5 Basic Probability Theory

Probability theory addresses the following fundamental question: *how do we reason?* Reasoning is central to many areas of human endeavor, including philosophy (what is the best way to make decisions?), cognitive science (how does the mind work?), artificial intelligence (how do we build reasoning machines?), and science (how do we test and develop theories based on experimental data?). In nearly all real-world situations, our data and knowledge about the world is incomplete, indirect, and noisy. As a consequence, uncertainty is a fundamental part of our decision-making process. **Bayesian** reasoning provides a formal and consistent way to reason in the presence of uncertainty. Bayesian probability theory is distinguished by defining probabilities as **degrees-of-belief**. This is in contrast to **Frequentist statistics**, where the probability of an event is defined as its frequency in the limit of an infinite number of repeated trials.

5.1 Classical Logic

Perhaps the most famous attempt to describe a formal system of reasoning is classical logic, originally developed by Aristotle. In classical logic, we have some statements that may be true or false, and we have a set of rules which allow us to determine the truth or falsity of new statements. For example, suppose we introduce two statements, named $\mathbf{A}$ and $\mathbf{B}$:

$\mathbf{A} \equiv$ "My car was stolen"
$\mathbf{B} \equiv$ "My car is not in the parking spot where I remember leaving it"

Moreover, let us assert the rule "$\mathbf{A}$ implies $\mathbf{B}$", which we will write as $\mathbf{A} \rightarrow \mathbf{B}$. Then, if $\mathbf{A}$ is known to be true, we may deduce logically that $\mathbf{B}$ must also be true (if my car is stolen then it won't be in the parking spot where I left it). Alternatively, if I find my car where I left it ("$\mathbf{B}$ is false," written $\bar{\mathbf{B}}$), then I may infer that it was not stolen ($\bar{\mathbf{A}}$) by the contrapositive $\bar{\mathbf{B}} \rightarrow \bar{\mathbf{A}}$.

Classical logic provides a model of how humans might reason, and a model of how we might build an "intelligent" computer. Unfortunately, classical logic has a significant shortcoming: it assumes that all knowledge is absolute. Logic requires that we know some facts about the world with absolute certainty, and then, we may deduce only those facts which must follow with absolute certainty.

In the real world, there are almost no facts that we know with absolute certainty. Most of what we know about the world we acquire indirectly, through our five senses, or from dialogue with other people. One can therefore conclude that most of what we know about the world is **uncertain**. (Finding something that we know with certainty has occupied generations of philosophers.)

For example, suppose I discover that my car is not where I remember leaving it ($\mathbf{B}$). Does this mean that it was stolen? No, there are many other explanations — maybe I have forgotten where I left it or maybe it was towed. However, the knowledge of $\mathbf{B}$ makes $\mathbf{A}$ more *plausible*. Even though we do not know it to be stolen, it becomes more likely a scenario than before. The actual degree of plausibility depends on other contextual information — did I park it in a safe neighborhood?, did I park it in a handicapped zone?, etc.

Predicting the weather is another task that requires reasoning with uncertain information. While we can make some predictions with great confidence (e.g. we can reliably predict that it will not snow in June, north of the equator), we are often faced with much more difficult questions (will it rain today?) which we must infer from unreliable sources of information (e.g., the weather report, clouds in the sky, yesterday's weather, etc.). In the end, we usually cannot determine for certain whether it will rain, but we do get a degree of certainty upon which to base decisions, like whether or not to carry an umbrella.

Another important example of uncertain reasoning occurs whenever you meet someone new — at this time, you immediately make hundreds of inferences (mostly unconscious) about who this person is and what their emotions and goals are. You make these decisions based on the person's appearance, the way they are dressed, their facial expressions, their actions, the context in which you meet, and what you have learned from previous experience with other people. Of course, you have no conclusive basis for forming opinions (e.g., the panhandler you meet on the street may be a method actor preparing for a role). However, we need to be able to make judgements about other people based on incomplete information; otherwise, normal interpersonal interaction would be impossible (e.g., how do you really *know* that everyone isn't out to get you?).

What we need is a way of discussing not just true or false statements, but statements that have varying levels of certainty. In addition, we would like to be able to use our beliefs to reason about the world and interpret it. As we gain new information, our beliefs should change to reflect our greater knowledge. For example, for any two propositions $\mathbf{A}$ and $\mathbf{B}$ (that may be true or false), if $\mathbf{A} \rightarrow \mathbf{B}$, then strong belief in $\mathbf{A}$ should increase our belief in $\mathbf{B}$. Moreover, strong belief in $\mathbf{B}$ may sometimes increase our belief in $\mathbf{A}$ as well.

