CSCI 5090/7090 Machine Learning, Spring 2018: Homework 1

Due: Monday, February 12^{rd}, beginning of class

Instructions There are 2 written questions and a third coding question on this assignment. You need to turn in the **hardcopy** of your answers (excluding the programming question) at the beginning of the class. Also, submit both your written answers (as a pdf) and your implementation to the Dropbox Folder on Folio. Please show your work for all questions.

Your written answers may be typed. You're welcome to type your solutions in LaTex if you know how. If you don't know LaTex but want to type, there a number of markdown editors with real-time preview and equation editing. Here are two: `https://stackedit.io/`, `http://marxi.co/`. Writing your solutions by hand is also fine as long as they're neat.

You are welcome to use any Python libraries for data munging, visualization, and numerical linear algebra. Examples includes Numpy, Pandas, and Matplotlib. You may NOT, however, use any Python machine learning libraries such as Scikit-Learn or TensorFlow. If in doubt, email the instructors.

1 Probability and Bayes' Rule [25 points]

1. Assume the probability of a certain disease is 0.01. The probability of test positive given that a person is infected with the disease is 0.95 and the probability of test positive given the person is not infected with the disease is 0.05.
   (a) Calculate the probability of test positive. [5pt]
   (b) Use Bayes Rule to calculate the probability of being infected with the disease given that the test is positive. [5pt]
2. A group of students were classified based on whether they are senior or junior and whether they are taking CSE446 or not. The folowing data was obtained.

	Junior	Senior
taking CSE446	23	34
no CSE446	41	53

   Suppose a student was randomly chosen from the group. Let J be the event that the student is junior, S be the event that the student is senior, C be the event that the student is taking CSE446, and $\bar{C}$ be the event that the student is not taking CSE446. Calculate the following probabilities. Show your work.
   (a) (5 points) $P(C, S)$
   (b) (5 points) $P(C|S)$
   (c) (5 points) $P(\bar{C}|J)$

2 MLE [20 points]

2.1 The Poisson distribution [12 points]

You're a Seahawks fan, and the team is six weeks into its season. The number touchdowns scored in each game so far are given below:

$$[1, 3, 3, 0, 1, 5].$$

Let's call these scores $x_1, \ldots, x_6$. Based on your data, you'd like to build a model to understand how many touchdowns the Seahaws are likely to score in their next game. You decide to model the number of touchdowns scored per game using a *Poisson distribution*. The Poisson distribution with parameter λ assigns every non-negative integer $x = 0, 1, 2, \ldots$ a probability given by

$$\text{Poi}(x|\lambda) = e^{-\lambda}\frac{\lambda^x}{x!}.$$

So, for example, if $\lambda = 1.5$, then the probability that the Seahawks score 2 touchdowns in their next game is $e^{-1.5} \times \frac{1.5^2}{2!} \approx 0.25$. To check your understanding of the Poisson, make sure you have a sense of whether raising λ will mean more touchdowns in general, or fewer.

1. *(8 points)* Derive an expression for the maximum-likelihood estimate of the parameter λ governing the Poisson distribution, in terms of your touchdown counts $x_1, \ldots, x_6$. (Hint: remember that the log of the likelihood has the same maximum as the likelihood function itself.)
2. *(4 points)* Given the touchdown counts, what is your numerical estimate of λ?

2.2 The Uniform Distribution [8 points]

Given a set of i.i.d samples $X_1, X_2, \cdots, X_n$ with uniform distributions $Uniform(0, \theta)$, find the maximum likelihood estimator of θ.

(a) *(4 points)* Write down the likelihood function

(b) *(4 points)* Find the maximum likelihood estimator

3 Programming Question [55 points]

In this assignment you will be implementing the C4.5 decision tree algorithm and running it on real emails to train a spam filter. All resources including all the instructions and skeleton of the code are provided in the file `https://sci2lab.github.io/mehdi/teaching/x90/hw/1/hw1_dt.zip`.