6.804/9.66/9.660 Recitation #1, Fall 2021

Basic probability theory refresher

September 15, 2021 / September 17, 2021

1 Probability

One of the central ideas of this course is that human cognitive behavior can be described as the result of principled probabilistic inference. We'll see plenty of examples supporting this claim over the semester.

This recitation is designed to give you the basic tools you'll need in this course for describing that probabilistic inference: statistics and probability.

1.1 Probability functions

We use **probability functions** to quantify our beliefs about states of the world. In class, we built a simple distribution describing our beliefs about the fairness of a coin:

$$\begin{aligned} P(\text{fair coin}) &= 999/1000 \\ P(\text{unfair coin}) &= 1/1000 \end{aligned}$$

This is a valid example of a probability mass function. It allocates some fractional mass to two possibilities. A formal requirement of any probability mass function is that it allocate a *fixed* total mass to possible events, and a non-negative mass to each event. In our case, the following must hold:

$$0 \leq P(\text{fair coin}) \leq 1 \tag{1}$$

$$0 \leq P(\text{unfair coin}) \leq 1 \tag{2}$$

$$P(\text{fair coin}) + P(\text{unfair coin}) = 1 \tag{3}$$

1.2 Random variables

We used the informal terms "fair coin" and "unfair coin" above to describe possible states of the world. More generally, we can name a **random variable** that captures these different states. A random variable is a variable whose value is the result of some random or stochastic process. They will become useful once we start describing more complex systems with many possible states.

In our case, we can let X be a binary random variable, which is 0 if the coin is fair and 1 otherwise. We can then restate the above probability function more formally as follows:

$$P(X = 0) = 999/1000 \tag{4}$$
$$P(X = 1) = 1/1000 \tag{5}$$

We typically use the lowercase form of the variable to denote specific assignments of the variable. Thus $P(X = x)$ is the probability that X take on the value x. For convenience, we often abbreviate $P(X = x)$ as $P_X(x)$ — or, in even more sloppy notation, simply $P(X)$ (where the x is now implicit). You should get used to reading these different notations, as they are used interchangeably in the literature we'll be reading in the class.

1.3 Conditional probability

A **conditional probability function** states the probability of a random variable taking on a particular value, conditioned on the assignments of other random variables. In class we defined a particular conditional probability function, called a *likelihood function*, to describe the probability of observing certain coin flip sequences conditioned on the fairness of the coin. Let D be a random sequence of 5 coin flips. We defined the following conditional probability function:

$$P(D = \texttt{HHTHT} \mid X = 0) = 1/2^5 \tag{6}$$
$$P(D = \texttt{HHTHT} \mid X = 1) = 0 \tag{7}$$

Intuitively speaking, the above function describes our beliefs about D taking on a certain value after finding out that X takes on a particular value.

1.4 Joint probability

A **joint probability function** states the probability of two or more random variables taking on particular values. We can define this abstractly in terms of conditional probability, for two random variables X and Y:

$$P(X = x \wedge Y = y) \triangleq P(Y = y)P(X = x \mid Y = y) \tag{8}$$

Exercise 1. What is $P(X = 0 \wedge X = 1)$?

Exercise 2. What is $P(D = \texttt{HHTHT} \wedge X = 0)$? Describe in words what this means, and how it differs from **??**.

1.5 Independence

We say two random variables X and Y are **independent** when the following holds for all x and y:

$$P(X = x \mid Y = y) = P(X = x) \tag{9}$$

We write $X \perp\!\!\!\perp Y$ to indicate that X and Y are independent.

Independence between two random variables holds when knowledge of one variable's value provides no information about the other variable's value.

Exercise 3. Show that, if $P(X = x \mid Y = y) = P(X = x)$, then $P(Y = y \mid X = x) = P(Y = y)$. This shows that independence is a *symmetric* relation.

1.6 Bayes' rule

Bayes' rule can be derived from the definitions above.

$$\begin{aligned} P(X \wedge Y) &= P(Y)P(X \mid Y) && (\textbf{??}) \\ P(X)P(Y \mid X) &= P(Y)P(X \mid Y) && \text{(expand LHS with } \textbf{??}\text{)} \\ P(Y \mid X) &= \frac{P(Y)P(X \mid Y)}{P(X)} && \text{(assuming } P(X) \neq 0\text{)} \end{aligned}$$

To be concrete, let's put that in terms of the coin-flipping model. Recall that X describes whether the coin is fair, and D describes the particular sequence we observe.

$$P(X \mid D) = \frac{P(D \mid X)P(X)}{P(D)} \tag{10}$$

Bayes' rule allows us to derive *beliefs over* X after observing some data D. This relation is central to all of the models we'll build in this class. It's so central that we've assign different parts of the equation separate names:

- The **prior** $P(X)$ describes our beliefs about the fairness of the coin *before* observing any data.

- The **likelihood** $P(D \mid X)$ describes our predictions about what sort of data we'll observe, assuming a particular world state holds (i.e. $X = 0$, the coin is fair, or $X = 1$, the coin is unfair).

- The **posterior** $P(X \mid D)$ describes how we should *update* our beliefs about X in virtue of the data D.

1.7 Marginalization

?? has a tricky element in its denominator. How are we supposed to calculate the term $P(D)$ — the probability of observing data D in the abstract? The **law of total probability** allows us to compute $P(D)$ in terms of definitions we already have:

$$\begin{aligned} P(D) &= \sum_x P(D \wedge X = x) \\ &= \sum_x P(X = x) P(D \mid X = x) \end{aligned} \tag{11}$$

We also call this process **marginalizing out** X from $P(D \mid X)$.

