Inference

In this course, we will focus on *probabilistic inference.* This is the process of inferring unknown properties of a system given observations via the mechanics of probability theory.

For example, suppose we want to understand some aspect of the population of the United States, such as the portion of adults who prefer Coca Cola over Pepsi. Let this unknown value be called θ. How can I gain insight into this value?

One thing I could do is ask people what they believe about θ, which they might reasonably compactly communicate via a probability distribution (or equivalently by a drawing curve on a piece of paper). Note that different people might give different answers to this question! (What do *you* think are plausible values of θ?)

Of course, the Coca Cola corporation probably wouldn't be too happy if I simply drew my personal beliefs about this value and submitted a report. Instead, we could conduct a survey to gain some more information about θ. So we contact some adults and ask them if they prefer Coca Cola over Pepsi, and record the results. Let's call the results of this experiment $\mathcal{D}$ (for "data").

What do we do with this data once we've measured it? The goal of probabilistic inference is to make some statements about θ given these observations.

Probability

There are just two laws of probability you need to know. The first is the *sum rule,* also called the *rule of total probability.* Suppose X is some event (an event is simply a collection of things that could be true, for example "the total value shown on two dice is less than six"), and suppose $\{Y_i\}_{i=1}^N$ are some mutually exclusive and exhaustive events ("for example, Y_i could be the event "the first rolled die has value i," for $i \in \{1, 2, \ldots, 6\}$). Then:

$$\Pr(X) = \sum_{i=1}^{N} \Pr(X, Y_i),$$

so the total probability of X is the sum of the probabilities of X occurring alongside each of the scenarios described by the $\{Y_i\}$.

The second law of probability to know is the *product rule:*

$$\Pr(X, Y) = \Pr(Y \mid X) \Pr(X).$$

Here $\Pr(Y \mid X)$, read "the probability of Y given X," is the probability of event Y when we restrict to only cases where X is true. This is easy to see from a Venn diagram. By manipulating the product rule, we arrive at *Bayes' theorem:*

$$\begin{aligned}\Pr(Y \mid X) &= \frac{\Pr(X \mid Y) \Pr(Y)}{\Pr(X)} \\ &= \frac{\Pr(X \mid Y) \Pr(Y)}{\sum \Pr(X \mid Y) \Pr(Y)}.\end{aligned}$$

One possible way to interpret this result is as follows. Suppose we are interested in the truth of Y. We begin with a prior belief about Y, $\Pr(Y)$. We then learn that X is true. We use the formula above to update our belief about Y by conditioning on this new information, giving $\Pr(Y \mid X)$. Bayes' theorem therefore gives a probabilistically consistent way to update one's beliefs given new

information. In this context the value $\Pr(Y)$ is called a *prior probability*, as it represents our beliefs prior to observing X. The output, $\Pr(Y \mid X)$ is called the *posterior probability.*

The process of updating beliefs in this way is often called *probability inversion,* because after observing X, we first calculate $\Pr(X \mid Y)$ ("how likely was it to observe X if Y were true?"), then invert it using Bayes' theorem to give the desired probability ("how likely is Y now that I've seen X?"). Many classical paradoxes arise because these probabilities can be quite different from each other, depending on the prior $\Pr(Y)$.

What is the prior probability that the outcome of rolling two dice is less than six? What if I told you that the value of the first die was a five, does that change anything?

The above results are only valid for discrete random variables X. Although this can be useful in machine learning (for example, if these variables correspond to different hypotheses we wish to compare), in practice, we will often be interested in continuous variables, such as θ above. Thankfully, we above formulas are perfectly valid for continuous random variables x and θ, where we replace probabilities $\Pr$ with probability density functions p and replace sums with integrals:

$$p(x) = \int p(x, y) \, \mathrm{d}y;$$
$$p(y \mid x) = \frac{p(x \mid y)p(y)}{\int p(x \mid y)p(y) \, \mathrm{d}y}.$$

We will always assume that probability density functions exist, because the cases where they don't are not typically encountered in machine learning. It also saves us from having to state "when the density exists" all the time.

