Handling Probabilities

When considering a particular algorithm (or experiment), we represent some performance measure as a variable with some set of possible outcomes. When the outcomes are stochastic in nature (meaning we cannot determine *a priori* what the outcome will be), then we refer to these variables as *random variables*. For a particular random variable X, we will represent the probability that X takes on some value x as $P(X = x) = p$, where $0 \leq p \leq 1$. This defines a probability distribution over the values of the random variable X. For example, suppose X is a *Boolean* random variable (meaning that there are only two possible values for X). Suppose X can take on the values TRUE and FALSE. Then we can define the Boolean probability distribution over X to be $P(X = \text{TRUE}) = p$ and $P(X = \text{FALSE}) = (1 - p)$.

There are many different philosophical approaches to handling probabilities. Most of the probabilities handled in this class will tend to follow the "frequentist" approach in that probabilities will be derived from explicit experiments. However, an alternative approach that offers tremendous power in algorithm design is the "Bayesian" approach. In Bayesian probability, proposal distributions and belief estimates are used.

Both approaches make a distinction between "unconditional" and "conditional" probabilities. An *unconditional probability* is a probability that is determined based on no other information. In Bayesian probability, these probabilities are sometimes called "prior" or a priori probabilities. A *conditional probability* is a probability whose value is determined based on the existence of other information. In Bayesian probability, conditional probabilities tend to appear in one of two roles—as a likelihood estimate or as a "posterior" or a posteriori probability. Suppose we have two events a and b. The unconditional probability of a is denoted as $P(a)$. The conditional probability of a given that we know b, however, is denoted as $P(a|b)$. Conditional probabilities are defined as $P(a|b) = P(a, b)/P(b)$.

Notice the numerator to the definition of the conditional probability. The notation $P(a, b)$ denotes a *joint probability* over the two events a and b. One common task when manipulating probabilities is extracting the distribution of a subset of variables over a single variable. The process of extracting the distribution is called either *marginalizing* or *conditioning* depending on the form used. (In general, we will use the term marginalize even when mathematically we are conditioning. This is because, as we will see, the two approaches are mathematically equivalent.)

- Marginalizing: $P(\mathbf{Y}) = \sum_{\mathbf{z}} P(\mathbf{Y}, \mathbf{z})$
- Conditioning: $P(\mathbf{Y}) = \sum_{\mathbf{z}} P(\mathbf{Y}|\mathbf{z})P(\mathbf{z})$

Consider marginalizing. We calculate the probability distribution of event $\mathbf{Y}$ from the joint distribution $P(\mathbf{Y}, \mathbf{z})$ by summing the probabilities over all possible values of $\mathbf{z}$. Recall the definition of conditional probability. Using this definition, we can rewrite $P(\mathbf{Y}, \mathbf{z})$ as $P(\mathbf{Y}|\mathbf{z})P(\mathbf{z})$. Thus we see that conditioning and marginalizing are equivalent.

Another common use of probabilities is in estimating new probabilities given other probabilities. One common rule used is Bayes' Rule (which forms the foundation of the Bayesian approach). Bayes' Rule can be derived directly from the definition of conditional probability. First, recall that $P(a, b) = P(a|b)(P(b)$. Observing that $P(a, b) = P(b, a)$, we can also show that $P(a, b) = P(b|a)P(a)$. Setting the right hand side of these two equations to be equal to each other, we derive Bayes' Rule:

$$\begin{aligned} P(a|b)P(b) &= P(b|a)P(a) \\ P(a|b) &= \frac{P(b|a)P(a)}{P(b)}. \end{aligned}$$

As an alternative definition, we can "reverse" the process of conditioning by noting that $P(b) = \sum_i P(b|a_i)P(a_i)$. So we can now rewrite Bayes' Rule as

$$P(a|b) = \frac{P(b|a)P(a)}{\sum_i P(b|a_i)P(a_i)}.$$

At times, we will also need to be concerned with the relationship between random variables. In particular, we will make use of the concept of independence. We say that two random variables X and Y and independent

if any of the following hold:

$$\begin{aligned} P(X|Y) &= P(X) \\ P(Y|X) &= P(Y) \\ P(X,Y) &= P(X)P(Y). \end{aligned}$$

We can apply the same type of definitions with conditional probabilities. Specifically, given three random variables, X, Y, and Z, then we can say X and Y are conditionally independent given Z if any of the following hold:

$$\begin{aligned} P(X|Y,Z) &= P(X|Z) \\ P(Y|X,Z) &= P(Y|Z) \\ P(X,Y|Z) &= P(X|Z)P(Y|Z). \end{aligned}$$

Given a probability distribution, other things we might want to know about that distribution are several summary statistics. Specifically, the *mean* or *expected value* of a distribution can be determined as $E[X] = \sum_x xP(X = x)$. Expectation has a nice property, called the *linearity of expectation*, in which we find that $E[X + Y] = E[X] + E[Y]$. It is interesting to note that this holds even when X and Y are not independent.

We do not want to end with expected value since the expected value says nothing about the shape of the distribution. There are several other statistics that can describe the shape (called "moments" of the distribution) such as variance, skew, and kurtosis. We will concern ourselves only with variance. (Actually, the mean is also a "moment" of the distribution—it is the first moment). Variance provides a measure of how "spread out" the distribution is. Variance is defined as an expected value over the amount of variation in the distribution as follows:

$$\begin{aligned} Var[X] &= E[(X - E[X])^2] \\ &= E[X^2 - 2XE[X] + E^2[X]] \\ &= E[X^2] - 2E[XE[X]] + E^2[X] \\ &= E[X^2] - 2E[X]E[X] + E^2[X] \\ &= E[X^2] - E^2[X]. \end{aligned}$$

Generally, it is not possible to determine the true mean and variance of a distribution based on a set of experiments that have been run. However, the sample mean and sample variance can be determined as approximations of the underlying mean and variance. These are defined as follows:

- Sample Mean: $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
- Sample Variance: $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$

Note that the square root of the variance is called the "standard deviation."