Note 10: Probability Basics

Math 198: Math for Machine Learning

1 Probability Measures

Probability theory allows us to assign likelihoods to events for processes with contain some element of random chance. Take, for instance, a dice roll. There are six possible outcomes; this set of possible outcomes is known as the *sample space* and is denoted Ω. In general, Ω can be an infinite set, for which the probability of any specific outcome is 0. Therefore we additionally consider *events*, which are subsets of Ω. The set of all events is denoted $\mathcal{F}$. We can then define a *probability measure* which associates events in $\mathcal{F}$ with probabilities between 0 and 1, that is, $\mathbb{P} : \mathcal{F} \to [0, 1]$. This measure must satisfy two basic properties: $\mathbb{P}(\Omega) = 1$, and for any countable collection of disjoint sets $\{A_i\} \subseteq \mathcal{F}$, $\mathbb{P}\big(\bigcup_i A_i\big) = \sum_i \mathbb{P}(A_i)$ (*countable additivity*). Together Ω, $\mathcal{F}$, and $\mathbb{P}$ constitute a *probability space*.

We will now consider a handful of basic and useful results which apply to all probability spaces. For some event A we can define the *complement* of A, $A^c = \Omega \backslash A$. Then $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$, as $A + A^c = \Omega$. Furthermore, if we have two events A, B such that $B \subseteq A$, then $\mathbb{P}(B) \leq \mathbb{P}(A)$:

$$\mathbb{P}(A) = \mathbb{P}(B \cup (A \backslash B)) = \mathbb{P}(B) + \mathbb{P}(A \backslash B) \geq \mathbb{P}(B)$$

If A and B are taken to be generic events, with possible overlap, then $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$:

$$\begin{aligned}
\mathbb{P}(A \cup B) &= \mathbb{P}((A \cap B) \cup (A \backslash B) \cup (B \backslash A)) \\
&= \mathbb{P}(A \cap B) + \mathbb{P}(A \backslash B) + \mathbb{P}(B \backslash A) \\
&= \mathbb{P}(A \cap B) + \mathbb{P}(A) - \mathbb{P}(A \cap B) + \mathbb{P}(B) - \mathbb{P}(A \cap B) \\
&= \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)
\end{aligned}$$

Generally, for any countable set of events $\{A_i\} \subseteq \mathcal{F}$, $\mathbb{P}\big(\bigcup_i A_i\big) \leq \sum_i \mathbb{P}(A_i)$; this is known as the *union bound*.

We denote the conditional probability of an event A given an event B occurred as $\mathbb{P}(A|B)$, and it is defined as

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$$

for $\mathbb{P}(B) > 0$. From this definition we can derive the equality $\mathbb{P}(A \cap B) = \mathbb{P}(A|B)\mathbb{P}(B) = \mathbb{P}(B|A)\mathbb{P}(A)$, and from this we arrive at *Bayes' rule*:

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(B|A)\mathbb{P}(A)}{\mathbb{P}(B)}$$

In this context we refer to $\mathbb{P}(A)$ as the *prior probability*, $\mathbb{P}(A|B)$ as the *posterior*, and $\mathbb{P}(B|A)$ as the *likelihood*.

2 Random Variables/Vectors

So far we have considered outcomes and events, but we will work more often with *random variables*, which are any uncertain quantities with an associated probability distribution over the values they can assume. For example, consider two dice rolls. Both dies rolling a 6 is an outcome, the sum of the dies equaling 7 is an event, and the sum of the dies is a random variable. Formally, a random variable on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ is a function $X : \Omega \to \mathbb{R}$.

We can define the probability that a random variable X takes on some value x by making reference to the outcomes in Ω:

$$\mathbb{P}(X = x) = \mathbb{P}(\{\omega \in \Omega : X(\omega) = x\})$$

For a random variable X we define the *cumulative distribution function*, which gives the probability that X is at most some value:

$$F(x) = \mathbb{P}(X \leq x)$$

The CDF can also be used to give us the probability that a variable lies within some range:

$$\mathbb{P}(a < X \leq b) = F(b) - F(a)$$

If X has a countable range and assumes each value in this range with positive probability, we describe it as a *discrete random variable*. We can then define a *probability density function* $p : X(\Omega) \to [0, 1]$ which satisfies

$$\sum_{x \in X(\Omega)} p(x) = 1$$

by just setting $p(x) = \mathbb{P}(X = x)$.

For a *continuous random variable* with an uncountable range, each value in the range is assumed with probability zero. In this case we define a PDF $p : \mathbb{R} \to [0, \infty)$ such that

$$F(x) = \int_{-\infty}^{x} p(z)\mathrm{d}z$$

with the requirement that

$$\int_{-\infty}^{\infty} p(x)\mathrm{d}x = 1$$

The values of p are not exactly probabilities (notably they can take on any positive value), but can be understood to represent the relative likelihood that the value of X falls in the neighborhood of x. In particular, for small $\epsilon > 0$,

$$\mathbb{P}(x - \epsilon \leq X \leq x + \epsilon) = \int_{x-\epsilon}^{x+\epsilon} p(z)\mathrm{d}z \approx 2\epsilon p(x)$$

3 Expected Value

We can define the average value of a random variable X – we refer to this as the *expected value* or *mean* $\mathbb{E}[X]$. For discrete X we have

$$\mathbb{E}[X] = \sum_{x \in X(\Omega} xp(x)$$

and for continuous X

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} xp(x)\mathrm{d}x$$

The mean of a distribution can be interpreted as the center of mass of its PDF.

