Maximum A Posteriori and Naïve Bayes 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 Maximum A Posteriori

a. Intuitively, what is MAP? What problem is it trying to solve? How does it differ from MLE?

b. Given a 6-sided die (possibly unfair), you roll the die $N$ times and observe the counts for each of the 6 outcomes as $n_1, \ldots, n_6$. What is the maximum a posteriori estimate of this distribution, using Laplace smoothing? Recall that the die rolls themselves follow a multinomial distribution.

1.2 Naive Bayes Review

Recall the classification setting: we have data vectors of the form $X = (X_1, \ldots, X_m)$ and we want to predict a label $Y \in \{0, 1\}$.

a. Recall in Naive Bayes, given a data point $x$, we compute $P(Y = 1|X = x)$ and predict $Y = 1$ provided this quantity is $\geq 0.5$, and otherwise we predict $Y = 0$. Decompose $P(Y = 1|X = x)$ into smaller terms, and state where the Naive Bayes assumption is used.

b. Suppose we are given example vectors with labels provided. Give a formula to estimate (using maximum likelihood) each quantity $P(X_i = x_i|Y = y)$ above, for $i \in \{1, \ldots, m\}$ and $y \in \{0, 1\}$. You can assume there is a function count which takes in any number of boolean conditions and returns a count over the data of the number of examples in which they are true. For example, $\text{count}(X_3 = 2, X_5 = 7)$ returns the number of examples where $X_3 = 2$ and $X_5 = 7$.

2 Problems

2.1 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 $\lambda$ 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 $f(\Lambda = \lambda) = K \cdot \lambda^{\frac{1}{2}} \cdot e^{-\lambda^{2}}$. What is the MLE estimate? What is inference of $\lambda$ given the observation? What is the Maximum a Posteriori (MAP) estimate of $\lambda$? Through your process try to identify what is a point-estimate, and what is a distribution.
2.2 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 = Haha)` returns the number of users where $X_1 = 1$ and $Y = Haha$.
- `count(Y = Love)` returns the number of users where $Y = Love$.
- `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.

2.3 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:

- If $Y = 0$, then $X_i \sim N(\mu_{i,0}, \sigma^2_{i,0})$
- If $Y = 1$, 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 $p = 0.6$:

<table>
<thead>
<tr>
<th>Feature $i$</th>
<th>$\mu_{i,0}$</th>
<th>$\mu_{i,1}$</th>
<th>$\sigma^2_{i,0}$</th>
<th>$\sigma^2_{i,1}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>5</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>3</td>
<td>1</td>
<td>4</td>
</tr>
</tbody>
</table>

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).
3 Ethics and Beta Distribution

While there won’t be any ethics material on the final exam, we’re including a problem that will not only exercise some probability, but hopefully provoke you to think about the impact that probability- and data-driven decisions have on society.

The Economist used a beta distribution to forecast results for the 2020 U.S. presidential election.\(^1\)

![Three steps of Bayesian inference](image)

Figure 1: Updated prediction of Democratic vote share is "Posterior" prediction.

1. Why is the beta distribution appropriate for modeling a presidential election?

2. Read the polling report published by The Economist. What should be considered when using this model and releasing its election predictions?

---