CS109 March 7, 2024

Beta and Parameter Estimation Solution

Before you leave lab, make sure you click here so that you're marked as having attended this week's section. The CA leading your discussion section can enter the password needed once you've submitted.

1 Warmups

1.1 Beta

a. Suppose you have a coin where you have no prior belief on its true probability of heads p. How can you model this belief as a Beta distribution?

b. Suppose you have a coin which you believe is fair, with "strength" α. That is, pretend you've seen α heads and α tails. How can you model this belief as a Beta distribution?

c. Now suppose you take the coin from the previous part and flip it 10 times. You see 8 heads and 2 tails. How can you model your posterior belief of the coin's probability of heads?

a. Beta$(1, 1)$ is a uniform prior, meaning that prior to seeing the experiment, all probabilities of heads are equally likely.

b. Beta$(\alpha + 1, \alpha + 1)$. This is our prior belief about the distribution.

c. Beta$(\alpha + 9, \alpha + 3)$

1.2 Beta Sum

What is the distribution of the sum of 100 iid Betas? Let X be the sum

$$X = \sum_{i=1}^{100} X_i \qquad \text{where each } X_i \sim \text{Beta}(a = 3, b = 4)$$

Note the variance of a Beta:

$$\text{Var}(X_i) = \frac{ab}{(a+b)^2(a+b+1)} \qquad \text{where } X_i \sim \text{Beta}(a, b)$$

By the Central Limit Theorem, the sum of equally weighted IID random variables will be Normally distributed. We calculate the expectation and variance of X_i using the beta

formulas:

$$
\begin{aligned}
E(X_i) &= \frac{a}{a+b} && \text{Expectation of a Beta} \\
&= \frac{3}{7} \approx 0.43 \\
\text{Var}(X_i) &= \frac{ab}{(a+b)^2(a+b+1)} && \text{Variance of a Beta} \\
&= \frac{3 \cdot 4}{(3+4)^2(3+4+1)} \\
&= \frac{12}{49 \cdot 8} \approx 0.03
\end{aligned}
$$

$$
\begin{aligned}
X &\sim N(\mu = n \cdot E[X_i], \sigma^2 = n \cdot \text{Var}(X_i)) \\
&\sim N(\mu = 43, \sigma^2 = 3)
\end{aligned}
$$

2 Problems

2.1 Vision Test MLE

You decide that the vision tests given by eye doctors would be more precise if we used an approach *inspired* by logistic regression, which we'll learn about on Friday. In a vision test a user looks at a letter with a particular font size and either correctly guesses the letter or incorrectly guesses the letter.

You assume that the probability that a particular patient is able to guess a letter correctly is:

$$p = \sigma(\theta + f)$$

Where θ is the user's vision score and f is the font size of the letter. This formula uses the sigmoid function, which we'll study more during Week 9 of the quarter!

$$
\begin{aligned}
\sigma(z) &= \frac{1}{1+e^{-z}} \\
\frac{\partial \sigma(z)}{\partial z} &= \sigma(z)[1-\sigma(z)]
\end{aligned}
$$

Explain how you could estimate a user's vision score (θ) based on their 20 responses $(f^{(1)}, y^{(1)}) \ldots (f^{(20)}, y^{(20)})$, where $y^{(i)}$ is an indicator variable for whether the user correctly identified the ith letter and $f^{(i)}$ is the font size of the ith letter. Solve for any and all partial derivatives required by the approach you describe in your answer.

We are going to solve this problem by finding the MLE estimate of θ. To find the MLE estimate, we are going to find the argmax of the log likelihood function. To calculate argmax we are going to use gradient ascent, which requires that we know the partial derivative of

the log likelihood function with respect to theta.
First we write the log likelihood:

$$L(\theta) = \prod_{i=1}^{20} p^{y_i}(1-p)^{[1-y_i]}$$
$$LL(\theta) = \sum_{i=1}^{20} (y_i \log(p) + (1-y_i)\log(1-p))$$

Then we find the derivative of log likelihood with respect to θ. We first do this for one data point:

$$\frac{\partial LL}{\partial \theta} = \frac{\partial LL}{\partial p} \cdot \frac{\partial p}{\partial \theta}$$

