Probability review

We will go through a review of probability concepts over here, all of the review materials have been adapted from CS229 Probability Notes.

1. Elements of probability

In order to define a probability on a set we need a few basic elements,

Sample space Ω: The set of all the outcomes of a random experiment. Here, each outcome $\omega \in \Omega$ can be thought of as a complete description of the state of the real world at the end of the experiment.

Set of events (or event space) F: A set whose elements $A \in F$ (called events) are subsets of Ω (i.e., $A \subseteq \Omega$ is a collection of possible outcomes of an experiment)

Probability measure: A function $P : F \to \mathbb{R}$ that satisfies the following **properties**

> $P(A) \geq 0$, for all $A \in F$
>
> If A_1, A_2, ... are disjoint events $(i.e., A_i \cap A_j = \emptyset$ whenever $i \neq j)$, then $P(\bigcup_i A_i) = \sum_i P(A_i)$
>
> $P(\Omega) = 1$.

These three **properties** are called the **Axioms of Probability**.

Example: Consider the event of tossing a six-sided die. The sample space is $\Omega = \{1, 2, 3, 4, 5, 6\}$. We can define different event spaces on this sample space. For example, the simplest event space is the trivial event space $F = \{\emptyset, \Omega\}$. Another event space is the set of all subsets of Ω. For the first event space, the unique probability measure satisfying the requirements above is given by $P(\emptyset) = 0$, $P(\Omega) = 1$. For the second event

space, one valid probability measure is to assign the probability of each set in the event space to be $\frac{i}{6}$ where i is the number of elements of that set; for example, $P(\{1,2,3,4\}) = \frac{4}{6}$ and $P(\{1,2,3\}) = \frac{3}{6}$.

Properties:

If $A \subseteq B \Rightarrow P(A) \leq P(B)$.

$P(A \cap B) \leq min(P(A), P(B))$.

Union Bound $P(A \cup B) \leq P(A) + P(B)$.

$$P(\Omega - A) = 1 - P(A).$$

Law of Total Probability If $A_1, \ldots, A_k$ are a set of disjoint events such that $\bigcup_{i=1}^{k} A_i = \Omega$ then $\sum_{i=1}^{k} P(A_i) = 1$.

1.1 Conditional probability

Let B be an event with non-zero probability. The conditional probability of any event A given B is defined as

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}.$$

In other words, $P(A \mid B)$ is the probability measure of the event A after observing the occurrence of event B.

1.2 Chain Rule

Let $S_1, \cdots, S_k$ be events, $P(S_i) > 0$. Then the chain rule states that:

$$\begin{aligned} & P(S_1 \cap S_2 \cap \cdots \cap S_k) \\ = & P(S_1)P(S_2|S_1)P(S_3|S_2 \cap S_1) \cdots P(S_k|S_1 \cap S_2 \cap \cdots S_{k-1}) \end{aligned}$$

Note that for $k = 2$ events, this is just the definition of conditional probability:

$$P(S_1 \cap S_2) = P(S_1)P(S_2|S_1)$$

In general, the chain rule is derived by applying the definition of conditional independence multiple times, as in the following example:

$$\begin{aligned} & P(S_1 \cap S_2 \cap S_3 \cap S_4) \\ = & P(S_1 \cap S_2 \cap S_3) P(S_4 \mid S_1 \cap S_2 \cap S_3) \\ = & P(S_1 \cap S_2) P(S_3 \mid S_1 \cap S_2) P(S_4 \mid S_1 \cap S_2 \cap S_3) \\ = & P(S_1) P(S_2 \mid S_1) P(S_3 \mid S_1 \cap S_2) P(S_4 \mid S_1 \cap S_2 \cap S_3) \end{aligned}$$

1.3 Independence

Two events are called independent if and only if $P(A \cap B) = P(A)P(B)$ (or equivalently, $P(A \mid B) = P(A)$). Therefore, independence is equivalent to saying that observing B does not have any effect on the probability of A.

2. Random variables

Consider an experiment in which we flip 10 coins, and we want to know the number of coins that come up heads. Here, the elements of the sample space Ω are 10-length sequences of heads and tails. For example, we might have $\omega_0 = \langle H, H, T, H, T, H, H, T, T, T \rangle \in \Omega$. However, in practice, we usually do not care about the probability of obtaining any particular sequence of heads and tails. Instead we usually care about real-valued functions of outcomes, such as the number of heads that appear among our 10 tosses, or the length of the longest run of tails. These functions, under some technical conditions, are known as **random variables**.

