Maximum A Posteriori and Naïve Bayes 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 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.

a. From the course notes: The paradigm of MAP is that we should choose the value for our parameters that is the most likely given the data. At first blush this might seem the same as MLE; however, remember that MLE chooses the value of parameters that makes the data most likely. One of the disadvantages of MLE is that it best explains data we have seen and makes no attempt to generalize to unseen data. In MAP, we incorporate prior belief about our parameters, and then we update our posterior belief of the parameters based on the data we have seen.

b. Using a prior which represents one imagined observation of each outcome is called “Laplace smoothing” and it guarantees that none of your probabilities are 0 or 1. The Laplace estimate for a Multinomial RV is \( p_i = \frac{n_i + 1}{N + 6} \) for \( i = 1, \ldots, 6 \).

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, 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 \( f(\Lambda = \lambda) = K \cdot \lambda^4 \cdot e^{-\lambda/4} \). 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.

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

\[
L(\lambda) = \frac{\lambda^4 \cdot e^{-\lambda}}{4!}
\]

\[
LL(\lambda) = 4 \log(\lambda) - \lambda - \log(4!)
\]

\[
\frac{\partial LL}{\partial \lambda} = \frac{4}{\lambda} - 1
\]

Set \( \frac{\partial LL}{\partial \lambda} \) to 0 and solve for \( \lambda \).

\[
\lambda = 4
\]

Inference of \( \lambda \) given the observation:

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

MAP estimate of \( \lambda \): we find the \( \lambda \) that maximizes the inference given the observation, i.e.
we want to solve:

\[
\arg\max_{\lambda} \quad f(\lambda | X = 4) = \arg\max_{\lambda} \frac{P(X = 4 | \lambda) \cdot f(\lambda)}{P(X = 4)} = \arg\max_{\lambda} P(X = 4 | \lambda) \cdot f(\lambda) = \arg\max_{\lambda} \frac{\lambda^4 \cdot e^{-\lambda}}{4!} \cdot K \cdot \lambda \cdot e^{-\frac{\lambda}{2}}
\]

Take log.

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

Differentiate with respect to \(\lambda\), set to 0 and solve.

\[
\frac{5}{\lambda} - 1 - \frac{1}{2} = 0
\]

\[
\lambda = \frac{10}{3}
\]

### 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 \texttt{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 \texttt{count}:

\[
\text{count}(X_1 = 1, Y = \text{Haha}) \quad \text{return the number of users where } X_1 = 1 \text{ and } Y = \text{Haha.}
\]

\[
\text{count}(Y = \text{Love}) \quad \text{return the number of users where } Y = \text{Love.}
\]

\[
\text{count}(X_1 = 0, X_3 = 0) \quad \text{return the number of users where } X_1 = 0, \text{ and } X_3 = 0.
\]

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 \textit{Laplace estimation} when calculating probabilities.

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

\[
\hat{Y} = \arg \max_y \frac{P(X_1 = 1, X_2 = 1, X_3 = 0 | Y = y) P(Y = y)}{P(X_1 = 1, X_2 = 1, X_3 = 0)}
\]

\[
= \arg \max_y P(X_1 = 1, X_2 = 1, X_3 = 0 | Y = y) P(Y = y)
\]

\[
= \arg \max_y P(X_1 = 1 | Y = y) P(X_2 = 1 | Y = y) P(X_3 = 0 | Y = y) P(Y = y), \text{ where:}
\]

\[
P(X_1 = 1 | Y = y) = \frac{\text{count}(X_1 = 1, Y = y) + 1}{\text{count}(Y = y) + 2}
\]

\[
P(X_2 = 1 | Y = y) = \frac{\text{count}(X_2 = 1, Y = y) + 1}{\text{count}(Y = y) + 2}
\]

\[
P(X_3 = 1 | Y = y) = \frac{\text{count}(X_3 = 1, Y = y) + 1}{\text{count}(Y = y) + 2}
\]

\[
P(X_1 = 0 | Y = y) = \frac{\text{count}(X_1 = 0, Y = y) + 1}{\text{count}(Y = y) + 2}
\]

\[
P(X_2 = 0 | Y = y) = \frac{\text{count}(X_2 = 0, Y = y) + 1}{\text{count}(Y = y) + 2}
\]

\[
P(X_3 = 0 | Y = y) = \frac{\text{count}(X_3 = 0, Y = y) + 1}{\text{count}(Y = y) + 2}
\]

\[
P(Y = y) = \frac{\text{count}(Y = y)}{10,000}
\]

### 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).
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 parameters presented above, we ultimately predict \( \hat{Y} = 0 \) if

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

is greater than

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

and otherwise predict \( \hat{Y} = 1 \).

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.\(^1\)

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?

1. Several features of the beta distribution map onto election modeling:
   
   - The beta requires number of successes and number of failures as parameters. These are easy to acquire for elections since they are the counts of the target candidate polling positively vs. negatively.
   
   - The beta distribution can be used to model quantities representing fractions or percentages, since it is a continuous random variable with a support of \([0, 1]\). Election outcomes are typically reported in terms of percentage vote share, which naturally lends itself to a beta.
   
   - Election results are highly variable and betas allow us to incorporate uncertainty!
   
   - Election predictions require priors and new data, especially as the election day approaches. In the Economist’s model, the expected distribution of potential vote shares in each state was used as the prior, and the state polls that trickled in during the course of the campaign were the ”new data”.

2. Possible answers:

**Limitations**

- The beta approach is bad at modeling multiparty systems because success/failure can’t split up vote shares per candidate unless it is a two-party system.
- Random drift causes uncertainty around the current polling average.
- In states that are heavily polled late in the race, the model will pay little attention to its prior forecast; conversely, it will emphasise the prior early in the race or in thinly-polled states.
• Different methods of turnout projection can produce a bias.

• *Partisan non-response bias*: The probability that a poll respondent will agree to participate in a survey varies in response to media coverage. When there is unusually bad news about a candidate, their supporters are not in the mood to tell pollsters what they think – even though their ultimate voting intention has not changed. This causes the other candidate’s vote share to be over-represented in polls.

**Ethical Considerations**

• Poll results influence how people will vote. When the public believes a candidate is extremely likely to win, some people are less likely to vote. This is why some analysts think election polls should be covered using margins of error rather than speculative win probabilities.

• Poll results act as a feedback mechanism that affect parties’ policy choices.