Will Monroe
CS 109

Lecture Notes #14
July 26, 2017

Conditional and Beta Distributions

Based on a chapter by Chris Piech

Conditional Distributions

Earlier, we looked at conditional probabilities for events. Here we formally go over conditional probabilities for random variables. The equations for both the discrete and continuous case are intuitive extensions of our understanding of conditional probability:

Discrete

The conditional probability mass function (PMF) for the discrete case:

$$p_{X|Y}(x|y) = P(X = x \mid Y = y) = \frac{P(X = x, Y = y)}{P(Y = y)} = \frac{P_{X,Y}(x, y)}{p_Y(y)}$$

The conditional cumulative density function (CDF) for the discrete case:

$$F_{X|Y}(a|y) = P(X \leq a \mid Y = y) = \frac{\sum_{x \leq a} p_{X,Y}(x, y)}{p_Y(y)} = \sum_{x \leq a} p_{X|Y}(x \mid y)$$

Continuous

The conditional probability density function (PDF) for the continuous case:

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

The conditional cumulative density function (CDF) for the continuous case:

$$F_{X|Y}(a \mid y) = P(X \leq a \mid Y = y) = \int_{-\infty}^{a} f_{X|Y}(x \mid y)dx$$

This is doing something a bit surprising: it is conditioning on an *exact* value of a continuous random variable, $Y = y$. But isn't $P(Y = y)$ equal to 0? And didn't we say you can't condition on a zero-probability event? Yes and yes. However, the notion of conditioning on an exact value of a continuous random variable is very useful: imagine trying to estimate someone's age knowing their (exact) height, for example. As a result, we define the conditional probability distribution to use the probability density function in the continuous case.

Mixing Discrete and Continuous

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:

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

Estimating Probabilities

At this point, 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).

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 heads 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: $X \sim \text{Uni}(0, 1)$.

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

$$
\begin{aligned}
f_{X|N}(x|n) &= \frac{P(N = n|X = x)f_X(x)}{P(N = n)} && \text{Bayes' Theorem} \\
&= \frac{\binom{n+m}{n}x^n(1-x)^m}{P(N = n)} && \text{binomial PMF, uniform PDF} \\
&= \frac{\binom{n+m}{n}}{P(N = n)}x^n(1-x)^m && \text{moving terms around} \\
&= \frac{1}{c} \cdot x^n(1-x)^m && \text{where } c = \int_0^1 x^n(1-x)^m dx
\end{aligned}
$$

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 random variable $X \sim \text{Beta}(a, b)$ is:

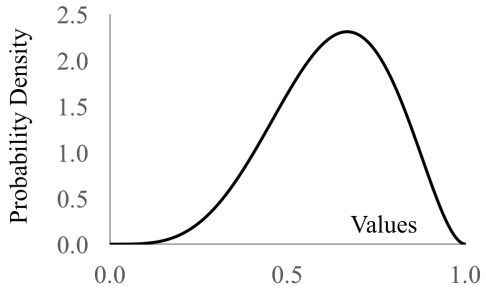
$$
f(x) = \begin{cases} \frac{1}{B(a,b)}x^{a-1}(1-x)^{b-1} & \text{if } 0 < x < 1 \\ 0 & \text{otherwise} \end{cases} \qquad \text{where } B(a, b) = \int_0^1 x^{a-1}(1-x)^{b-1}dx
$$

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

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 $X \sim \text{Beta}(5, 3)$ (note the addition of 1 to both heads and tails) 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 $\text{Beta}(1, 1) = \text{Uni}(0, 1)$. 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 to update*.

Beta as a Prior

You can set $X \sim \text{Beta}(a, b)$ as a prior to reflect how biased you think the coin is before you flip it. This is a subjective judgment that represents $a + b - 2$ "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 $X \mid (n \text{ heads in } n + m \text{ trials}) \sim \text{Beta}(a + n, b + m)$. Using the prior $\text{Beta}(1, 1) = \text{Uni}(0, 1)$ is the same as saying we haven't seen any "imaginary" trials, so a priori, we know nothing about the coin. This form of thinking about probabilities is representative of the "Bayesian" field of thought, in which probabilities are explicitly represented as distributions (with prior beliefs). That school of thought is often thought of in opposition to the "frequentist" school, which tries to calculate probabilities as single numbers evaluated by the ratio of successes to experiments.

Example 1

Distributions of student grades in a course are well modeled by a beta distribution. Suppose we are given a set of student grades for a single exam and we find that it is best fit by a Beta distribution: $X \sim \text{Beta}(a = 8.28, b = 3.16)$. What is the probability that a student is above the mean (i.e. expectation)?

The answer to this question requires two steps. First, we calculate the mean of the distribution, then we calculate the probability that the random variable takes on a value greater 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 1 minus the CDF of X evaluated at $E[X]$. We don't have a closed form for the CDF of a beta distribution, but all modern programming languages have a beta CDF function. In Python, we can execute `scipy.stats.beta.cdf`, which takes the x value first, followed by the a and b parameters of your Beta distribution.

$$\begin{aligned} P(X > E[X]) &= 1 - F_X(0.7238) = \texttt{1 - scipy.stats.beta.cdf(0.7238, 8.28, 3.16)} \\ &\approx 0.54 \end{aligned}$$