More formally, a random variable X is a function $X : \Omega \to \mathbb{R}$. Typically, we will denote random variables using upper case letters $X(\omega)$ or more simply X (where the dependence on the random outcome ω is implied). We will denote the value that a random variable may take on using lower case letters x. Thus, $X = x$ means that we are assigning the value $x \in \mathbb{R}$ to the random variable X.

Example: In our experiment above, suppose that $X(\omega)$ is the number of heads which occur in the sequence of tosses ω. Given that only 10 coins are tossed, $X(\omega)$ can take only a finite number of values, so it is known as a discrete random variable. Here, the probability of the set associated with a random variable X taking on some specific value k is
$P(X = k) := P(\{\omega : X(\omega) = k\})$.

Example: Suppose that $X(\omega)$ is a random variable indicating the amount of time it takes for a radioactive particle to decay. In this case, $X(\omega)$ takes on a infinite number of possible values, so it is called a continuous random variable. We denote the probability that X takes on a value between two real constants a and b (where $a < b$) as
$P(a \leq X \leq b) := P(\{\omega : a \leq X(\omega) \leq b\})$.

2.1 Cumulative distribution functions

In order to specify the probability measures used when dealing with random variables, it is often convenient to specify alternative functions (CDFs, PDFs, and PMFs) from which the probability measure governing an experiment immediately follows. In this section and the next two sections, we describe each of these types of functions in turn. A cumulative distribution function (CDF) is a function $F_X : \mathbb{R} \to [0, 1]$ which specifies a probability measure as,

$$F_X(x) = P(X \leq x).$$

By using this function one can calculate the probability of any event.

Properties:

$0 \leq F_X(x) \leq 1$.

$lim_{x \to -\infty} F_X(x) = 0$.

$lim_{x \to +\infty} F_X(x) = 1$.

$x \leq y \Rightarrow F_X(x) \leq F_X(y)$.

2.2 Probability mass functions

When a random variable X takes on a finite set of possible values (i.e., X is a discrete random variable), a simpler way to represent the probability measure associated with a random variable is to directly specify the probability of each value that the random variable can assume. In particular, a probability mass function (PMF) is a function $p_X : \Omega \to \mathbb{R}$ such that $p_X(x) = P(X = x)$.

In the case of discrete random variable, we use the notation $Val(X)$ for the set of possible values that the random variable X may assume. For example, if $X(\omega)$ is a random variable indicating the number of heads out of ten tosses of coin, then $Val(X) = \{0, 1, 2, \ldots, 10\}$.

Properties:

$0 \leq p_X(x) \leq 1$.

$\sum_{x \in Val(X)} p_X(x) = 1$.

$\sum_{x \in A} p_X(x) = P(X \in A)$.

2.3 Probability density functions

For some continuous random variables, the cumulative distribution function $F_X(x)$ is differentiable everywhere. In these cases, we define the Probability Density Function or PDF as the derivative of the CDF, i.e.,

$$f_X(x) = \frac{dF_X(x)}{dx}.$$

Note here, that the PDF for a continuous random variable may not always exist (i.e., if $F_X(x)$ is not differentiable everywhere).

According to the **properties** of differentiation, for very small δx,

$P(x \leq X \leq x + \delta x) \approx f_X(x)\delta x$.

Both CDFs and PDFs (when they exist!) can be used for calculating the probabilities of different events. But it should be emphasized that the value of PDF at any given point x is not the probability of that event, i.e., $f_X(x) \neq P(X = x)$. For example, $f_X(x)$ can take on values larger than one (but the integral of $f_X(x)$ over any subset of $\mathbb{R}$ will be at most one).

Properties:

$f_X(x) \geq 0$.

$\int_{-\infty}^{\infty} f_X(x) = 1$.

$\int_{x \in A} f_X(x)dx = P(X \in A)$.

2.4 Expectation

Suppose that X is a discrete random variable with **PMF** $p_X(x)$ and $g : \mathbb{R} \to \mathbb{R}$ is an arbitrary function. In this case, $g(X)$ can be considered a random variable, and we define the expectation or expected value of $g(X)$ as:

$$E[g(X)] = \sum_{x \in Val(X)} g(x)p_X(x).$$

If X is a continuous random variable with PDF $f_X(x)$, then the expected value of g(X) is defined as,

$$E[g(X)] = \int_{-\infty}^{\infty} g(x)f_X(x)dx$$

.

