Lisa Yan
CS 109

Section #8
June 3-5, 2020

Section 8 Solutions

---

1. **MLE and MAP**

   To start, let us look briefly at how we calculate the parameters for Maximum Likelihood and Maximum a Posteriori Estimation. As usual, we let f be some probability distribution function, and we let g some prior probability distribution function.

$$\theta_{MLE} = \arg\max_{\theta} \prod_{i=1}^{n} f(X_i|\theta)$$

$$\theta_{MAP} = \arg\max_{\theta} \prod_{i=1}^{n} f(X_i|\theta)g(\theta)$$

   These look suspiciously similar, which begs the question: what is the difference between MLE and MAP?

   (a) The only difference between the MLE and MAP updates is the introduction of an additional equation $g(\theta)$. In words, what are we doing when we add $g(\theta)$?

   (b) Write the log likelihood function $LL(\theta)$ for both MLE and MAP, being sure to use the properties of log to simplify your work.

   (c) Let us now look at $\theta = \arg\max LL(\theta)$ for θ_{MAP} and θ_{MLE}. Notably, there is one additional term for θ_{MAP} — this is not surprising, given our earlier observation. What does that additional term need to equal so that θ_{MAP} and θ_{MLE} are equal?

   (d) If we do the above so that $\theta_{MAP} = \theta_{MLE}$, what kind of probability distribution function should $g(\theta)$ be equal to? And what does that say about the relationship between θ_{MAP} and θ_{MLE}?

   (a) When we add the prior $g(\theta)$ we are essentially giving our model additional information to make predictions. Of course, there is a question of what information makes sense and what information is computationally feasible. In other words, what calculations can we give our model, based on the data, that will be helpful?

(b) The log likelihood functions are as follows.

$$LL(\theta_{MLE}) = \sum_{i=1}^{n} \log f(X_i|\theta)$$

$$LL(\theta_{MAP}) = \sum_{i=1}^{n} \log(f(X_i|\theta)g(\theta))$$

$$= \sum_{i=1}^{n} \log f(X_i|\theta) + \log g(\theta)$$

As a practical piece of advice, if you do not know where to start when making calculations with log likelihood, a good place to start is to apply the definition, and then separate the terms as much as possible. It makes things clearer!

(c) θ_{MAP} and θ_{MLE} are given below.

$$\theta_{MAP} = \arg\max_{\theta} \sum_{i=1}^{n} \log f(X_i|\theta) + \log g(\theta)$$

$$\theta_{MLE} = \arg\max_{\theta} \sum_{i=1}^{n} \log f(X_i|\theta)$$

Intuitively, we want $\log g(\theta)$ to not matter. The key insight here is that, when we optimize, constants do not matter. Thus if $g(\theta)$ were a constant, then we could drop $\log g(\theta)$ from the θ_{MAP} calculation, thus making $\theta_{MAP} = \theta_{MLE}$.

(d) We saw that $g(\theta)$ needs to be constant, which means that $g(\theta)$ needs to be the probability distribution function for a uniform distribution. Intuitively, if g is dependent on θ in any way, when we take the gradient with respect to θ, the term corresponding to $\log g(\theta)$ would not go away. However, this is equivalent to not having a prior at all — which is exactly what we have with MLE. (Make sure you think about why a uniform prior would be the same as having $g(\theta) = 1$ for all θ. One way to think about this formally is to take the next step and calculate the gradient.)

To conclude, we have proven that MLE is a special case of MAP, where we have a uniform prior, i.e. no additional information is given to the model.

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$(X_1 = 1, Y = \text{Haha})$	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.

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

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

you don't need to use MAP to estimate $P(Y)$: $P(Y = y) = \mathbf{count}(Y = y)/10,000$

3. **Vision Test**

You decide that the vision tests given by eye doctors would be more precise if we used an approach inspired by logistic regression. 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.

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.

