$$

The *variance* is defined as the expected value of the square of the variable minus its expectation:

$$Var(X \mid B) = \sum_{i=1}^{N} P(x_i \mid B)(x_i - E(X))^2 \tag{2.10}$$

Intuitively, the variance gives a measure of how *wide* or *narrow* the probabilities are distributed for a certain random variable. The square root of the variance is known as the standard deviation, which is usually more intuitive as its units are the same as those of the variable.

So far we have considered discrete variables, however, the rules of probability can be extended to continuous variables. If we have a continuous variable X, we can divide it into a set of mutually exclusive and exhaustive intervals, such that $P = (a < X \leq b)$ is a proposition, thus the rules derived so far apply to it. A *continuous random variable* can be defined in terms of a *probability density function*, $f(X \mid B)$, such that:

$$P(a < X \leq b \mid B) = \int_a^b f(X \mid B) dx \tag{2.11}$$

The probability density function must satisfy $\int_{-\infty}^{\infty} f(X \mid B) dx = 1$.

An example of a continuous probability distribution is the *Normal* or Gaussian distribution. This distribution plays an important role in many applications of probability and statistics, as many phenomena in nature have an approximately normal distribution; it is also prevalent in probabilistic graphical models due to its mathematical properties.

A normal distribution is denoted as $N(\mu, \sigma^2)$, where μ is the *mean* (center) and σ is the *standard deviation* (spread); and it is defined as:

$$f(X \mid B) = \frac{1}{\sigma\sqrt{2\pi}} exp\{-\frac{1}{2\sigma^2}(x-\mu)^2\} \tag{2.12}$$

The density function of a Gaussian distribution is depicted in Fig. 2.1.

Another important continuous distribution is the exponential distribution; for example, the time it takes for a certain piece of equipment to fail is usually modeled by an exponential distribution. The exponential distribution is denoted as $Exp(\beta)$; it has a single parameter $\beta > 0$, and it is defined as:

$$f(X \mid B) = \frac{1}{\beta} e^{-x/\beta}, x > 0 \tag{2.13}$$

An example of an exponential density function is shown in Fig. 2.2.

It is common to represent probability distributions, in particular for continuous variables, using the cumulative distribution function, F. The cumulative distribution function of a random variable, X, is the probability that $X \leq x$. For a continuous variable, it is defined in terms of the density function as:

$$F(x) = \int_{-\infty}^{x} f(X) \tag{2.14}$$

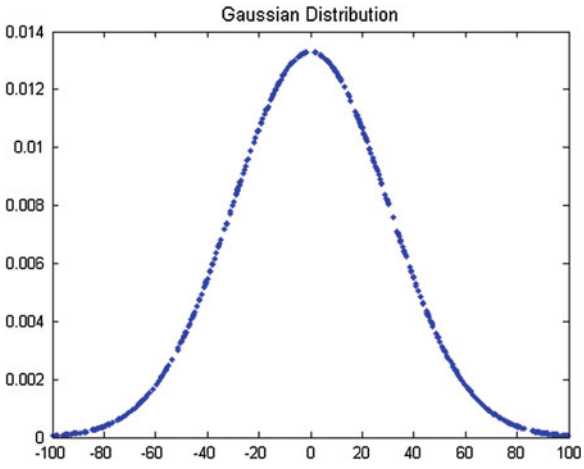


Fig. 2.1 Probability density function of the Gaussian distribution

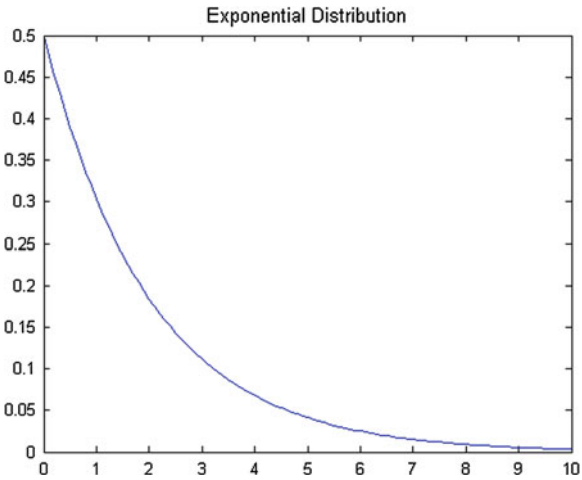


Fig. 2.2 Probability density function of the exponential distribution

The following are some properties of cumulative distribution functions:

- In the interval [0, 1]: $0 \leq F(X) \leq 1$
- Nondecreasing: $F(X_1) < F(X_2)$ if $X_1 < X_2$
- Limits: $lim_{x \to -\infty} = 0$ and $lim_{x \to \infty} = 1$

In the case of discrete variables, the cumulative probability, $P(X \leq x)$ is defined as:

$$\mathbf{P}(x) = \sum_{x=-\infty}^{X=x} P(X) \tag{2.15}$$

It has similar properties as the cumulative distribution function.