We can calculate both the smaller partial derivatives independently:

$$\frac{\partial LL}{\partial p} = \frac{y_i}{p} - \frac{1-y_i}{1-p}$$
$$\frac{\partial p}{\partial \theta} = p[1-p]$$

Putting it all together for one letter:

$$\begin{aligned}\frac{\partial LL}{\partial \theta} &= \frac{\partial LL}{\partial p} \cdot \frac{\partial p}{\partial \theta} \\ &= \left[\frac{y_i}{p} - \frac{1-y_i}{1-p}\right] p[1-p] \\ &= y_i(1-p) - p(1-y_i) \\ &= y_i - p \\ &= y_i - \sigma(\theta - f)\end{aligned}$$

For all twenty examples:

$$\frac{\partial LL}{\partial \theta} = \sum_{i=1}^{20} y_i - \sigma(\theta + f^{(i)})$$

2.2 Why Boba Cares About MAP

You don't understand why there's no boba place within walking distance around campus, so you decide to start one. In order to estimate the amount of ingredients needed and the time you will spend in the business (you still need to study), you want to estimate how many orders you will receive per hour. After taking CS109, you are pretty confident that incoming orders can be considered as independent events and the process can be modeled with a Poisson.

Now the question is - what is the λ parameter of the Poisson? In the first hour of your soft opening, you are visited by 4 curious students, each of whom made an order. You have a prior belief that $f(\Lambda = \lambda) = K \cdot \lambda \cdot e^{-\frac{\lambda}{2}}$. What is the MLE estimate? What is inference of λ given the observation? What is the Maximum a Posteriori (MAP) estimate of λ? Through your process try to identify what is a point-estimate, and what is a distribution.

To find the MLE, we start from finding the likelihood function (i.e. joint probability of observed events) and find the λ that maximizes the likelihood function.

$$\begin{aligned} L(\lambda) &= \frac{\lambda^4 \cdot e^{-\lambda}}{4!} \\ LL(\lambda) &= 4\log(\lambda) - \lambda - \log(4!) \\ \frac{\partial LL}{\partial \lambda} &= \frac{4}{\lambda} - 1 \end{aligned}$$

Set $\frac{\partial LL}{\partial \lambda}$ to 0 and solve for λ.

$$\lambda = 4$$

Inference of λ given the observation:

$$f(\lambda|X = 4) = \frac{P(X = 4|\lambda) \cdot f(\lambda)}{P(X = 4)}$$

MAP estimate of λ: we find the λ that maximizes the inference given the observation, i.e. we want to solve:

$$\begin{aligned} \arg\max_{\lambda} f(\lambda|X = 4) &= \arg\max_{\lambda} \frac{P(X = 4|\lambda) \cdot f(\lambda)}{P(X = 4)} \\ &= \arg\max_{\lambda} P(X = 4|\lambda) \cdot f(\lambda) \\ &= \arg\max_{\lambda} \frac{\lambda^4 \cdot e^{-\lambda}}{4!} \cdot K \cdot \lambda \cdot e^{-\frac{\lambda}{2}} \end{aligned}$$

Take log.

$$log(\frac{\lambda^4 \cdot e^{-\lambda}}{4!} \cdot K \cdot \lambda \cdot e^{-\frac{\lambda}{2}}) = 4log(\lambda) - \lambda + 1 + log(K) + log(\lambda) - \frac{\lambda}{2}$$

Differentiate with respect to λ, set to 0 and solve.

$$\begin{aligned} \frac{5}{\lambda} - 1 - \frac{1}{2} &= 0 \\ \lambda &= \frac{10}{3} \end{aligned}$$

2.3 Dirk/Evan Showdown

Dirk and Evan are set to compete in a match of seven games, and the player winning the most will be rewarded with a medal from the Ruler of all of Bayestopia, Queen Doris. The outcomes of each of the seven games are independent of one another, and Dirk wins each game with a probability p (so that Evan wins with probability $1 - p$). Unfortunately, p is unknown, so you decide to model p itself as a Beta random variable such that $p \sim Beta(1, 1)$.