5.2 Basic Definitions and Rules

The rules of probability theory provide a system for reasoning with uncertainty.There are a number of justifications for the use of probability theory to represent logic (such as Cox's Axioms) that show, for certain particular definitions of common-sense reasoning, that probability theory is the only system that is consistent with common-sense reasoning. We will not cover these here (see, for example, Wikipedia for discussion of the Cox Axioms).

The basic rules of probability theory are as follows.

- The probability of a statement $\mathbf{A}$ — denoted $P(\mathbf{A})$ — is a real number between 0 and 1, inclusive. $P(\mathbf{A}) = 1$ indicates absolute certainty that $\mathbf{A}$ is true, $P(\mathbf{A}) = 0$ indicates absolute certainty that $\mathbf{A}$ is false, and values between 0 and 1 correspond to varying degrees of certainty.

- The **joint probability** of two statements $\mathbf{A}$ and $\mathbf{B}$ — denoted $P(\mathbf{A}, \mathbf{B})$ — is the probability that both statements are true. (i.e., the probability that the statement "$\mathbf{A} \wedge \mathbf{B}$" is true). (Clearly, $P(\mathbf{A}, \mathbf{B}) = P(\mathbf{B}, \mathbf{A})$.)

- The **conditional probability** of $\mathbf{A}$ given $\mathbf{B}$ — denoted $P(\mathbf{A}|\mathbf{B})$ — is the probability that we would assign to $\mathbf{A}$ being true, **if** we knew $\mathbf{B}$ to be true. The conditional probability is defined as $P(\mathbf{A}|\mathbf{B}) = P(\mathbf{A}, \mathbf{B})/P(\mathbf{B})$.

- **The Product Rule:**
$$P(\mathbf{A}, \mathbf{B}) = P(\mathbf{A}|\mathbf{B})P(\mathbf{B}) \tag{5.1}$$

In words, the probability that $\mathbf{A}$ and $\mathbf{B}$ are both true is given by the probability that $\mathbf{B}$ is true, multiplied by the probability we would assign to $\mathbf{A}$ if we knew $\mathbf{B}$ to be true. Similarly, $P(\mathbf{A}, \mathbf{B}) = P(\mathbf{B}|\mathbf{A})P(\mathbf{A})$. This rule follows directly from the definition of conditional probability.

- **The Sum Rule:**

  $$P(\mathbf{A}) + P(\bar{\mathbf{A}}) = 1 \tag{5.2}$$

  In words, the probability of a statement being true and the probability that it is false must sum to 1. In other words, our certainty that $\mathbf{A}$ is true is in inverse proportion to our certainty that it is not true. A consequence: given a set of mutually-exclusive statements $\mathbf{A}_i$, exactly one of which must be true, we have

  $$\sum_i P(\mathbf{A}_i) = 1 \tag{5.3}$$

- All of the above rules can be made conditional on additional information. For example, given an additional statement $\mathbf{C}$, we can write the Sum Rule as:

  $$\sum_i P(\mathbf{A}_i|\mathbf{C}) = 1 \tag{5.4}$$

  and the Product Rule as

  $$P(\mathbf{A}, \mathbf{B}|\mathbf{C}) = P(\mathbf{A}|\mathbf{B}, \mathbf{C})P(\mathbf{B}|\mathbf{C}) \tag{5.5}$$

From these rules, we further derive many more expressions to relate probabilities. For example, one important operation is called **marginalization:**

$$P(\mathbf{B}) = \sum_i P(\mathbf{A}_i, \mathbf{B}) \tag{5.6}$$

if $\mathbf{A}_i$ are mutually-exclusive statements, of which exactly one must be true. In the simplest case — where the statement $\mathbf{A}$ may be true or false — we can derive:

$$P(\mathbf{B}) = P(\mathbf{A}, \mathbf{B}) + P(\bar{\mathbf{A}}, \mathbf{B}) \tag{5.7}$$

