A

Probability Theory

Probability theory provides a formal framework for the discussion of chance or uncertainty. This appendix reviews some key concepts of the theory and establishes notation. However, it glosses over some details (e.g., pertaining to measure theory). Therefore, the interested reader is encouraged to consult a textbook on the topic for a more comprehensive picture.

A.1 Probabilistic models

A probabilistic model is defined as a tuple $(\Omega, \mathcal{F}, P)$, where:

sample space

event space

probability density function

- Ω is the *sample space*, also called the *event space*;
- $\mathcal{F}$ is a σ-algebra over Ω; that is, $\mathcal{F} \subseteq 2^{\Omega}$ and is closed under intersection and countable union; and
- $P : \mathcal{F} \mapsto [0, 1]$ is the *probability density function* (PDF).

Intuitively, the sample space is a set of things that can happen in the world according to our model. For example, in a model of a six-sided die, we might have $\Omega = \{1, 2, 3, 4, 5, 6\}$. The σ-field $\mathcal{F}$ is a collection of measurable events. $\mathcal{F}$ is required because some outcomes in Ω may not be measurable; thus, we must define our probability density function P over $\mathcal{F}$ rather than over Ω. However, in many cases, such as the six-sided die example, all outcomes *are* measurable. In those cases we can equate $\mathcal{F}$ with 2^{Ω} and view the probability space as the pair (Ω, P) and P as $P : 2^{\Omega} \mapsto [0, 1]$. We assume this in the following.

A.2 Axioms of probability theory

The probability density function P must satisfy the following axioms.

1. For any $A \subseteq \Omega$, $P(\emptyset) = 0 \leq P(A) \leq P(\Omega) = 1$.
2. For any pair of disjoint sets $A, A' \subset \Omega$, $P(A \cup A') = P(A) + P(A')$.

That is, all probabilities must be bounded by 0 and 1; 0 is the probability of the empty set and 1 the probability of the whole sample space. Second, when sets of outcomes from the sample space are nonoverlapping, the probability of achieving an outcome from either of the sets is the sum of the probabilities of achieving an outcome from each of the sets. We can infer from these rules that if two sets $A, A' \subseteq \Omega$ are not disjoint, $P(A \cup A') = P(A) + P(A') - P(A \cap A')$.

A.3 Marginal probabilities

We are often concerned with sample spaces Ω that are defined as the Cartesian product of a set of random variables $X_1, \ldots, X_n$ with domains $\mathcal{X}_1, \ldots, \mathcal{X}_n$ respectively. Thus, in this setting, $\Omega = \prod_{i=1}^{n} \mathcal{X}_i$. Our density function P is thus defined over full assignments of values to our variables, such as $P(X_1 = x_1, \ldots, X_n = x_n)$. However, sometimes we want to ask about *marginal probabilities*: the probability that a single variable X_i takes some value $x_i \in \mathcal{X}_i$. We define

marginal probability

$$P(X_i = x_i) = \sum_{x_1 \in \mathcal{X}_1} \cdots \sum_{x_{i-1} \in \mathcal{X}_{i-1}} \sum_{x_{i+1} \in \mathcal{X}_{i+1}} \cdots \sum_{x_n \in \mathcal{X}_n} P(X_1 = x_1, \ldots, X_n = x_n).$$

From this definition and from the axioms given earlier we can also infer that, for example,

$$P(X_i = x_i) = \sum_{x_j \in \mathcal{X}_j} P(X_i = x_i \text{ and } X_j = x_j).$$

A.4 Conditional probabilities

We say that two random variables X_i and X_j are *independent* when $P(X_i = x_i$ and $X_j = x_j) = P(X_i = x_i) \cdot P(X_j = x_j)$ for all values $x_i \in \mathcal{X}_i$, $x_j \in \mathcal{X}_j$.

Often, random variables are not independent. When this is the case, it can be important to know the probability that X_i will take some value x_i given that $X_j = x_j$ has already been observed. We define this probability as

$$P(X_i = x_i | X_j = x_j) = \frac{P(X_i = x_i \text{ and } X_j = x_j)}{P(X_j = x_j)}.$$

conditional probability

joint probability

Bayes' rule

We call $P(X_i = x_i | X_j = x_j)$ a *conditional probability*; $P(X_i = x_i$ and $X_j = x_j)$ is called a *joint probability* and $P(X_j = x_j)$ a marginal probability, already discussed previously.

Finally, *Bayes' rule* is an important identity that allows us to reverse conditional probabilities. Specifically,

$$P(X_i = x_i | X_j = x_j) = \frac{P(X_j = x_j | X_i = x_i) P(X_i = x_i)}{P(X_j = x_j)}.$$