Intuitively, the expectation of $g(X)$ can be thought of as a "weighted average" of the values that $g(x)$ can taken on for different values of x, where the weights are given by $p_X(x)$ or $f_X(x)$. As a special case of the above, note that the expectation, $E[X]$ of a random variable itself is found by letting $g(x) = x$; this is also known as the mean of the random variable X.

Properties:

$E[a] = a$ for any constant $a \in \mathbb{R}$.

$E[af(X)] = aE[f(X)]$ for any constant $a \in \mathbb{R}$.

(Linearity of Expectation)
$E[f(X) + g(X)] = E[f(X)] + E[g(X)]$.

For a discrete random variable X,
$E[\mathbf{1}\{X = k\}] = P(X = k)$.

2.5 Variance

The variance of a random variable X is a measure of how concentrated the distribution of a random variable X is around its mean. Formally, the variance of a random variable X is defined as
$Var[X] = E[(X - E[X])^2]$.

Using the **properties** in the previous section, we can derive an alternate expression for the variance:

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

where the second equality follows from linearity of expectations and the fact that $E[X]$ is actually a constant with respect to the outer expectation.

Properties:

$Var[a] = 0$ for any constant $a \in \mathbb{R}$.

$Var[af(X)] = a^2 Var[f(X)]$ for any constant $a \in \mathbb{R}$.

Example Calculate the mean and the variance of the uniform random variable X with PDF
$f_X(x) = 1, \forall x \in [0, 1], 0$ elsewhere.

$$E[X] = \int_{-\infty}^{\infty} x f_x(x) dx = \int_0^1 x dx = \frac{1}{2}$$

$$E[X^2] = \int_{-\infty}^{\infty} x^2 f_X(x) dx = \int_0^1 x^2 dx = \frac{1}{3}$$

$$Var[X] = E[X^2] - E[X]^2 = \frac{1}{3} - \frac{1}{4} = \frac{1}{12}$$

Example : Suppose that $g(x) = \mathbf{1}\{x \in A\}$ for some subset $A \subseteq \Omega$. What is $E[g(X)]$?

Discrete case:

$$E[g(X)] = \sum_{x \in Val(X)} \mathbf{1}\{x \in A\} P_X(x) dx = \sum_{x \in A} P_X(x) dx = P(X \in A)$$

Continuous case:

$$E[g(X)] = \int_{-\infty}^{\infty} \mathbf{1}\{x \in A\} f_X(x) dx = \int_{x \in A} f_X(x) dx = P(X \in A)$$

2.6 Some common random variables

Discrete random variables

• $X \sim$ **Bernoulli(p)** (where $0 \leq p \leq 1$): one if a coin with heads probability p comes up heads, zero otherwise.

$$p(x) = \begin{cases} p, & \text{if } x = 1. \\ 1 - p, & \text{if } x = 0. \end{cases}$$

• $X \sim$ **Binomial(n, p)** (where $0 \leq p \leq 1$): the number of heads in n independent flips of a coin with heads probability p.

$$p(x) = \binom{n}{x} \cdot p^x (1-p)^{n-x}$$

• $X \sim$ **Geometric(p)** (where $p > 0$): the number of flips of a coin with heads probability p until the first heads.

$$p(x) = p(1-p)^{x-1}$$

• $X \sim$ **Poisson(λ)** (where $\lambda > 0$): a probability distribution over the nonnegative integers used for modeling the frequency of rare events.

$$p(x) = e^{-\lambda} \frac{\lambda^x}{x!}$$

Continuous random variables

• $X \sim$ **Uniform**(a, b) (where $a < b$): equal probability density to every value between a and b on the real line.

$$f(x) = \begin{cases} \frac{1}{b-a}, & \text{if } a \leq b. \\ 0, & \text{otherwise.} \end{cases}$$

• $X \sim$ **Exponential**(λ) (where $\lambda > 0$): decaying probability density over the nonnegative reals.

$$f(x) = \begin{cases} \lambda e^{-\lambda x}, & \text{if } x \geq 0. \\ 0, & \text{otherwise.} \end{cases}$$

• $X \sim$ **Normal**(μ, σ^2): also known as the Gaussian distribution

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

3. Two random variables

Thus far, we have considered single random variables. In many situations, however, there may be more than one quantity that we are interested in knowing during a random experiment. For instance, in an experiment where we flip a coin ten times, we may care about both $X(\omega) =$ the number of heads that come up as well as $Y(\omega) =$ the length of the longest run of consecutive heads. In this section, we consider the setting of two random variables.