Perhaps the nicest thing about taking expected values is that they are linear:

$$\mathbb{E}\big[\sum_{i=1}^{n} \alpha_i X_i + \beta\big] = \sum_{i=1}^{n} \alpha_i \mathbb{E}[X_i] + \beta$$

which holds even if the X_i are not independent.

4 Variance

Just as we can use expectation as a measure of the center of a distribution, the *variance* gives us a measure of the spread about the center. The variance $\text{Var}(X)$ of a random variable X is the average squared deviation of the value of X from its expected value:

$$\text{Var}(X) = \mathbb{E}\big[(X - \mathbb{E}[X])^2\big]$$

It is straightforward to show that $\text{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2$:

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

Variance is not linear, but $\text{Var}(\alpha X + \beta) = \alpha^2 \text{Var}(X)$. However, given n independent random variables $X_1 \ldots X_n$, then $\text{Var}(X_1 + \ldots + X_n) = \text{Var}(X_1) + \ldots + \text{Var}(X_n)$. This is indeed true of any uncorrelated X_n; we will expand on this identity in the note on correlation.

Finally, as the variance is not in the same units as the random variable itself (due to the squaring in the definition), we additionally define the *standard deviation* $\sigma(X) = \sqrt{\text{Var}(X)}$. This value is of the same scale as X itself, and can be used to normalize X:

$$\bar{X} = \frac{X - \mathbb{E}[X]}{\sigma(X)}$$

Applications: Chebyshev's Inequality

Much like the union bound, we are able to define fundamental inequalities which constrain probabilities for arbitrary probability spaces. These properties will come in handy when working with probabilities, regardless of the structure of the problem.

First, let's consider a weaker result, Markov's inequality: if X is a nonnegative random variable and $a > 0$,

$$\mathbb{P}(X \geq a) \leq \frac{\mathbb{E}[X]}{a}$$

We show the proof for continuous X; the discrete case is similar:

$$\begin{aligned}
\mathbb{E}[X] &= \int_{-\infty}^{\infty} xp(x)\mathrm{d}x \\
&= \int_{0}^{\infty} xp(x)\mathrm{d}x \\
&= \int_{0}^{a} xp(x)\mathrm{d}x + \int_{a}^{\infty} xp(x)\mathrm{d}x \\
&\geq \int_{a}^{\infty} xp(x)\mathrm{d}x \\
&\geq \int_{a}^{\infty} ap(x)\mathrm{d}x \\
&= a\int_{a}^{\infty} p(x)\mathrm{d}x \\
&= a\mathbb{P}(X \geq a)
\end{aligned}$$

Chebyshev's inequality is more general – it applies to all random variables X, and gives a more concrete notion of how the variance σ^2 of X measures its spread around $\mathbb{E}[X]$. For any real number $k > 0$, Chebyshev's inequality gives us that

$$\mathbb{P}(|X - \mathbb{E}[X]| \geq k\sigma) \leq \frac{1}{k^2}$$

Chebyshev's inequality follows from Markov's inequality; let $Y = (X - \mathbb{E}[X])^2$ and $a = (k\sigma)^2$, then

$$\mathbb{P}(|X - \mathbb{E}[X]| \geq k\sigma) = \mathbb{P}((X - \mathbb{E}[X])^2 \geq k^2\sigma^2) \leq \frac{\mathbb{E}[X - \mathbb{E}[X]]^2}{k^2\sigma^2} = \frac{\sigma^2}{k^2\sigma^2} = \frac{1}{k^2}$$

Chebyshev's inequality gives us a decent bound on these probabilities for any possible probability space, but often for specific distributions we can improve on these bounds.

Applications: Law of Large Numbers

How do we know that probability theory is valid at all? Theory is useless if it does not agree with practice. Fortunately, the Law of Large Numbers ensures that, after a sufficiently large number of trials, practice is guaranteed to match theory. More formally, if we continue to take independent, identically distributed samples, the sample average will converge to the true average of the distribution.

The statement of the (weak[1]) law of large numbers is as follows. Let $X_1, X_2, \ldots$ be a series of independent, identically distributed random variables with mean μ. The sample average of n such random variables is $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$. Then for any $\epsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}(|\bar{X}_n - \mu| < \epsilon) = 0$$

If the variance σ^2 is finite, we can prove the weak law of large numbers using Chebyshev's inequality. Since the X_i are independent, we have

$$\mathrm{Var}(\bar{X}_n) = \mathrm{Var}\big(\frac{1}{n}\sum_{i=1}^{n} X_i\big) = \frac{1}{n^2}\sum_{i=1}^{n} \mathrm{Var}(X_i) = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$$

We then apply Chebyshev's inequality to $\bar{X}_n$:

$$\mathbb{P}(|\bar{X}_n - \mu| \geq \epsilon) \leq \frac{\sigma^2}{n\epsilon^2}$$

Clearly, as n approaches infinity, this upper bound for this probability approaches 0.

[1]The strong version differs in the strength of the convergence; both laws essentially say the same thing.