Lisa Yan and Jerry Cain
CS109

Section #8
November 10-11, 2020

# Section #8 Solutions

Based on the work of many CS109 staffs

## 1 Warmups

### 1.1 Parameters and MLE

Suppose $x_1, \ldots, x_n$ are i.i.d. (independent and identically distributed) values sampled from some distribution with density function $f(x|\theta)$, where $\theta$ is unknown. Recall that the likelihood of the data is

$$L(\theta) = f(x_1, x_2, \ldots, x_n|\theta) = \prod_{i=1}^{n} f(x_i|\theta)$$

Recall we solve an optimization problem to find $\hat{\theta}$ which maximizes $L(\theta)$, i.e., $\hat{\theta} = \arg\max_\theta L(\theta)$.

1. Write an expression for the log-likelihood, $LL(\theta) = \log L(\theta)$.
2. Why can we optimize $LL(\theta)$ rather than $L(\theta)$?
3. Why do we optimize $LL(\theta)$ rather than $L(\theta)$?

1. $LL(\theta) = \sum_{i=1}^{n} \log f(x_i|\theta)$
2. The logarithm (for bases > 1) is a monotonically increasing function. This means that if $f(a) > f(b)$, then $\log(f(a)) > \log(f(b))$, so the arg max function is preserved across a logarithmic transformation, i.e., $\arg\max L(\theta) = \arg\max LL(\theta)$.
3. Logs turn products into sums, which makes taking the derivative for maximization or minimization much simpler.

### 1.2 Beta

1. 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?
2. Suppose you have a coin which you believe is fair, with “strength” $\alpha$. That is, pretend you’ve seen $\alpha$ heads and $\alpha$ tails. How can you model this belief as a Beta distribution?
3. 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. $\text{Beta}(1, 1)$ is a uniform prior, meaning that prior to seeing the experiment, all probabilities of heads are equally likely.
2. $\text{Beta}(\alpha + 1, \alpha + 1)$. This is our prior belief about the distribution.
3. $\text{Beta}(\alpha + 9, \alpha + 3)$

### 1.3 Maximum A Posteriori

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

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

1. 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.

2. 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 $i = 1, ..., 6$.

### 1.4 Naive Bayes

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

1. 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.

2. 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 $i \in \{1, \ldots, d\}$ 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 $X_5 = 7$.

1.

$$\begin{aligned} P(Y = 1|X = x) &= \frac{P(Y = 1)P(X = x|Y = 1)}{P(Y = 1)P(X = x|Y = 1) + P(Y = 0)P(X = x|Y = 0)} && \text{(Bayes+LTP)} \\ &= \frac{P(Y = 1)\prod_{i=1}^{d} P(X_i = x_i|Y = 1)}{P(Y = 1)\prod_{i=1}^{d} P(X_i = x_i|Y = 1) + P(Y = 0)\prod_{i=1}^{d} P(X_i = x_i|Y = 0)} && \text{(NB Assumption)} \end{aligned}$$

2. $P(X_i = x_i|Y = y) = \dfrac{\text{count}(X_i = x_i, Y = y)}{\text{count}(Y = y)}$

### *1.5 Gradient Ascent and Linear Regression*

Let $f : \mathbb{R}^n \to \mathbb{R}$ be a function which maps vectors $x \in \mathbb{R}^n$ to scalars $f(x) \in \mathbb{R}$.

1. What is the gradient ascent update step, with learning rate $\eta$?
2. Intuitively, what problem is gradient ascent trying to solve numerically?
3. What are some tradeoffs between a high and low learning rate ($\eta$)?

1. $x \leftarrow x + \eta \nabla f(x)$
2. We are attempting to numerically find the value of $x$ that maximizes $f(x)$ by incrementally taking small steps in the direction of steepest ascent (according the the derivative).
3. A small learning rate might require more steps until convergence, while a large learning rate might overshoot and miss the absolute maximum.

## 2 Problems

### *2.1 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.

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.

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

$$P(X_1 = 1|Y = y) = \frac{\textbf{count}(X_1 = 1, Y = y) + 1}{\textbf{count}(Y = y) + 2}$$
$$P(X_2 = 1|Y = y) = \frac{\textbf{count}(X_2 = 1, Y = y) + 1}{\textbf{count}(Y = y) + 2}$$
$$P(X_3 = 1|Y = y) = \frac{\textbf{count}(X_3 = 1, Y = y) + 1}{\textbf{count}(Y = y) + 2}$$
$$P(X_1 = 0|Y = y) = \frac{\textbf{count}(X_1 = 0, Y = y) + 1}{\textbf{count}(Y = y) + 2}$$
$$P(X_2 = 0|Y = y) = \frac{\textbf{count}(X_2 = 0, Y = y) + 1}{\textbf{count}(Y = y) + 2}$$
$$P(X_3 = 0|Y = y) = \frac{\textbf{count}(X_3 = 0, Y = y) + 1}{\textbf{count}(Y = y) + 2}$$

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