The derivation of this formula is straightforward, using the basic rules of probability theory:

$$\begin{aligned}
P(\mathbf{A}) + P(\bar{\mathbf{A}}) &= 1, && \text{Sum rule} && (5.8)\\
P(\mathbf{A}|\mathbf{B}) + P(\bar{\mathbf{A}}|\mathbf{B}) &= 1, && \text{Conditioning} && (5.9)\\
P(\mathbf{A}|\mathbf{B})P(\mathbf{B}) + P(\bar{\mathbf{A}}|\mathbf{B})P(\mathbf{B}) &= P(\mathbf{B}), && \text{Algebra} && (5.10)\\
P(\mathbf{A}, \mathbf{B}) + P(\bar{\mathbf{A}}, \mathbf{B}) &= P(\mathbf{B}), && \text{Product rule} && (5.11)
\end{aligned}$$

Marginalization gives us a useful way to compute the probability of a statement $\mathbf{B}$ that is intertwined with many other uncertain statements.

Another useful concept is the notion of **independence**. Two statements are independent if and only if $P(\mathbf{A}, \mathbf{B}) = P(\mathbf{A})P(\mathbf{B})$. If $\mathbf{A}$ and $\mathbf{B}$ are independent, then it follows that $P(\mathbf{A}|\mathbf{B}) = P(\mathbf{A})$ (by combining the Product Rule with the defintion of independence). Intuitively, this means that, whether or not $\mathbf{B}$ is true tells you nothing about whether $\mathbf{A}$ is true.

In the rest of these notes, we will always use probabilities as statements about variables. For example, suppose we have a variable x that indicates whether there are one, two, or three people in a room (i.e., the only possibilities are $x = 1$, $x = 2$, $x = 3$). Then, by the sum rule, we can derive $P(x = 1) + P(x = 2) + P(x = 3) = 1$. Probabilities can also describe the range of a real variable. For example, $P(y < 5)$ is the probability that the variable y is less than 5. (We'll discuss continuous random variables and probability densities in more detail in the next chapter.)

To summarize:

The basic rules of probability theory:
• $P(\mathbf{A}) \in [0...1]$
• **Product rule:** $P(\mathbf{A}, \mathbf{B}) = P(\mathbf{A}|\mathbf{B})P(\mathbf{B})$
• **Sum rule:** $P(\mathbf{A}) + P(\bar{\mathbf{A}}) = 1$
• Two statements $\mathbf{A}$ and $\mathbf{B}$ are **independent** iff: $P(\mathbf{A}, \mathbf{B}) = P(\mathbf{A})P(\mathbf{B})$
• **Marginalizing:** $P(\mathbf{B}) = \sum_i P(\mathbf{A}_i, \mathbf{B})$
• Any basic rule can be made conditional on additional information.
For example, it follows from the product rule that $P(\mathbf{A}, \mathbf{B}|\mathbf{C}) = P(\mathbf{A}|\mathbf{B}, \mathbf{C})P(\mathbf{B}|\mathbf{C})$

Once we have these rules — and a suitable model — we can derive *any* probability that we want. With some experience, you should be able to derive any desired probability (e.g., $P(\mathbf{A}|\mathbf{C})$) given a basic model.

5.3 Discrete Random Variables

In ML, and other areas of engineering and science, it is convenient to describe systems in terms of random variables. For example, to describe the weather, we might define a discrete variable $\mathbf{w}$ that can take on two values sunny or rainy, and then try to determine $P(\mathbf{w} = \mathsf{sunny})$, i.e., the probability that it will be sunny today. **Discrete (Categorical) distributions** describe these types of probabilities.

A Bernoulli distribution is a special case of a Categorical distribution when there are only two outcomes. The canonical example is the random coint oss. Let $\mathbf{c}$ be a variable that specifies the result of the flip: $\mathbf{c} = \mathsf{heads}$ if the coin lands on its head, and $\mathbf{c} = \mathsf{tails}$ otherwise. In this chapter and the rest of these notes, we will use probabilities specifically to refer to values of variables, e.g., $P(\mathbf{c} = \mathsf{heads})$ is the probability that the coin lands heads.

What is the probability that the coin lands heads? This probability should be some real number θ, $0 \leq \theta \leq 1$. For most coins, we would say $\theta = .5$. What does this number mean? The number θ is a representation of our belief about the possible values of $\mathbf{c}$. Some examples:

$\theta = 0$	we are absolutely certain the coin will land tails
$\theta = 1/3$	we believe that tails is twice as likely as heads
$\theta = 1/2$	we believe heads and tails are equally likely
$\theta = 1$	we are absolutely certain the coin will land heads

Formally, we denote the probability of the coin coming up heads as $P(\mathbf{c} = \textsf{heads})$, so $P(\mathbf{c} = \textsf{heads}) = \theta$. In general, we denote the probability of a specific event event as $P(\textsf{event})$. By the Sum Rule, we know $P(\mathbf{c} = \textsf{heads}) + P(\mathbf{c} = \textsf{tails}) = 1$, and thus $P(\mathbf{c} = \textsf{tails}) = 1 - \theta$.

