CSCI 447/547: Homework 1

Due: 09/23/2021

Please submit your responses to the following questions on Moodle. You may work with a partner, and are encouraged to develop solution strategies together, but your actual submitted responses must be entirely your own. Using the internet is okay, but try to solve these on your own first, and be sure not to plagiarize! The objective of these exercises is not to get the correct answer, but rather to demonstrate that you understand the answer and the question to which you are responding. As such, please use complete sentences and develop your arguments fully. In particular, don't leave it to me to try to divine your meaning!

Problem 1: Legal reasoning

(Source: Murphy) Suppose a crime has been committed. Blood is found at the scene for which there is no innocent explanation. It is of a type which is present in 1% of the population.

1. The prosecutor claims: "There is a 1% chance that the defendant would have the crime blood type if he were innocent. Thus there is a 99% chance that he guilty". This is known as the prosecutor's fallacy. What is wrong with this argument?
2. The defender claims: "The crime occurred in a city of 800,000 people. The blood type would be found in approximately 8000 people. The evidence has provided a probability of just 1 in 8000 that the defendant is guilty, and thus has no relevance." This is known as the defender's fallacy. What is wrong with this argument?

Problem 2: Poisson distribution MLE

The Poisson distribution is a useful model when we would like to model a count as a random variable. For example, the number of letters a person gets in a day is an example of something that could be Poisson-distributed (other examples: the number of deer that appear on a game camera in a day, the number of meteorites to strike the earth in a year). The probability mass function of a Poisson distribution is

$$P(X = k) = \frac{\lambda^k \mathrm{e}^{-\lambda}}{k!},$$

where k is the number of times an event occurs, and λ is the rate parameter. Given some dataset $X_1, X_2, \ldots, X_m$, derive the maximum likelihood estimator for λ. Evaluate your estimator for the data set $[1, 2, 2, 3]$. (HINT: the procedure here is just like for the Bernoulli distribution: logarithm, derivative, set to zero, solve).

Problem 3: Naive Bayes

When utilizing naive Bayes, we have relied upon maximum likelihood estimation to estimate the parameters of categorical distributions over words conditioned on a class, i.e.

$$P(X = x|C = c)$$

for x in a vocabulary and c in a set of classes. Describe what happens if, having adopted this strategy, we attempt to apply naive Bayes to a message that contains a word that did not appear in the training corpus. (FOR GRAD STUDENTS) Derive a strategy to ameliorate this (potential) problem and defend your reasoning for why it's a good solution.

Problem 4: Monty Hall (FOR GRAD STUDENTS)

(Source: Murphy) On a game show, a contestant is told the rules as follows:

> There are three doors, labelled 1, 2, 3. A single prize has been hidden behind one of them. You get to select one door. Initially your chosen door will not be opened. Instead, the game show host will open one of the other two doors, and she will do so in such a way as not to reveal the prize. For example, if you first choose door 1, she will then open one of doors 2 and 3, and it is guaranteed that she will choose which one to open so that the prize will not be revealed.
>
> At this point, you will be given a fresh choice of door: you can either stick with your first choice, or you can switch to the other closed door. All the doors will then be opened and you will receive whatever is behind your final choice of door.

Imagine that the contestant chooses door 1 first; then the host opens door 3, revealing nothing behind the door, as promised. Should the contestant

(a) stick with door 1

(b) switch to door 2

(c) does it make no difference?

You may assume that initially, the prize is equally likely to be behind any of the 3 doors. (HINT: use Bayes rule.)