1.2: Basic Concepts in Probability Theory

Fundamental definitions

The natural mathematical setting is set theory. *Sets* are generalized collections of *objects*. The basics: $\omega \in A$ is a binary relation which says that the object ω is an *element* of the set A. Another binary relation is *set inclusion*. If all members of A are in B, we write $A \subseteq B$. The *union* of sets A and B is denoted $A \cup B$ and the *intersection* of A and B is denoted $A \cap B$. The *Cartesian product* of A and B, denoted $A \times B$, is the set of all ordered elements (a, b) where $a \in A$ and $b \in B$.

Some details: If ω is not in A, we write $\omega \notin A$. Sets may also be objects, so we may speak of sets of sets, but typically the sets which will concern us are simple discrete collections of numbers, such as the possible rolls of a die {1,2,3,4,5,6}, or the real numbers $\mathbb{R}$, or Cartesian products such as $\mathbb{R}^N$. If $A \subseteq B$ but $A \neq B$, we say that A is a *proper subset* of B and write $A \subset B$. Another binary operation is the *set difference* $A \backslash B$, which contains all ω such that $\omega \in A$ and $\omega \notin B$.

In probability theory, each object ω is identified as an *event*. We denote by Ω the set of all events, and $\emptyset$ denotes the set of no events. There are three basic axioms of probability:

- To each set A is associated a non-negative real number $P(A)$, which is called the probability of A.
- $P(\Omega) = 1$.
- If $\{A_i\}$ is a collection of disjoint sets, if $A_i \cap A_j = \emptyset$ for all $i \neq j$, then

$$P\Big(\bigcup_i A_i\Big) = \sum_i P(A_i)\,. \tag{1.2.1}$$

From these axioms follow a number of conclusions. Among them, let $\neg A = \Omega \backslash A$ be the *complement* of A, the set of all events *not* in A. Then since $A \cup \neg A = \Omega$, we have $P(\neg A) = 1 - P(A)$. Taking $A = \Omega$, we conclude $P(\emptyset) = 0$.

The meaning of $P(A)$ is that if events ω are chosen from Ω *at random*, then the relative frequency for $\omega \in A$ approaches $P(A)$ as the number of trials tends to infinity. But what do we mean by 'at random'? One meaning we can impart to the notion of randomness is that a process is random if its outcomes can be accurately modeled using the axioms of probability. This entails the identification of a *probability space* Ω as well as a *probability measure* P. For example, in the microcanonical ensemble of classical statistical physics, the space Ω is the collection of phase space points $\varphi = \{q_1, \ldots, q_n, p_1, \ldots, p_n\}$ and the probability measure is $d\mu = \Sigma^{-1}(E) \prod_{i=1}^n dq_i\, dp_i\, \delta\big(E - H(q,p)\big)$, so that for $A \in \Omega$ the probability of A is $P(A) = \int d\mu\, \chi_A(\varphi)$, where $\chi_A(\varphi) = 1$ if $\varphi \in A$ and $\chi_A(\varphi) = 0$ if $\varphi \notin A$ is the *characteristic function* of A. The quantity $\Sigma(E)$ is determined by normalization: $\int d\mu = 1$.

Bayesian Statistics

We now introduce two additional probabilities. The *joint probability* for sets A and B together is written $P(A \cap B)$. That is, $P(A \cap B) = Prob[\omega \in A \; and \; \omega \in B]$. For example, A might denote the set of all politicians, B the set of all American citizens, and C the set of all living humans with an IQ greater than 60. Then $A \cap B$ would be the set of all politicians who are also American citizens, *Exercise: estimate* $P(A \cap B \cap C)$.

The *conditional probability* of B *given* A is written $P(B|A)$. We can compute the joint probability $P(A \cap B) = P(B \cap A)$ in two ways:

$$P(A \cap B) = P(A|B) \cdot P(B) = P(B|A) \cdot P(A)\,. \tag{1.2.2}$$

Thus,

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}\,, \tag{1.2.3}$$

a result known as *Bayes' theorem*. Now suppose the 'event space' is partitioned as $\{A_i\}$. Then

$$P(B) = \sum_i P(B|A_i)\,P(A_i)\,. \tag{1.2.4}$$

We then have

$$P(A_i|B) = \frac{P(B|A_i)\,P(A_i)}{\sum_j P(B|A_j)\,P(A_j)}\ , \tag{1.2.5}$$

a result sometimes known as the *extended form of Bayes' theorem*. When the event space is a 'binary partition' $\{A, \neg A\}$, we have

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B|A)\,P(A) + P(B|\neg A)\,P(\neg A)}\ . \tag{1.2.6}$$

Note that $P(A|B) + P(\neg A|B) = 1$ (which follows from $\neg\neg A = A$).