To learn more about p, you read up on all of their previous games, find that they've already competed 12 times, and learn that Evan has won 7 of those 12 games in this order: WLLLWWLWWLWW. Even if Evan's edge is ever so slight, he appears to be the a priori favorite.

a. Find the posterior distribution of p given Dirk and Evan's prior history competing against one another.

b. Recall that the PDF of a $Beta$ distribution on integer parameters a and b is given as:

$$f(x|a, b) = \frac{1}{B(a, b)} x^{a-1}(1 - x)^{b-1}, \text{ where } B(a, b) \text{ is a normalization constant}$$

Assuming our hyperparameters a and b are positive integers, explain why the expected value of $p^m(1 - p)^n$—that is, $E[p^m(1 - p)^n]$—is given by

$$E[p^m(1 - p)^n] = \frac{B(a + m, b + n)}{B(a, b)}$$

You can assume that m and n are nonnegative integers as well.

c. Relying only on the posterior distribution of p that you computed for part a and the result from part b, compute the expected probability that the best-of-seven match between Dirk and Evan isn't settled after the first six games and therefore requires the seventh be played?

a. Because the prior is $p \sim Beta(1, 1)$, and because the likelihood function is a Binomial with $n = 12$ and unknown p, our posterior is also a Beta, but with parameters of $a = 1 + 5 = 6$ and $b = 1 + 7 = 8$. More compactly, $p|\text{data} \sim Beta(6, 8)$.

b. LOTUS tells us that the expected value of an arbitrary function $f(x)$ is:

$$E[f(x)] = \frac{1}{B(a, b)} \int_0^1 f(x)\, x^{a-1}(1 - x)^{b-1}.$$

$f(x) = x^m(1 - x)^n$ blends beautifully with the PDF of our Beta posterior:

$$E[f(p)] = E[p^m(1 - p)^n] = \frac{1}{B(a, b)} \int_0^1 x^m(1 - x)^n x^{a-1}(1 - x)^{b-1}$$
$$= \frac{1}{B(a, b)} \int_0^1 x^{m+a-1}(1 - x)^{n+b-1} = \frac{B(a + m, b + n)}{B(a, b)}$$

c. The probability that a seventh game is played is $\binom{6}{3}p^3(1-p)^3$. So the expected value is $\binom{6}{3}p^3(1-p)^3$ would be:

$$\binom{6}{3}\frac{B(9,11)}{B(6,8)}.$$

2.4 Multiclass Bayes

In this problem we are going to explore how to write Naive Bayes for multiple output classes. We want to predict a single output variable Y which represents how a user feels about a book. Unlike in your homework, the output variable Y can take on one of the *four* values in the set {Like, Love, Haha, Sad}. We will base our predictions off of three binary feature variables $X_1, X_2,$ and X_3 which are indicators of the user's taste. All values $X_i \in \{0, 1\}$.

We have access to a dataset with 10,000 users. Each user in the dataset has a value for X_1, X_2, X_3 and Y. You can use a special query method **count** that returns the number of users in the dataset with the given *equality* constraints (and only equality constraints). Here are some example usages of **count**:

count$(X_1 = 1, Y = \text{Haha})$	returns the number of users where $X_1 = 1$ and $Y = \text{Haha}$.
count$(Y = \text{Love})$	returns the number of users where $Y = \text{Love}$.
count$(X_1 = 0, X_3 = 0)$	returns the number of users where $X_1 = 0$, and $X_3 = 0$.

You are given a new user with $X_1 = 1$, $X_2 = 1$, $X_3 = 0$. What is the best prediction for how the user will feel about the book (Y)? You may leave your answer in terms of an argmax function. You should explain how you would calculate all probabilities used in your expression. Use **Laplace estimation** when calculating probabilities.

