Homework 5 Fall 2024

04650 Mathematical Foundations of Machine Learning (CMU-Africa)

Release Date: November 12th, 2024
DUE: **November 22nd, 2024, 11:59 PM CAT**

- **Collaboration policy:** You can discuss HW problems with other students, but the work you submit must be written in your own words and not copied from anywhere else. However, do write down (at the top of the first page of your HW solutions) the names of all the people with whom you discussed this HW assignment. We strongly encourage you to type your solution using LaTeX, but this is not required

- **Submitting your work:** Assignment will be submitted using Gradescope. You will submit your notebook file (Homework5.ipynb) to "Homework 5 - Code" and your PDF solutions to "Homework 5 - Solutions"

- **Getting Help**: Please use office hours and Piazza to ask any questions related to this assigments.

- **Refrain from using ChatGPT**: These problems are easy and can be easiliy solved with the use of generative AI's like ChatGPT. However, you will not learn by using these tools. Not learning will impact your performance in other courses which require this knowledge like introduction to deep learning and introduction to machine learning for engineers. Ask questions on Piazza about anything you don't understand, the TA's and instructors will respond to you as fast as possible. Moreover, you will not have access to internet during the exam.

Learning Objectives

- Probability: Learn how to estimate paremeters using MLE and MAP. Learn how to use probabilistic reasoning to build a simple classifier.

1 Probability (70 points)

1.1 MAP (30 Points)

In this question, we will use the Geometric distribution (the distribution of independent Bernoulli trials until the first success) to model an airport baggage claim. Let B be the number of your bag (meaning you watched $B-1$ bags pass before yours arrived). The probability distribution of B is given by

$$Pr(B=b) = (1-p)^{b-1}p \quad \text{for } b = 1, 2, \ldots$$

where $0 < p \leq 1$ is the parameter of the distribution.

We now will use a Beta prior to get a MAP estimate of the distribution parameter p. Specifically,

$$Pr(p|\alpha, \beta) = p^{\alpha-1}(1-p)^{\beta-1}$$

where $\alpha > 0$ and $\beta > 0$ are fixed, given constants.

1. Assume that over n trips, you noted your bag's number as $b_1, b_2, \ldots, b_n$ respectively, and the number of your bag on each of these trips is independent. Derive the log posterior probability of recording these numbers $b_1, b_2, \ldots, b_n$:

   $$\log Pr(p|b_1, \ldots, b_n)$$

   (10 Points)

   Hint: you can use Bayes' Rule.

2. Use your answer in part (1) to derive the maximum-a-posterior estimate (MAP) of the parameter:

   $$\hat{p} = \arg\max_{0<p\leq 1} \log Pr(p|b_1, \ldots, b_n)$$

   (10 Points)

3. Suppose that $n = 5$ and the values of $b_1, b_2, \ldots, b_n$ are 10, 9, 5, 28, 7. Also, $\alpha = 14$ and $\beta = 590$. Using your answer in part (2), what is the MAP estimate of the p parameter for this data and prior? (10 Points)

1.2 Naive Bayes (20 Points)

You are asked to build a Naïve Bayes classifier using the training dataset in Table 1, where each instance is assigned to one out of 3 classes ("healthy" (H), "influenza" (I), or "salmonella poisoning" (S)).

Training	Fever (F)	Vomiting (V)	Diarrhea (D)	Classification
D1	no	no	no	Healthy (H)
D2	average	no	no	Influenza (I)
D3	high	no	no	Influenza (I)
D4	high	yes	yes	Salmonella poisoning (S)
D5	average	no	yes	Salmonella poisoning (S)

Table 1: Health Classification based on Symptoms

1. Using the Naïve Bayes model, find the prior probabilities $P(H)$, $P(I)$, and $P(S)$ given the training data above. (4 points)

2. Fill in Table 2 below to finish building your Naïve Bayes classifier. Use Laplace smoothing with parameter $\alpha = 2$ for your conditional probability estimates. (8 points)

3. Apply your Naïve Bayes Classifier to a person who is vomiting but has no fever or diarrhea. Determine the probabilities of this person being healthy, suffering from influenza, and salmonella poisoning. (8 points)

$P(X\|Y)$	$Y = H$	$Y = I$	$Y = S$
$X =$ (high fever)			
$X =$ (average fever)			
$X = V$			
$X = D$			

Table 2: Conditional Probability Table for $P(X|Y)$

1.3 MLE (20 points)

- A variable y is called a count if it only takes non-negative integer values, i.e., $y \in \{0, 1, 2, \ldots\}$. A common distribution for handling such variables is the Poisson distribution, which has the following form:

$$\text{Poisson}(y; \lambda) = \frac{e^{-\lambda}\lambda^y}{y!}, \quad y \in \{0, 1, 2, \ldots\}$$

  where $\lambda > 0$ is some known constant that characterizes the Poisson distribution.

  Given independent identically distributed (iid) observations $\{y_i\}_{i=1}^N$, $y_i \in \{0, 1, 2, \ldots\}$, calculate the MLE estimate of λ. (8 Points)

- In this problem, you will examine the task of estimating the probability density of the maximum height obtained by waves in the ocean. Scientists have recorded the maximum wave height on n days, obtaining samples $x_1, x_2, \ldots, x_n \in \mathbb{R}$. It is known that these are i.i.d. random variables following the Rayleigh distribution with parameter θ. Consider the following probability density function for the Rayleigh distribution:

$$f_X(x; \theta) = \frac{x}{\theta^2} \exp\left(-\frac{x^2}{2\theta^2}\right)$$

  The likelihood function for your estimate is then $L_X(\theta) = f_X(x_1, \ldots, x_n|\theta)$. Your task is to estimate $\hat{\theta}_{\text{MLE}}$, the maximum likelihood estimate of θ. (12 Points)

2 Text classification (30 Points)

Consider a text classification problem. In this case, you will try to classify text as either spam or ham. To do this, you will apply concepts of Likelihood, prior, and posterior given a dataset comprising pairs of text and labels. There are two types of labels: 1 (spam) and 0 (ham). Your goal is to create a simple classifier that, when given, determines if the text is spam or ham. You have been provided with the starter code and the data

1. Find the priors. What are the priors in this distribution? i.e find $P(ham)$ and $P(spam)$
2. Find the likelihoods for each word. For each word in the dataset, find the likelihood that the word is in spam and ham. This will represent the conditional probability $P(w|\text{spam})$ and $P(w|\text{ham})$ for w where $w \in V$. V is the vocabulary of the dataset.
3. Define a function that, when given a text sequence, returns the probability of the text being in spam. I.e., it returns $P(\text{spam}|\text{text})$. Note that this function calculates the likelihood using the Bayes rule. Do the same for ham.
4. Perform inference, i.e., given a string of text, determine if it is ham or spam based on the posterior probabilities calculated from the previous steps. Your function will determine the posterior probability of your text being in ham and spam and classify it as being the larger of the two.
5. Evaluate the data based on your test set and report the accuracy of your classifier. Your accuracy must be greater than 85%.