Returning to our survey example, we might begin with a prior belief about the value of θ, represented by a prior probability distribution $p(\theta)$. To couple our observations $\mathcal{D}$ to the value of interest, we construct a probabilistic model $\Pr(\mathcal{D} \mid \theta)$, which describes how likely we would see a particular survey result $\mathcal{D}$ given a particular value of θ. Note that this model could have any form and we are free to make it as complicated as we'd like: was there bias in the sampling mechanism? Do people always tell the truth?

Finally, we use these to compute the posterior probability of θ given the survey results, $p(\theta \mid \mathcal{D})$. This posterior distribution encapsulates our entire belief about θ! We can use it to answer various questions we might be about θ.

The Bayesian Method

There are four main steps to the Bayesian approach to probabilistic inference:[1]

- **Likelihood.** First, we construct the likelihood (or *model*), $p(\mathcal{D} \mid \theta)$. This serves to describe the mechanism giving rise our observations $\mathcal{D}$ given a particular value of the parameter of interest θ.
- **Prior.** Next, we summarize our prior beliefs about the parameters θ, which we encode via a probability distribution $p(\theta)$.
- **Posterior.** Given some observations $\mathcal{D}$, we obtain the posterior distribution $p(\theta \mid \mathcal{D})$ using Bayes' theorem.

[1] These are summarized from Tony O'Hagan and Jonathan Forster's eloquent introduction in *Kendall's Advanced Theory of Statistics Volume 2B.*

- **Inference.** We now use the posterior distribution to draw further conclusions as required.

The last step is purposely open-ended. For example, we can use it to make predictions about new data (as in supervised learning), we can summarize it in various ways (for example, point estimation if we must report a single "best guess" of θ), use it to compare alternative models (giving rise to Bayesian model comparison), determine which data to obtain next (optimal design of experiments, called "active learning" in machine learning), and more. We will consider several of these in this course.

Issues

Bayesian inference is a completely consistent system for probabilistic reasoning. Unfortunately, it is not without its issues, some of which we list below.

Origin of priors

In contrast to the model $p(\mathcal{D} \mid \theta)$, it is not usually clear where the prior $p(\theta)$ should come from. There is an entire branch of study concerning prior elicitation, but for now we will simply treat it as given. We will continue to discuss this throughout the course. Indeed, we will see that several "tricks" often encountered in alternative approaches (such as *regularization*) can be interpreted as implicitly placing particular prior beliefs on θ.

The meaning of probability

Another problem is what exactly *probability* means. The dominant statistical practice for many years (known as the *classical* or *frequentist* theory) defines probability in terms of the limit of conducting infinitely many random experiments. Therefore it is impossible to consider the "probability" of a statement such as "at least 50% of adults prefer Coca Cola." This statement is either true or false, so its frequentist probability is either zero or one (but we might not know which). In the Bayesian interpretation, we allow probabilities instead to describe *degrees of belief* in such a proposition. In this way, we can treat everything as a random variable and use the tools of probability to carry out all inference. That is, in Bayesian probability, parameters, data, and hypotheses are all treated the same. This viewpoint is not universally accepted, and there is a lot of fascinating philosophical writing on the subject, which we will largely avoid.

Note that the two interpretations of probability agree on the axioms and theorems of probability theory. No one argues the truth of Bayes' theorem. The main difference is that a frequentist would not allow a probability distribution to be placed on parameters such as θ, so the use of Bayes' theorem to update beliefs about parameters in light of data is not allowed in that framework.

Intractable integrals

Unfortunately, the integral in the denominator of Bayes' theorem:

$$p(x) = \int p(x \mid y) p(y) \, \mathrm{d}y$$

is not in general tractable for arbitrary combinations of priors and likelihoods. For this reason, we will spend a lot of time discussing various schemes to approximate the posterior distribution in such cases. Sometimes this can be more of an art than a science.