

Beta and Parameter Estimation

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

- 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?
- 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?
- 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?

1.2 Beta Sum

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

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

Note the variance of a Beta:

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

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!

$$\sigma(z) = \frac{1}{1 + e^{-z}}$$

$$\frac{\partial \sigma(z)}{\partial z} = \sigma(z)[1 - \sigma(z)]$$

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

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 $f(\Lambda = \lambda) = K \cdot \lambda \cdot e^{-\lambda}$. 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.

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 $p \sim \text{Beta}(1, 1)$.

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.

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

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 $\{\text{Like, Love, Haha, Sad}\}$. 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.

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 i th feature given the label, $P(X_i|Y)$. The ubiquitous solution is to make the *Gaussian Input Assumption* that:

$$\begin{aligned} \text{If } Y = 0, \text{ then } X_i &\sim N(\mu_{i,0}, \sigma_{i,0}^2) \\ \text{If } Y = 1, \text{ then } X_i &\sim N(\mu_{i,1}, \sigma_{i,1}^2) \end{aligned}$$

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	$\mu_{i,0}$	$\mu_{i,1}$	$\sigma_{i,0}^2$	$\sigma_{i,1}^2$
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).