We can make the Naive Bayes assumption of independence and simplify argmax of $P(Y|\mathbf{X})$ to get an expression for $\hat{Y}$, the predicted output value, and evaluate it using the provided **count** function.

$$\begin{aligned}
\hat{Y} &= \arg\max_y \frac{P(X_1 = 1, X_2 = 1, X_3 = 0|Y = y)P(Y = y)}{P(X_1 = 1, X_2 = 1, X_3 = 0)} \\
&= \arg\max_y P(X_1 = 1, X_2 = 1, X_3 = 0|Y = y)P(Y = y) \\
&= \arg\max_y P(X_1 = 1|Y = y)P(X_2 = 1|Y = y)P(X_3 = 0|Y = y)P(Y = y), \text{ where:}
\end{aligned}$$

$$\begin{aligned}
P(X_1 = 1|Y = y) &= [\mathbf{count}(X_1 = 1, Y = y) + 1]/[\mathbf{count}(Y = y) + 2] \\
P(X_2 = 1|Y = y) &= [\mathbf{count}(X_2 = 1, Y = y) + 1]/[\mathbf{count}(Y = y) + 2] \\
P(X_3 = 1|Y = y) &= [\mathbf{count}(X_3 = 1, Y = y) + 1]/[\mathbf{count}(Y = y) + 2] \\
P(X_1 = 0|Y = y) &= [\mathbf{count}(X_1 = 0, Y = y) + 1]/[\mathbf{count}(Y = y) + 2] \\
P(X_2 = 0|Y = y) &= [\mathbf{count}(X_2 = 0, Y = y) + 1]/[\mathbf{count}(Y = y) + 2] \\
P(X_3 = 0|Y = y) &= [\mathbf{count}(X_3 = 0, Y = y) + 1]/[\mathbf{count}(Y = y) + 2]
\end{aligned}$$

$$P(Y = y) = \textbf{count}(Y = y)/10,000$$

2.5 *Gaussian Naïve Bayes*

The version of Naïve Bayes that we used in class worked great when the feature values were all binary. If instead they are continuous, we are going to have to rethink how we estimate of the probability of the ith feature given the label, $P(X_i|Y)$. The ubiquitous solution is to make the *Gaussian Input Assumption* that:

$$\text{If } Y = 0, \text{ then } X_i \sim N(\mu_{i,0}, \sigma^2_{i,0})$$
$$\text{If } Y = 1, \text{ then } X_i \sim N(\mu_{i,1}, \sigma^2_{i,1})$$

For each feature, there are 4 parameters (mean and variance for both class labels). There is a final parameter, p, which is the estimate of $P(Y = 1)$. Assume that you have trained on data with **two** input features and have **already estimated** all 9 parameter values, including that $p = 0.6$:

Feature i	$\mu_{i,0}$	$\mu_{i,1}$	$\sigma^2_{i,0}$	$\sigma^2_{i,1}$
1	5	0	1	1
2	0	3	1	4

Write an inequality to predict whether $Y = 1$ for input $[X_1 = 5, X_2 = 3]$. Use the Naïve Bayes assumption and the Gaussian Input Assumption. Your expression should be in terms of the learned parameters (either using numbers or symbols is fine).

Fundamentally, you need to compute two probabilities: $P(X_1 = 5, X_2 = 3, Y = 0)$ and $P(X_1 = 5, X_2 = 3, Y = 1)$ and then predict $\hat{Y}$ to be whichever of 0 and 1 leads to a higher joint probability. However, when some of the input variables are continuous, the probability those continuous values takes on any specific value is 0.

However, you can still compute and compare probability **densities**, as with: $f(X_1 = 5, X_2 = 3, Y = 0)$ and $f(X_1 = 5, X_2 = 3, Y = 1)$. The Naïve Bayes assumption allows us to rewrite those densities as:

$$f(X_1 = 5, X_2 = 3, Y = 0) = f(X_1 = 5|Y = 0) \cdot f(X_2 = 3|Y = 0) \cdot P(Y = 0)$$

and

$$f(X_1 = 5, X_2 = 3, Y = 1) = f(X_1 = 5|Y = 1) \cdot f(X_2 = 3|Y = 1) \cdot P(Y = 1)$$

When the input variables are specifically guided by Gaussians with the learned paramaters presented above, we ultimately predict $\hat{Y} = 0$ if

$$\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{5-5}{1})^2} \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{3-0}{1})^2} \cdot 0.4 = \frac{1}{5\pi} e^{-\frac{9}{2}}$$

is greater than

$$\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}(\frac{5-0}{1})^2} \cdot \frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}(\frac{3-3}{4})^2} \cdot 0.6 = \frac{3}{10\pi}e^{-\frac{25}{2}}$$

and otherwise predict $\hat{Y} = 1$.