CS 3501/ECE 3502: Foundations of Data Analysis **Due Feb. 26**

Homework 2

Instructions: Submit a single Jupyter notebook (.ipynb) of your work to Collab by 11:59pm on the due date. All code should be written in Python. **Be sure to show all the work involved in deriving your answers! If you just give a final answer without explanation, you may not receive credit for that question.**

You may discuss the concepts with your classmates, but write up the answers entirely on your own. Do not look at another student's answers, do not use answers from the internet, and do not show your answers to anyone.

1. **Maximum Likelihood Estimation (MLE).** You are maintaining a web server and need a probability model for the traffic on it. You've settled on a model for the time in between server requests with the following probability density function (pdf):

$$p(x;\lambda) = 2\lambda x \exp(-\lambda x^2).$$

   (a) Given a sample of data $x_1, x_2, ..., x_N$, derive an equation for the MLE of the parameter λ.

   (b) Download the data file "traffic.csv", which contains $10,000$ samples from the above distribution of hypothetical server request time intervals. Use your MLE equation from above to compute the MLE of λ for this data.

   (c) Plot a histogram of the data sample. Then plot the above pdf, $p(x;\lambda)$, using your MLE found in part (b) for the λ parameter. Describe what you found from comparing the two plots.

2. **Hypothesis Testing.** Here we are going to test a hypotheses about cardiac measurements from the provided "cardiac.csv" file. You want to test the hypothesis that women are more likely to have hypertension (high blood pressure) than men. Hypertension is the variable `hxofHT` (be careful, `hxofHT = 0` indicates they **do** have hypertension) and `gender` is male = 0, female = 1.

   (a) What is the 2×2 contingency table for this data? The rows of your table should be `gender` and the columns should be `hxofHT`. The four entries of the table will be counts from the data. For example, one entry will count the number of people who are both women (`gender = 1`) and have hypertension (`hxofHT = 0`), etc.

   (b) Write a Python function to compute the probability of getting *exactly* this table.

   (c) If you want to test if women have hypertension more frequently than men, what is the *null hypothesis*?

   (d) Again, using the function built in (b), perform the *Fisher exact test* to get a p value for the hypothesis that women have hypertension more frequently than men. Can you "reject the null hypothesis" with the threshold $p \leq 0.05$?

3. **K-means.** We experiment with unsupervised learning-Kmeans. Your implementation objectives are: (1) a python function to randomly initialize the centers, and (2) test kmeans function on a simple 2D data (provided as "2D data.txt"). If you don't see reasonable clusters coming out, something is probably wrong. Your results should include:

- A plot of the data, unclustered.
- The data clustered with $K = 2$. Plot your clustered data points with different colors.
- The data clustered with $K = 3$ (run your random initialization for 10 times... you should see a small amount of variability in the outputs). Plot (i) all 10 distance scores (the summation of distances between each point to its center), and (ii) clustered data points with different colors.
- A plot of $K \in \{4, 6, 8, 10, 15, 20\}$ versus distance score. Any random initialization is fine. No need to repeat different initializations.

4. **Bonus Points (Optional).** In this experiment, test your Kmeans clustering on the MNIST hand written digits database (provided as "trainX.txt", "trainY.txt", "testX.txt", and "testY.txt") with $K = 10$. Report your estimated mean images (centers). Note that you will need to re-shape each row (representing a digit image) of the "trainX" or "testX" as a 28^2 matrix and then plot as an image using Python functions (e.g., matplotlib.pyplot.imshow).