Once we flip the coin and observe the result, then we can be pretty sure that we know the value of $\mathbf{c}$. There is no practical need to model the uncertainty in this measurement. However, suppose we do not observe the coin flip, but instead hear about it from a friend (who may be forgetful or untrustworthy). Let $\mathbf{f}$ be a variable indicating how the friend claims the coin landed. That is, $\mathbf{f} = \textsf{heads}$ means the friend says that the coin came up heads. Suppose the friend says the coin landed heads. Do we believe him? And if so, with how much certainty? As we shall see, probabilistic reasoning obtains quantitative values that qualitatively matches our common sense reasonably well.

Suppose we know something about our friend's behaviour. We can represent our beliefs with probabilities. For example, $P(\mathbf{f} = \textsf{heads} \,|\, \mathbf{c} = \textsf{heads})$ represents our belief that the friend says "heads" when the the coin landed heads. Because the friend can only say one thing, we can apply the Sum Rule to get:

$$P(\mathbf{f} = \textsf{heads} \,|\, \mathbf{c} = \textsf{heads}) + P(\mathbf{f} = \textsf{tails} \,|\, \mathbf{c} = \textsf{heads}) \;=\; 1 \tag{5.12}$$

$$P(\mathbf{f} = \textsf{heads} \,|\, \mathbf{c} = \textsf{tails}) + P(\mathbf{f} = \textsf{tails} \,|\, \mathbf{c} = \textsf{tails}) \;=\; 1 \tag{5.13}$$

If our friend always tells the truth, then we know $P(\mathbf{f} = \textsf{heads} \,|\, \mathbf{c} = \textsf{heads}) = 1$ and $P(\mathbf{f} = \textsf{tails} \,|\, \mathbf{c} = \textsf{heads}) = 0$. But if our friend *usually* lies, then we might have $P(\mathbf{f} = \textsf{heads} \,|\, \mathbf{c} = \textsf{heads}) = .3$.

5.4 Binomial and Multinomial Distributions

A binomial distribution is the distribution over the number of positive outcomes for a binary (yes/no) experiment, where on each trial the probability of a positive outcome is $p \in [0, 1]$. For example, for n tosses of a coin for which the probability of heads on a single trial is p, the distribution over the number of heads we might observe is a binomial distribution. The binomial distribution over the number of positive outcomes, denoted K, given n trials, each having a positive outcome with probability p is given by

$$P(K = k) \;=\; \binom{n}{k}\, p^k\, (1-p)^{n-k} \tag{5.14}$$

for $k = 0, 1, \ldots, n$, where

$$\binom{n}{k} = \frac{n!}{k!\,(n-k)!}\ . \tag{5.15}$$

A multinomial distribution is a natural extension of the binomial distribution to an experiment with k mutually exclusive outcomes, having probabilities p_j, for $j = 1, \ldots, k$. Of course, to be valid probabilities $\sum p_j = 1$. For example, rolling a die can yield one of six values, each with probability 1/6 (assuming the die is fair). Given n trials, the multinomial distribution specifies the distribution over the number of each of the possible outcomes. Given n trials, k possible outcomes with probabilities p_j, the distribution over the event that outcome j occurs x_j times (and of course $\sum x_j = n$), is the multinomial distribution given by

$$P(X_1 = x_1, X_2 = x_2, \ldots, X_k = x_k) = \frac{n!}{x_1!\, x_2!\, \ldots\, x_k!}\, p_1^{x_1}\, p_2^{x_2}\, \ldots\, p_k^{x_k} \tag{5.16}$$

5.5 Mathematical Expectation

Suppose each outcome r_i has an associated real value $x_i \in \mathbb{R}$. Then, the *expected value* of x is:

$$\mathbb{E}[x] \;=\; \sum_i P(r_i)\, x_i\ . \tag{5.17}$$

The expected value of $f(x)$ is given by

$$\mathbb{E}[f(x)] \;=\; \sum_i P(r_i)\, f(x_i)\ . \tag{5.18}$$