Lisa Yan and Jerry Cain
CS109

Lecture Notes #21
October 30, 2020

The Beta Distribution

Based on a chapter by Chris Piech

Pre-recorded lecture: Sections 1 and 3.1
In-lecture: Sections 2, 3.2, 4.1
Not covered: Section 4.2

In this chapter we are going to have a very meta discussion about how we represent probabilities. Until now probabilities have just been numbers in the range 0 to 1. However, if we have uncertainty about our probability, it would make sense to represent our probabilities as random variables (and thus articulate the relative likelihood of our belief).

1 Mixing Discrete and Continuous Random Variables

In order to characterize probabilities as random variables (recall that according to Axiom 1, probabilities are real values between 0 and 1) in the context of discrete experiment outcomes (e.g., number of heads in a certain number of coin flips), we must introduce one more concept: Bayes' theorem that mixes discrete PMFs with continuous PDFs.

These equations are straightforward once you have your head around the notation for probability density functions ($f_X(x)$) and probability mass functions ($p_X(x)$).

Let X be continuous random variable and let N be a discrete random variable. The conditional probabilities of X given N and N given X respectively are:

$$f_{X|N}(x \mid n) = \frac{p_{N|X}(n \mid x) f_X(x)}{p_N(n)} \qquad\qquad p_{N|X}(n \mid x) = \frac{f_{X|N}(x \mid n) p_N(n)}{f_X(x)}$$

2 Estimating Probabilities

Imagine we have a coin and we would like to know its probability of coming up heads (p). We flip the coin $(n + m)$ times and it comes up head n times. One way to calculate the probability is to assume that it is exactly $p = \frac{n}{n+m}$. That number, however, is a coarse estimate, especially if $n + m$ is small. Intuitively it doesn't capture our uncertainty about the value of p. Just like with other random variables, it often makes sense to hold a distributed belief about the value of p.

To formalize the idea that we want a distribution for p we are going to use a random variable X to represent the probability of the coin coming up heads. Before flipping the coin, we could say that our belief about the coin's success probability is uniform: $X \sim Uni(0, 1)$.

If we let N be the number of heads that came up, given that the coin flips are independent, $(N|X) \sim Bin(n+m, x)$. We want to calculate the probability density function for $X|N$. We can start by applying Bayes Theorem:

$$
\begin{aligned}
f(X = x|N = n) &= \frac{P(N = n|X = x)f(X = x)}{P(N = n)} && \text{Bayes Theorem} \\
&= \frac{\binom{n+m}{n}x^n(1-x)^m}{P(N = n)} && \text{Binomial PMF, Uniform PDF} \\
&= \frac{\binom{n+m}{n}}{P(N = n)}x^n(1-x)^m && \text{Moving terms around} \\
&= \frac{1}{c} \cdot x^n(1-x)^m && \text{where } c = \int_0^1 x^n(1-x)^m dx
\end{aligned}
$$

3 Beta Random Variable

3.1 The Beta Distribution

The equation that we arrived at when using a Bayesian approach to estimating our probability defines a probability density function and thus a random variable. The random variable is called a Beta distribution, and it is defined as follows:

The Probability Density Function (PDF) for a Beta $X \sim Beta(a, b)$ is:

$$
f(X = x) = \begin{cases} \frac{1}{B(a,b)}x^{a-1}(1-x)^{b-1} & \text{if } 0 < x < 1 \\ 0 & \text{otherwise} \end{cases} \qquad \text{where } B(a, b) = \int_0^1 x^{a-1}(1-x)^{b-1}dx
$$

A Beta distribution has $E[X] = \frac{a}{a+b}$ and $Var(X) = \frac{ab}{(a+b)^2(a+b+1)}$. All modern programming languages have a package for calculating Beta CDFs. You will not be expected to compute the CDF by hand in CS109.

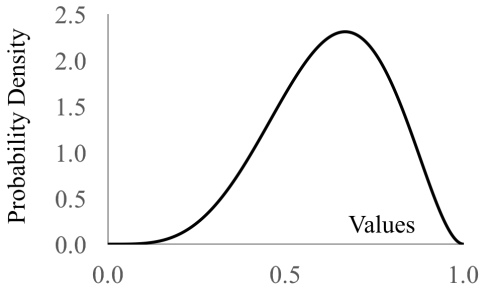
3.2 Flipping a coin with unknown probability

To model our estimate of the probability of a coin coming up heads as a beta set $a = n + 1$ and $b = m + 1$. Beta is used as a random variable to represent a belief distribution of probabilities in contexts beyond estimating coin flips. It has many desirable properties: it has a support range that is exactly $(0, 1)$, matching the values that probabilities can take on and it has the expressive capacity to capture many different forms of belief distributions.