Formula reference for Logistic Regression:

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

$$LL(\theta) = \sum_{i=0}^{n} y^{(i)} \log \sigma(\theta^T \mathbf{x}^{(i)}) + (1 - y^{(i)}) \log \left[1 - \sigma(\theta^T \mathbf{x}^{(i)})\right]$$

$$\frac{\partial LL(\theta)}{\partial \theta_j} = \sum_{i=0}^{n} \left[y^{(i)} - \sigma(\theta^T \mathbf{x}^{(i)}) \right] x_j^{(i)}$$

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.

We first write the log likelihood. Note that, below, we write p for $p^{(i)} = \sigma(\theta^T \mathbf{x}^{(i)})$, to make our notation cleaner.

$$L(\theta) = \prod_{i=1}^{20} p^{y^{(i)}} (1-p)^{1-y^{(i)}}$$

$$LL(\theta) = \sum_{i=1}^{20} (y^{(i)} \log(p) + (1 - y^{(i)}) \log(1-p))$$

Then we find the derivative of log likelihood with respect to θ for one datapoint:

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

We can calculate both the smaller partial derivatives independently:

$$\frac{\partial LL}{\partial p} = \frac{y^{(i)}}{p} - \frac{1 - y^{(i)}}{1-p} \text{ and } \frac{\partial p}{\partial \theta} = p[1-p]$$

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:

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

The Most Important Features

Let's explore saliency, a measure of how important a feature is for classification. We define the saliency of the ith input feature for a given example $(\mathbf{x}, y)$ to be the absolute value of the partial derivative of the log likelihood of the sample prediction, with respect to that input feature $\left|\frac{\partial LL}{\partial x_i}\right|$. In the images below, we show both input images and the corresponding saliency of the input features (in this case, input features are pixels):

First consider a trained logistic regression classifier with weights θ. Like the logistic regression classifier that you wrote in your homework it predicts binary class labels. In this question we allow the values of $\mathbf{x}$ to be real numbers, which doesn't change the algorithm (neither training nor testing).

a. What is the Log Likelihood of a single training example $(\mathbf{x}, y)$ for a logistic regression classifier?

$$LL(\theta) = y \cdot \log \sigma\Big(\theta^T \cdot \mathbf{x}\Big) + \Big(1 - y\Big)\log \Big[1 - \sigma\Big(\theta^T \cdot \mathbf{x}\Big)\Big]$$

b. Calculate is the saliency of a single feature (x_i) in a training example $(\mathbf{x}, y)$.

We can calculate the saliency for a single feature as follows.

$$\begin{aligned}
LL(\theta) &= y \log z + \Big(1 - y\Big) \log \Big(1 - z\Big) && \text{where } z = \sigma\Big(\theta^T \cdot \mathbf{x}\Big) \\
\frac{\partial LL}{\partial x_i} &= \frac{\partial LL}{\partial z} \cdot \frac{\partial z}{\partial x_i} && \text{chain rule} \\
&= \Big(\frac{y}{z} - \frac{1 - y}{1 - z}\Big) \cdot \Big(z(1 - z)\theta_i\Big) && \text{partial derivatives} \\
\text{saliency} &= \Big| \Big(\frac{y}{z} - \frac{1 - y}{1 - z}\Big) z(1 - z)\theta_i \Big|
\end{aligned}$$

Show that the ratio of saliency for features i and j is the ratio of the absolute value of their weights $\frac{|\theta_i|}{|\theta_j|}$.

We can take the ratio as follows using our expression above.

$$\text{saliency for feature } i, S_i = \left| \left(\frac{y}{z} - \frac{1-y}{1-z}\right) z(1-z)\theta_i \right| \text{, and same for } S_j$$

$$\frac{S_i}{S_j} = \frac{\left| \left(\frac{y}{z} - \frac{1-y}{1-z}\right) z(1-z)\theta_i \right|}{\left| \left(\frac{y}{z} - \frac{1-y}{1-z}\right) z(1-z)\theta_i \right|} = \frac{S_i}{S_j} = \frac{\left| \theta_i \right|}{\left| \theta_j \right|} \text{ by elimination}$$