As an example, consider the following problem in epidemiology. Suppose there is a rare but highly contagious disease A which occurs in 0.01% of the general population. Suppose further that there is a simple test for the disease which is accurate 99.99% of the time. That is, out of every 10,000 tests, the correct answer is returned 9,999 times, and the incorrect answer is returned only once. Now let us administer the test to a large group of people from the general population. Those who test positive are quarantined. Question: what is the probability that someone chosen at random from the quarantine group actually has the disease? We use Bayes' theorem with the binary partition $\{A, \neg A\}$. Let B denote the event that an individual tests positive. Anyone from the quarantine group has tested positive. Given this datum, we want to know the probability that that person has the disease. That is, we want $P(A|B)$. Applying Equation [Bayesbinary] with

$$P(A) = 0.0001 \quad , \quad P(\neg A) = 0.9999 \quad , \quad P(B|A) = 0.9999 \quad , \quad P(B|\neg A) = 0.0001\ , \tag{1.2.7}$$

we find $P(A|B) = \frac{1}{2}$. That is, there is only a 50% chance that someone who tested positive actually has the disease, despite the test being 99.99% accurate! The reason is that, given the rarity of the disease in the general population, the number of false positives is statistically equal to the number of true positives.

In the above example, we had $P(B|A) + P(B|\neg A) = 1$, but this is not generally the case. What is true instead is $P(B|A) + P(\neg B|A) = 1$. Epidemiologists define the *sensitivity* of a binary classification test as the fraction of actual positives which are correctly identified, and the *specificity* as the fraction of actual negatives that are correctly identified. Thus, $\mathsf{se} = P(B|A)$ is the sensitivity and $\mathsf{sp} = P(\neg B|\neg A)$ is the specificity. We then have $P(B|\neg A) = 1 - P(\neg B|\neg A)$. Therefore,

$$P(B|A) + P(B|\neg A) = 1 + P(B|A) - P(\neg B|\neg A) = 1 + \mathsf{se} - \mathsf{sp}\ . \tag{1.2.8}$$

In our previous example, $\mathsf{se} = \mathsf{sp} = 0.9999$, in which case the RHS above gives 1. In general, if $P(A) \equiv f$ is the fraction of the population which is afflicted, then

$$P(\mathsf{infected}\,|\,\mathsf{positive}) = \frac{f\cdot\mathsf{se}}{f\cdot\mathsf{se} + (1-f)\cdot(1-\mathsf{sp})}\ . \tag{1.2.9}$$

For continuous distributions, we speak of a probability *density*. We then have

$$P(y) = \int\! dx\ P(y|x)\,P(x) \tag{1.2.10}$$

and

$$P(x|y) = \frac{P(y|x)\,P(x)}{\int\! dx'\ P(y|x')\,P(x')}\ . \tag{1.2.11}$$

The range of integration may depend on the specific application.

The quantities $P(A_i)$ are called the *prior distribution*. Clearly in order to compute $P(B)$ or $P(A_i|B)$ we must know the priors, and this is usually the weakest link in the Bayesian chain of reasoning. If our prior distribution is not accurate, Bayes' theorem will generate incorrect results. One approach to approximating prior probabilities $P(A_i)$ is to derive them from a *maximum entropy construction*.

Random variables and their averages

Consider an abstract probability space $\mathcal{X}$ whose elements (events) are labeled by x. The average of any function $f(x)$ is denoted as $\mathbb{E}f$ or $\langle f\rangle$, and is defined for discrete sets as

$$\mathbb{E}f = \langle f\rangle = \sum_{x\in\mathcal{X}} f(x)\,P(x)\ , \tag{1.2.12}$$

where $P(x)$ is the probability of x. For continuous sets, we have

$$\mathbb{E}f = \langle f \rangle = \int\limits_{\mathcal{X}} dx\ f(x)\,P(x)\ . \tag{1.2.13}$$

Typically for continuous sets we have $\mathcal{X} = \mathbb{R}$ or $\mathcal{X} = \mathbb{R}_{\geq 0}$. Gardiner and other authors introduce an extra symbol, X, to denote a *random variable*, with $X(x) = x$ being its value. This is formally useful but notationally confusing, so we'll avoid it here and speak loosely of x as a random variable.

When there are two random variables $x \in \mathcal{X}$ and $y \in \mathcal{Y}$, we have $\Omega = \mathcal{X} \times \mathcal{Y}$ is the product space, and

$$\mathbb{E}f(x,y) = \langle f(x,y) \rangle = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} f(x,y)\,P(x,y)\ , \tag{1.2.14}$$

with the obvious generalization to continuous sets. This generalizes to higher rank products, $x_i \in \mathcal{X}_i$ with $i \in \{1, \ldots, N\}$. The *covariance* of x_i and x_j is defined as

$$C_{ij} \equiv \big\langle \big(x_i - \langle x_i \rangle\big)\big(x_j - \langle x_j \rangle\big) \big\rangle = \langle x_i x_j \rangle - \langle x_i \rangle \langle x_j \rangle\ . \tag{1.2.15}$$

If $f(x)$ is a convex function then one has

$$\mathbb{E}f(x) \geq f(\mathbb{E}x)\ . \tag{1.2.16}$$

For continuous functions, $f(x)$ is convex if $f''(x) \geq 0$ everywhere[4]. If $f(x)$ is convex on some interval $[a,b]$ then for $x_{1,2} \in [a,b]$ we must have

$$f\big(\lambda x_1 + (1-\lambda)x_2\big) \leq \lambda f(x_1) + (1-\lambda) f(x_2)\ , \tag{1.2.17}$$

where $\lambda \in [0,1]$. This is easily generalized to

$$f\Big(\sum_n p_n x_n\Big) \leq \sum_n p_n f(x_n)\ , \tag{1.2.18}$$

where $p_n = P(x_n)$, a result known as *Jensen's theorem*.