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 I want to understand some aspect of the CMU student population, such as the portion of students who believe that cats are better than dogs. Let's call this unknown value 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). Different people might give very different answers to this question! (What do you think are plausible values of θ?)

Of course, asking individual people to guess about θ probably won't give very accurate results. Instead, we could conduct a survey to gain some more information about θ. So I reach out to some CMU students and ask them if they think cats are better than dogs 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, where an event is simply a collection of things that could be true (e.g., "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 (e.g., 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 $\Pr(Y)$.

The above results are only valid for discrete random variables X. Although this can be useful in machine learning (e.g., if these variables correspond to different hypotheses we wish to compare), in practice, we will often be interested in continuous variables, such as θ from the example above. Thankfully, the above formulas are perfectly valid for continuous random variables if 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 some bias in the sampling mechanism that we need to account for? Do we assume that respondents 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

To summarize, there are four main steps to the Bayesian approach to probabilistic inference (these are summarized from Tony O'Hagan and Jonathan Forster's excellent introduction in *Kendall's Advanced Theory of Statistics, Volume 2B*):

- **Likelihood.** First, we construct the likelihood (or **model**), $p(\mathcal{D} \mid \theta)$. This describes the mechanism giving rise to our observations $\mathcal{D}$ given a setting of the parameters 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.
- **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 (e.g., point estimation if we must report a single "best guess" of θ), use it to compare alternative models (a.k.a. Bayesian model comparison), determine which data to obtain next (optimal design of experiments or "active learning" as its more commonly referred to 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 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 *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 CMU students prefer cats to dogs." 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 entirely 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.

Coin flipping

Suppose there is a coin that may be biased – this coin has unknown probability θ of giving a "heads." If we repeatedly flip this coin and observe the outcomes, how can we maintain our belief about θ?

Note that the coin-flipping problem can be seen as a simplification of the survey problem we discussed last time, where we assume that people always tell the truth, are sampled uniformly at random, and whose opinions are generated independently (perhaps by flipping a coin!).

Before we select a prior for θ, we write down the likelihood. For a particular problem, it is almost always easier to derive an appropriate likelihood than it is to identify an appropriate prior distribution.

Suppose we flip the coin n times and observe x "heads." Every statistician, regardless of philosophy, would agree that the probability of this observation, given the value of θ, comes from a binomial distribution:

$$\Pr(x \mid n, \theta) = \binom{n}{x} \theta^x (1-\theta)^{n-x}.$$

Classical method

Before we continue with the Bayesian approach, we pause to discuss how a classical statistician would proceed with this problem. Recall that in the frequentist approach, the value θ can only be considered in terms of the frequency of success ("heads") seen during an infinite number of trials. It is not valid in this framework to represent a "belief" about θ in terms of probability.

Rather, the frequentist approach to reasoning about θ is to construct an *estimator* for θ, which in theory can be any function of the observed data: $\hat{\theta}(x, n)$. Estimators are then analyzed in terms of their behavior as the number of observations goes to infinity (for example, we might prove that $\hat{\theta} \to \theta$ as $n \to \infty$). The classical estimator in this case is the empirical frequency $\hat{\theta} = x/n$.

Bayesian method

An interesting thing to note about the frequentist approach is that it ignores all prior information, opting instead to only look at the observed data. To a Bayesian, every such problem is different and should be analyzed contextually given the known information.

With the likelihood decided, we must now choose a prior distribution $p(\theta)$. A convenient prior in this case is the **beta distribution,** which has two parameters α and β:

$$p(\theta \mid \alpha, \beta) = \mathcal{B}(\theta; \alpha, \beta) = \frac{1}{B(\alpha, \beta)} \theta^{\alpha-1} (1-\theta)^{\beta-1}.$$

Here the normalizing constant $B(\alpha, \beta)$ is the **beta function:**

$$B(\alpha, \beta) = \int_0^1 \theta^{\alpha-1} (1-\theta)^{\beta-1} \, \mathrm{d}\theta.$$

The support of the beta distribution is $\theta \in (0, 1)$, and by selecting various values of α and β, we can control its shape to represent a variety of different prior beliefs.

Given our observations $\mathcal{D} = (x, n)$, we can now compute the posterior distribution of θ:

$$p(\theta \mid x, n, \alpha, \beta) = \frac{\Pr(x \mid n, \theta) p(\theta \mid \alpha, \beta)}{\int \Pr(x \mid n, \theta) p(\theta \mid \alpha, \beta) \, \mathrm{d}\theta}.$$

First we handle the normalization constant $\Pr(x \mid n, \alpha, \beta)$:

$$\begin{aligned}\int \Pr(x \mid n, \theta) p(\theta \mid \alpha, \beta)\, \mathrm{d}\theta &= \binom{n}{x} \frac{1}{B(\alpha, \beta)} \int_0^1 \theta^{\alpha+x-1}(1-\theta)^{\beta+n-x-1}\, \mathrm{d}\theta \\ &= \binom{n}{x} \frac{B(\alpha+x, \beta+n-x)}{B(\alpha, \beta)}.\end{aligned}$$

