Lisa Yan and Jerry Cain
CS109

Section #8

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 θ 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 $L(\theta)$?

1.2 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" α. That is, pretend you've seen α heads and α 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?

1.3 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.

1.4 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 ≥ 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 $X_5 = 7$.

1.5 *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 η?
2. Intuitively, what problem is gradient ascent trying to solve numerically?
3. What are some tradeoffs between a high and low learning rate (η)?

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