2.3.1 Two-Dimensional Random Variables

The concept of a random variable can be extended to two or more dimensions. Given two random variables, X and Y, their joint probability distribution is defined as $P(x, y) = P(X = x \wedge Y = y)$. For example, X might represent the number of products completed in one day in product line one, and Y the number of products completed in one day in product line two, thus $P(x, y)$ corresponds to the probability of producing x products in line one and y products in line two. $P(X, Y)$ must follow the axioms of probability, in particular: $0 \leq P(x, y) \leq 1$ and $\sum_x \sum_y P(X, Y) = 1$.

The distribution for two-dimensional discrete random variables (known as the *bivariate* distribution) can be represented in tabular form. For instance, consider the example of the two product lines, and assume that line one (X) may produce 1, 2, or 3 products per day, and line two (Y), 1 or 2 products. Then a possible joint distribution, $P(X, Y)$ is shown in Table 2.1.

Table 2.1 An example of a two-dimensional discrete probability distribution

	X = 1	X = 2	X = 3
Y = 1	0.1	0.3	0.3
Y = 2	0.2	0.1	0

Given the joint probability distribution, $P(X, Y)$, we can obtain the distribution for each individual random variable, what is known as the marginal probability:

$$P(x) = \sum_{y} P(X, Y);\ P(y) = \sum_{x} P(X, Y) \tag{2.16}$$

For instance, if we consider the joint distribution of Table 2.1, we can obtain the marginal probabilities for X and Y. For example, $P(X = 2) = 0.3 + 0.1 = 0.4$ and $P(Y = 1) = 0.1 + 0.3 + 0.3 = 0.7$.

We can also calculate the conditional probabilities of X given Y and vice versa:

$$P(x \mid y) = P(x, y)/P(y);\ P(y \mid x) = P(x, y)/P(x) \tag{2.17}$$

Following the example in Table 2.1:

$$P(X = 3 \mid Y = 1) = P(X = 3, Y = 1)/P(Y = 1) = 0.3/0.7 = 0.4286$$

The concept of independence can be applied to two-dimensional random variables. Two random variables, X, Y are independent if their joint probability distribution is equal to the product of their marginal distributions (for all values of X and Y):

$$P(X, Y) = P(X)P(X) \rightarrow \textit{Independent}\ (X, Y) \tag{2.18}$$

Another useful measure is called *correlation*—it is a measure of the degree of linear relation between two random variables, X, Y and is defined as:

$$\rho(X, Y) = E\{[X - E(X)][Y - E(Y)]\}/(\sigma_x \sigma_y) \tag{2.19}$$

where $E(X)$ is the expected value of X and σ_x its standard deviation. The correlation is in the interval $[-1, 1]$; a positive correlation indicates that as X increases, Y tends to increase; and a negative correlation that as X increases, Y tends to decrease.

Note that a correlation of zero does not necessarily imply independence, as the correlation only measures a linear relationship. So it could be that X and Y have a zero correlation but are related through a higher order function, and thus are not independent.

2.4 Information Theory

Information theory originated in the area of communications, although it is relevant for many different fields. In the case of probabilistic graphical models, it is mainly applied in learning. In this section we will cover the basic concepts of information theory.

Assume that we are communicating the occurrence of a certain event. Intuitively, we can think that the amount of *information* from communicating an event is inverse to the probability of the event. For example, consider that a message is sent informing about one of the following events:

1. It is raining in New York.
2. There was an earthquake in New York.
3. A meteorite fell over New York City.

The probability of the first event is higher than the second, and that of the second is higher than the third. Thus, the message for event 1 has the lowest amount of information and the message for event 3 gives the highest amount of information.

Let us now see how we can formalize the concept of information. Assume we have a source of information that can send q possible messages, $m_1, m_2, \ldots, m_q$; where each message corresponds to an event with probabilities $P_1, P_2, \ldots, P_q$. We want to find a function $I(m)$ based on the probability of m. The function must satisfy the following properties:

- The information ranges from zero to infinity: $I(m) \geq 0$.
- The information increases as the probability decreases: $I(m_i) > I(m_j)$ if $P(m_i) < P(m_j)$.
- The information tends to infinity as the probability tends to zero: $I(m) \to \infty$ if $P(m) \to 0$.
- The information of two messages is equal to the sum of that of the individual messages if these are independent: $I(m_i + m_j) = I(m_i) + I(m_j)$ if m_i independent of m_j.

A function that satisfies the previous properties is the logarithm of the inverse of the probability, that is,

$$I(m_k) = log(1/P(m_k)) \tag{2.20}$$

It is common to use base two logarithms, so the information is measured in "bits":

$$I(m_k) = log_2(1/P(m_k)) \tag{2.21}$$

For example, if we assume that the probability of the message m_r "raining in New York" is $P(m_r) = 0.25$, the corresponding information is $I(m_r) = log_2(1/0.25)$=2.

