Chris Piech
CS109

Handout #22
April 27th, 2016

Beta Distribution

Our last variable type is the Beta random variable. We waiting until this point in the class to introduce Beta distributions because to really understand Beta distributions you must first understand joint distributions. Beta random variables often semantically represent probabilities.

Properties

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

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

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.

Mixing Discrete and Continuous

In order to understand a common use of the Beta function, we will need to know how to compute conditional probabilities when we mix continuous and discrete random variables. 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:

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

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, $(N|X)$ $simBin(n+m,x)$. We want to calculate the probability density function for $X|N$. We can start by applying Bayes Theorem:

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

That fits the format of the beta distribution if you set $a = n + 1$ and $b = m + 1$. It also 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.

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 judgement 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, $X|(n$ heads in $n+m$ 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 seperate from the "Frequentist" school which tries to calculate probabilities as single numbers evaluated by the ratio of successes to experiments.

Beta as a General Distribution for Probabilities

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.

Assignment Example

In class 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 bellow 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 JavaScript we can execute: jStat.beta.cdf which takes the x parameter first followed by the alpha and beta parameters of your Beta distribution.

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

Expectation

At the end of lecture (if we have time) we are going to review expectation of sums and use it to derive the expectation of different random variables. This is foreshadowing for next class...