Let's imagine that we had observed $n = 4$ heads and $m = 2$ tails. The probability density function for $X \sim \text{Beta}(5, 3)$ is:

Notice how the most likely belief for the probability of our coin is when the random variable, which represents the probability of getting a heads, is $4/6$, the fraction of heads observed. This distribution shows that we hold a non-zero belief that the probability could be something other than $4/6$. It is unlikely that the probability is 0.01 or 0.09, but reasonably likely that it could be 0.5.

It works out that $Beta(1, 1) = Uni(0, 1)$. As a result the distribution of our belief about p before ("prior") and after ("posterior") can both be represented using a Beta distribution. When that happens we call Beta a "conjugate" distribution. Practically conjugate means easy update.



4 Beta as a Prior

You can set $X \sim Beta(a, b)$ as a prior to reflect how biased you think the coin is apriori to flipping it. This is a subjective judgment that represent $a+b-2$ "imaginary" trials with $a-1$ heads and $b-1$ tails. If you then observe $n+m$ real trials with n heads you can update your belief. Your new belief would be, $X|(n \text{ heads in } n+m \text{ trials}) \sim Beta(a+n, b+m)$. Using the prior $Beta(1,1) = Uni(0,1)$ is the same as saying we haven't seen any "imaginary" trials, so apriori we know nothing about the coin. This form of thinking about probabilities is representative of the "Bayesian" field of thought where computer scientists explicitly represent probabilities as distributions (with prior beliefs). That school of thought is separate from the "Frequentist" school which tries to calculate probabilities as single numbers evaluated by the ratio of successes to experiments.

4.1 Example: Medicine

Before being tested, a medicine is believed to be effective 80% of the time. The medicine is tried on 20 patients and it works for 14 but doesn't work for 6. What is your new belief that the drug works?

Frequentist Solution: Since a frequentist view does not incorporate prior belief about probability, using Maximum Likelihood Estimation, the probability that your drug works can be estimated as $\theta = 14/20 = 0.7$.

Bayesian Solution: We need to choose some distribution of our prior belief that the medicine works 80% of the time. While there are many choices, one that will help us out is to choose a Beta prior for θ, our probability of success. This is because after we incorporate the likelihood that results from a Bernoulli experiment, a Beta prior will result in a Beta posterior on θ.

Choosing a prior: Even after we have decided on the shape of θ, we still have many choices of the exact parameters of Beta based on how strong we want to make our prior belief. We could choose $\theta \sim \text{Beta}(a = 5, b = 2)$ (where we imagine there were 4 successes in 5 imaginary prior trials), $\theta \sim \text{Beta}(a = 81, 21)$ (80 successes in 100 imaginary trials), or anything that maintains the ratio of 80% probability of success.

Computing a posterior: Suppose we choose $\text{Beta}(5, 2)$ as our prior. If our actual experiment gives 14 successes and 6 failures, we update our belief on θ to $\theta|data \sim \text{Beta}(a = 5 + 14 = 19, b = 2 + 6 = 8)$, which corresponds to 18 imaginary and real successes and 8 imaginary and real failures. The reason our updates are so easy is because Beta is a conjugate distribution to our Bernoulli experiment distribution. Woohoo!

Choosing a single value to report: The Frequentist approach reports a single value for probability of success, but the Bayesian approach reports a distribution on our probability of success. How would we get a single value from the Bayesian approach? One reasonable value to report would be the expected value of our posterior, which turns out to be $E[\theta] = a/(a+b) \approx 0.70$.

In general, we choose to report the **most likely** value of our posterior distribution—also called the **mode of** X—which is defined as $\text{mode}(\theta) = \arg\max_\theta f(\theta)$. The **mode of** $\text{Beta}(a, b) = (a-1)/(a+b-2)$, and therefore $\text{mode}(\theta) = 18/(18+7) \approx 0.72$. More on this next time!

4.2 Example: Course assignments

In one particular iteration of this course, we talked about reasons why grade distributions might be well suited to be described as a Beta distribution. Let's say that we are given a set of student grades for a single exam and we find that it is best fit by a Beta distribution: $X \sim Beta(a = 8.28, b = 3.16)$. What is the probability that a student is below the mean (i.e. expectation)?

The answer to this question requires two steps. First calculate the mean of the distribution, then calculate the probability that the random variable takes on a value less than the expectation.

$$\begin{aligned} E[X] &= \frac{a}{a+b} = \frac{8.28}{8.28+3.16} \\ &\approx 0.7238 \end{aligned}$$

Now we need to calculate $P(X < E[X])$. That is exactly the CDF of X evaluated at $E[X]$. We don't have a formula for the CDF of a Beta distribution but all modern programming languages will have a Beta CDF function. In Python using the scipy stats library we can execute stats.beta.cdf which takes the x parameter first followed by the alpha and beta parameters of your Beta distribution.

$$\begin{aligned} P(X < E[X]) &= F_X(0.7238) = \text{stats.beta.cdf(0.7238, 8.28, 3.16)} \\ &\approx 0.46 \end{aligned}$$