3.1 Joint and marginal distributions

Suppose that we have two random variables X and Y . One way to work with these two random variables is to consider each of them separately. If we do that we will only need $F_X(x)$ and $F_Y(y)$. But if we want to know about the values that X and Y assume simultaneously during outcomes of a random experiment, we require a more complicated structure known as the joint cumulative distribution function of X and Y, defined by

$$F_{XY}(x, y) = P(X \leq x, Y \leq y).$$

It can be shown that by knowing the joint cumulative distribution function, the probability of any event involving X and Y can be calculated.

The joint CDF $F_{XY}(x, y)$ and the joint distribution functions $F_X(x)$ and $F_Y(y)$ of each variable separately are related by

$$F_X(x) = lim_{y \to \infty} F_{XY}(x, y)$$

$$F_Y(y) = lim_{x \to \infty} F_{XY}(x, y)$$

Here, we call $F_X(x)$ and $F_Y(y)$ the **marginal cumulative distribution functions** of $F_{XY}(x, y)$.

Properties:

$0 \leq F_{XY}(x, y) \leq 1$.

$lim_{x,y \to \infty} F_{XY}(x, y) = 1$.

$lim_{x,y \to -\infty} F_{XY}(x, y) = 0$.

$F_X(x) = lim_{y \to \infty} F_{XY}(x, y)$.

3.2 Joint and marginal probability mass functions

If X and Y are discrete random variables, then the joint probability mass function $p_{XY} : Val(X) \times Val(Y) \to [0, 1]$ is defined by

$$p_{XY}(x, y) = P(X = x, Y = y).$$

Here, $0 \leq P_{XY}(x, y) \leq 1$ for all x, y, and $\sum_{x \in Val(X)} \sum_{y \in Val(Y)} P_{XY}(x, y) = 1$.

How does the joint PMF over two variables relate to the probability mass function for each variable separately? It turns out that

$$p_X(x) = \sum_y p_{XY}(x, y).$$

and similarly for $p_Y(y)$. In this case, we refer to $p_X(x)$ as the **marginal probability mass function** of X. In statistics, the process of forming the marginal distribution with respect to one variable by summing

out the other variable is often known as "marginalization".

3.3 Joint and marginal probability density functions

Let X and Y be two continuous random variables with joint distribution function F_{XY} . In the case that $F_{XY}(x, y)$ is everywhere differentiable in both x and y, then we can define the joint probability density function,

$$f_{XY}(x, y) = \frac{\partial^2 F_{XY}(x, y)}{\partial x \partial y}$$

Like in the single-dimensional case, $f_{XY}(x, y) \neq P(X = x, Y = y)$, but rather

$$\int \int_{(x,y) \in A} f_{XY}(x, y) dx dy = P((X, Y) \in A).$$

Note that the values of the probability density function $f_{XY}(x, y)$ are always nonnegative, but they may be greater than 1. Nonetheless, it must be the case that $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{XY}(x, y) = 1$.

Analagous to the discrete case, we define

$$f_X(x) = \int_{-\infty}^{\infty} f_{XY}(x, y) dy$$

as the **marginal probability density function** (or **marginal density**) of X, and similarly for $f_Y(y)$.

3.4 Conditional distributions

Conditional distributions seek to answer the question, what is the probability distribution over Y, when we know that X must take on a certain value x? In the discrete case, the conditional probability mass function of X given Y is simply

$$p_{Y|X}(y \mid x) = \frac{p_{XY}(x, y)}{p_X(x)},$$

assuming that $p_X(x) \neq 0$.

In the continuous case, the situation is technically a little more complicated because the probability that a continuous random variable X takes on a specific value x is equal to zero. Ignoring this technical point, we simply define, by analogy to the discrete case, the *conditional probability density* of Y given $X = x$ to be

$$f_{Y|X}(y \mid x) = \frac{f_{XY}(x,y)}{f_X(x)}$$

provided $f_X(x) \neq 0$.