Once we have defined information for a particular message, another important concept is the average information for the q messages; that is, the expected value of the information also known as *entropy*. Given the definition of expected value, the average information of q message or entropy is defined as:

$$H(m) = E(I(m)) = \sum_{i=1}^{i=q} P(m_i) log_2(1/P(m_i)) \tag{2.22}$$

This can be interpreted as that on average H bits of information will be sent.

An interesting question is: When will H have its maximum and minimum values? Consider a binary source such that there are only two messages, m_1 and m_2; with $P(m_1) = p_1$ and $P(m_2) = p_2$. Given that there are only two possible messages, $p_2 = 1 - p_1$, so H only depends on one parameter, p_1 (or just p). Figure 2.3 shows a graph of H with respect to p. Observe that H is at its maximum when $p = 0.5$ and at its minimum (zero) when $p = 0$ and $p = 1$. In general, the entropy is at its maximum when there is a uniform probability distribution for the events; it is at its minimum when there is one element that has a probability of one and the rest have a probability of zero.

If we consider the conditional probabilities, we can extend the concept of entropy to *conditional entropy*:

$$H(X \mid y) = \sum_{i=1}^{i=q} P(X_i \mid y) log_2[1/P(X_i \mid y)] \tag{2.23}$$

Another extension of entropy is the *cross entropy*:

$$H(X, Y) = \sum_{X} \sum_{Y} P(X, Y) log_2[P(X, Y)/P(X)P(Y)] \tag{2.24}$$

The cross entropy provides a measure of the mutual information (dependency) between two random variables; it is zero when the two variables are independent.

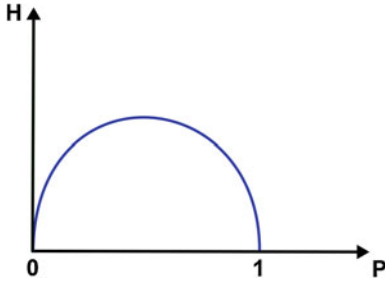


Fig. 2.3 Entropy versus probability for a binary source. The entropy is at its maximum when the probability is 0.5, and at its minimum when the probability is zero and one

2.5 Additional Reading

Donald Gillies [1] provides a comprehensive account of the different philosophical approaches to probability. An excellent book on probability theory from a logical perspective is [2]. Wasserman [5] gives a concise course on probability and statistics oriented for computer science and engineering students. There are several books on information theory; one that relates it to machine learning and inference is [3].

2.6 Exercises

1. If we throw two dice, what is the probability that the outcomes add to exactly seven? Seven or more?
2. If we assume that the height of a group of students follows a normal distribution with a mean of 1.7 m and a standard deviation of 0.1 m, how probable is it that there is a student of height above 1.9 m?
3. Demonstrate by mathematical induction the chain rule.
4. Assume that a person has one of two possible diseases, hepatitis (H) or typhoid (T). There are two symptoms associated to these diseases: headache (D) and fever (F), which could be TRUE or FALSE. Given the following probabilities: $P(T) = 0.5$, $P(D \mid T) = 0.7$, $P(D \mid \neg T) = 0.4$, $P(F \mid T) = 0.9$, $P(F \mid \neg T) = 0.5$. Describe the sampling space and complete the partial probability tables.
5. Given the data for the previous problem, and assuming that the symptoms are independent given the disease, obtain the probability that the person has hepatitis given that she does not have a headache and does have a fever.
6. Given the two-dimensional probability distribution in the table below, obtain: (a) $P(X_1)$, (b) $P(Y_2)$, and (c) $P(X_1 \mid Y_1)$.

	Y_1	Y_2	Y_3
X_1	0.1	0.2	0.1
X_2	0.3	0.1	0.2

7. In the previous problem, are X and Y independent?
8. In a certain place, the statistics show that in a year the weather behaves in the following way. From 365 days, 200 are sunny, 60 cloudy, 40 rainy, 20 snowy, 20 with thundershowers, 10 with hail, 10 windy, and 5 with drizzle. If on each day a message is sent about the weather, what is the information of the message for each type of weather?

9. Considering the information for each type of weather in the previous problem, what is the average information (entropy) of the message?
10. ***Investigate the different philosophical interpretations of probability, and discuss the advantages and limitations of each one of them. Which one do you consider the most appropriate? Why?

References

1. Gillies, D.: Philosophical Theories of Probability. Routledge, London (2000)
2. Jaynes, E.T.: Probability Theory: The Logic of Science. Cambridge University Press, Cambridge (2003)
3. MacKay, D.J.: Information Theory, Inference and Learning Algorithms. Cambridge University Press, Cambridge (2004)
4. Sucar, L.E., Gillies, D.F., Gillies, D.A.: Objective Probabilities in Expert Systems. Artif. Intell. **61**, 187–208 (1993)
5. Wasserman, L.: All of Statistics: A Concise Course in Statistical Inference. Springer, New York (2004)

http://www.springer.com/978-1-4471-6698-6

Probabilistic Graphical Models
Principles and Applications
Sucar, L.E.
2015, XXIV, 253 p. 117 illus., 4 illus. in color., Hardcover
ISBN: 978-1-4471-6698-6