Exercise 4. Show that the above equality holds for arbitrary independent random variables $X \perp\!\!\!\perp Y$. Concretely, let X be a binary random variable which is 1 if it is currently raining and 0 otherwise. Let Y be a binary variable which is 1 if Josh drank more than two cups of coffee yesterday and 0 otherwise. Show that the above equality holds for these two random variables.

Note that **??** allows us to now compute $P(D)$ tractably. Thus we can exactly compute $P(X \mid D)$ for this example problem.

Exercise 5. Use the likelihoods given in **????** and the priors in **????**

1.8 Distributions

So far we've explicitly specified probability values for all possible settings of a random variable. Many of our models instead rely on *probability distributions*, parameterized functions which can account for much more complex types of variables.

In our analyses, we will often assume that observed quantities are distributed according to one of these models. When a random variable X follows a distribution $\mathcal{D}$, we often say that X is *drawn* from $\mathcal{D}$, X is $\mathcal{D}$*-distributed*, or write $X \sim \mathcal{D}$.

At the highest level, these distributions factor into two types:

- **Discrete** distributions specify probability functions over discrete random variables (e.g. whether a coin is fair or unfair, how many students attended this recitation today). Our examples so far have only worked with discrete types.
- **Continuous** distributions specify probability functions over continuous random variables (e.g. the heights of students in this class).

Each distribution defines its own probability function, parameterized by one or more inputs.[1] We will introduce one common example of each type below.

1.8.1 Bernoulli distribution

The Bernoulli distribution can capture the behavior of a binary random variable, of just the type we've been working with so far. Suppose X is Bernoulli-distributed with parameter $0 \leq \theta \leq 1$. Its probability function is thus

$$P(X = x; \theta) = \begin{cases} \theta & \text{if x = 1} \\ 1 - \theta & \text{otherwise} \end{cases} \tag{12}$$

Exercise 6. Verify that this Bernoulli probability function satisfies the requirements for a probability function given in **??**.

1.8.2 Gaussian distribution

The Gaussian distribution, also called the normal distribution, defines a probability function over continuous random variables. Its probability function is defined as follows:

$$p(x) = \mathcal{N}(\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma}} e^{\frac{-(x-\mu)^2}{2\sigma^2}} \tag{13}$$

Where μ is called the *mean* parameter and σ^2 is the *variance* parameter. (We also say that σ is the *standard deviation*.

Why do we care about the Gaussian distribution? It turns out a lot of random variables in the world are Gaussian-distributed (e.g. heights of male and female adults, IQ, etc.). A handy theorem from probability theory, the *central limit theorem*, shows that any random variable which is a sum or average of other independent random variables will be nearly Gaussian. This explains why Gaussianity is so common (and so useful) a baseline assumption.[2]

Exercise 7. Challenge: Verify that the mean of a Gaussian random variable is μ, and that the variance is σ^2.

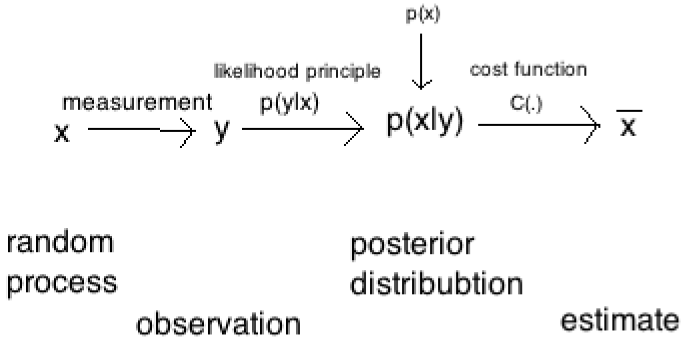
[1] We typically call probability functions on discrete-valued random variables **probability mass functions**, while functions on continuous-valued variables are **probability density functions**.

[2] Quote from some person on Quora.com: "The Central Limit Theorem says (roughly) that if something results from a lot of small influences that are not too correlated with each other, youll get a Normal distribution. Height, for example, is controlled by lots of genes, plus nutrition and other factors that work more or less independently."

2 Inference

Inference is the process of extracting information about one random variable (latent) from a set of other random variables (observed). It is important for a range of different tasks:

- Estimation / learning (infer the value of a continuous RV)
- Prediction (best estimate for next observation given past observations)
- Hypothesis testing / detection (binary)
- Modeling / model selection (choose among a set of models)



2.1 Maximum a posteriori (MAP)

A **maximum *a posteriori*** estimate is the parameter setting with the highest posterior probability given some observed data. For example, suppose we observe some Gaussian-distributed variables D and wish to estimate a new Gaussian parameter μ given some prior mean parameter μ_0. The MAP estimate μ^{MAP} is given by

$$\mu^{\text{MAP}} = \arg\max_{\mu} p(\mu \mid D) = \arg\max_{\mu} \frac{p(D \mid \mu) p(\mu \mid \mu_0)}{p(D)} \tag{14}$$

Notice that the MAP value specifies an entirely new posterior distribution.

You will derive the MAP estimate for a Gaussian distribution in the following practice problems.

2.2 Maximum likelihood

The MAP formula above lets us specify an entire posterior distribution. We are also often interested in the **maximum likelihood estimate** (MLE) for a parameter — that is, the point μ at which the likelihood $p(D \mid \mu)$ is maximized:

$$\mu^{\text{MLE}} = \arg\max_{\mu} p(D \mid \mu) \tag{15}$$

We will also derive the MLE equation for Gaussians in the following problems.