3.5 Chain rule

The chain rule we derived earlier for events can be applied to random variables as follows:

$$\begin{aligned} & p_{X_1,\cdots X_n}(x_1,\cdots,x_n) \\ = & p_{X_1}(x_1)p_{X_2|X_1}(x_2 \mid x_1)\cdots p_{X_n|X_1,\cdots,X_{n-1}}(x_n \mid x_1,\cdots,x_{n-1}) \end{aligned}$$

3.6 Bayes's rule

A useful formula that often arises when trying to derive expression for the conditional probability of one variable given another, is **Bayes's rule**.

In the case of discrete random variables X and Y ,

$$P_{Y|X}(y \mid x) = \frac{P_{XY}(x,y)}{P_X(x)} = \frac{P_{X|Y}(x \mid y)P_Y(y)}{\sum_{y' \in Val(Y)} P_{X|Y}(x \mid y')P_Y(y')}$$

If the random variables X and Y are continuous,

$$f_{Y|X}(y \mid x) = \frac{f_{XY}(x,y)}{f_X(x)} = \frac{f_{X|Y}(x \mid y)f_Y(y)}{\int_{-\infty}^{\infty} f_{X|Y}(x \mid y')f_Y(y')dy'}$$

3.7 Independence

Two random variables X and Y are independent if $F_{XY}(x,y) = F_X(x)F_Y(y)$ for all values of x and y. Equivalently,

For discrete random variables,
$p_{XY}(x,y) = p_X(x)p_Y(y)$ for all

$x \in Val(X)$, $y \in Val(Y)$.

For discrete random variables,
$p_{Y|X}(y \mid x) = p_Y(y)$ whenever $p_X(x) \neq 0$
for all $y \in Val(Y)$.

For continuous random variables,
$f_{XY}(x, y) = f_X(x) f_Y(y)$ for all $x, y \in \mathbb{R}$.

For continuous random variables,
$f_{Y|X}(y \mid x) = f_Y(y)$ whenever $f_X(x) \neq 0$
for all $y \in \mathbb{R}$.

Informally, two random variables X and Y are independent if "knowing" the value of one variable will never have any effect on the conditional probability distribution of the other variable, that is, you know all the information about the pair (X, Y) by just knowing $f(x)$ and $f(y)$. The following lemma formalizes this observation:

Lemma 3.1. If X and Y are independent then for any subsets $A, B \subseteq \mathbb{R}$, we have,

$$P(X \in A, Y \in B) = P(X \in A)P(Y \in B)$$

By using the above lemma one can prove that if X is independent of Y then any function of X is independent of any function of Y.

3.8 Expectation and covariance

Suppose that we have two discrete random variables X, Y and $g : \mathbb{R}^2 \to \mathbb{R}$ is a function of these two random variables. Then the expected value of g is defined in the following way,

$$E[g(X, Y)] = \sum_{x \in Val(X)} \sum_{y \in Val(Y)} g(x, y) p_{XY}(x, y).$$

For continuous random variables X, Y, the analogous expression is

$$E[g(X, Y)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x, y) f_{XY}(x, y) dx dy.$$

We can use the concept of expectation to study the relationship of two random variables with each other. In particular, the covariance of two random variables X and Y is defined as

$$Cov[X,Y] = E[(X - E[X])(Y - E[Y])]$$

Using an argument similar to that for variance, we can rewrite this as,

$$\begin{aligned} Cov[X,Y] &= E[(X - E[X])(Y - E[Y])] \\ &= E[XY - XE[Y] - YE[X] + E[X]E[Y]] \\ &= E[XY] - E[X]E[Y] - E[Y]E[X] + E[X]E[Y] \\ &= E[XY] - E[X]E[Y]. \end{aligned}$$

Here, the key step in showing the equality of the two forms of covariance is in the third equality, where we use the fact that $E[X]$ and $E[Y]$ are actually constants which can be pulled out of the expectation. When $Cov[X,Y] = 0$, we say that X and Y are uncorrelated.

Properties:

(Linearity of expectation)
$E[f(X,Y) + g(X,Y)] = E[f(X,Y)] + E[g(X,Y)]$.

$Var[X+Y] = Var[X] + Var[Y] + 2Cov[X,Y]$.

If X and Y are independent, then
$Cov[X,Y] = 0$.

If X and Y are independent, then
$E[f(X)g(Y)] = E[f(X)]E[g(Y)]$.