Now we apply Bayes theorem:

$$\begin{aligned}p(\theta \mid x, n, \alpha, \beta) &= \frac{\Pr(x \mid n, \theta) p(\theta \mid \alpha, \beta)}{\int \Pr(x \mid n, \theta) p(\theta \mid \alpha, \beta)\, \mathrm{d}\theta} \\ &= \left[\binom{n}{x} \frac{B(\alpha+x, \beta+n-x)}{B(\alpha, \beta)}\right]^{-1} \left[\binom{n}{x} \theta^x (1-\theta)^{n-x}\right] \left[\frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha, \beta)}\right] \\ &= \frac{1}{B(\alpha+x, \beta+n-x)} \theta^{\alpha+x-1}(1-\theta)^{\beta+n-x-1} \\ &= \mathcal{B}(\alpha+x, \beta+n-x).\end{aligned}$$

The posterior is therefore another beta distribution with parameters $(\alpha+x, \beta+n-x)$! Specifically, we have added the number of successes to the first parameter and the number of failures to the second.

The rather convenient fact that the posterior remains a beta distribution is because the beta distribution satisfies a property known as **conjugacy** with the binomial likelihood. This fact also leads to a common interpretation of the parameters α and β: they serve as "pseudocounts," or fake observations we pretend to have seen before seeing the data.

Figure 1 shows the relevant functions for the coin flipping example for $(\alpha, \beta) = (3, 5)$ and $(x, n) = (7, 8)$. Notice that the likelihood favors higher values of θ, whereas the prior had favored lower values of θ. The posterior, taking into account both sources of information, lies in between these extremes. Notice also that the posterior has support over a narrower range of plausible θ values than the prior; this is because we can draw more confident conclusions from having access to more information.

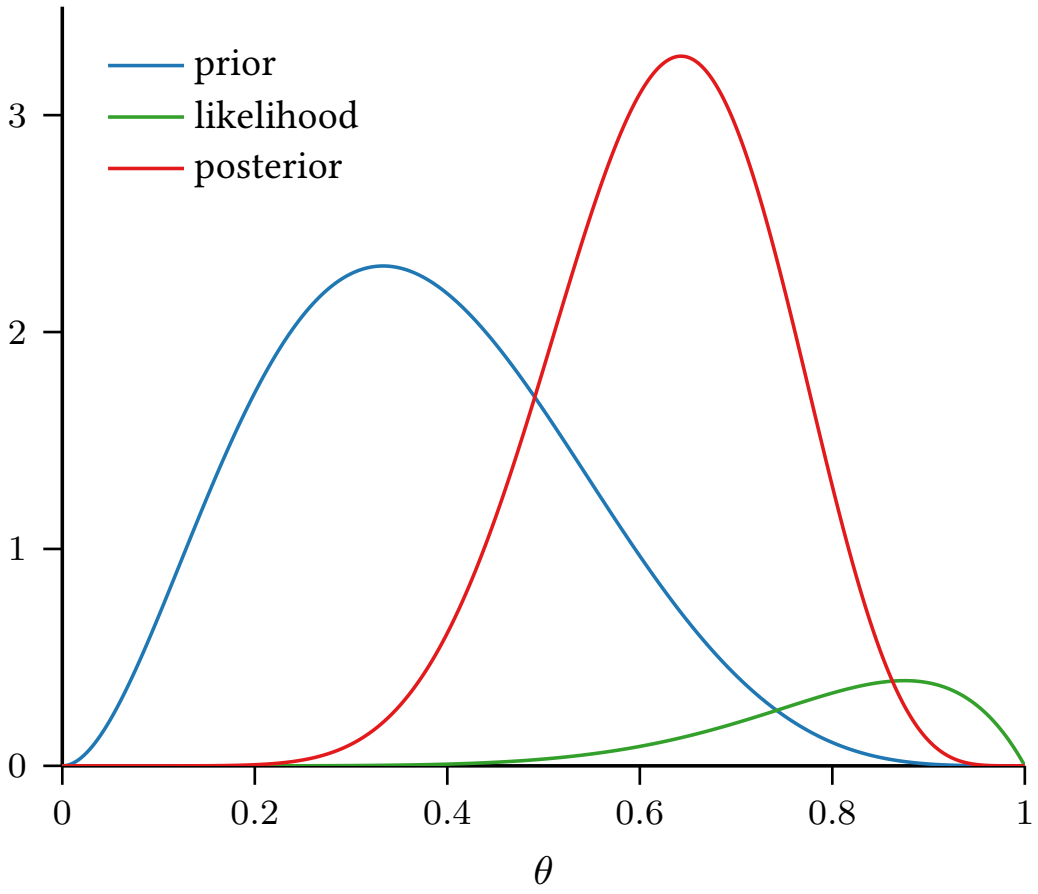

Figure 1: An example of Bayesian updating for coin flipping.