ECE 3502 / CS 3501: Foundations of Data Analysis **Due Tu 3/23**

Homework 2: Hypothesis Testing

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 or other sources, and do not show your answers to anyone. **Cite any sources you used outside of the class material (webpages, etc.), and list any fellow students with whom you discussed the homework concepts.**

1. In this problem you will experiment with unsupervised learning using K-means. You should implement two functions: (1) `clusterInit`, which should initialize random clusters according to the "furthest first" heuristic (select the first center randomly from the data, select subsequent centers as far as possible from any of the previous centers, e.g., if you've already selected four centers, the fifth center should be the point whose distance to any of the 4 centers is maximum.), and (2) `kmeans` to run K-means on a set of input points.

   Now test your K-means implementation on a simple set of 2D data (provided as `2D_data.csv`).

   (a) Plot the data unclustered (in other words, all points are the same color). How many clusters do you think there are? Use your answer to set K for all the experiments below.

   (b) Plot the data clustered with your K-means algorithm and random initialization (center points are picked uniformly randomly from the input data points). Use different colors for each cluster of points. Repeat the previous clustering and plot the results for a total of 5 times (using different random initialization each time). What do you notice about the repeated results?

   (c) Now run the K-means algorithm on the data 5 more times, with the same K, but now using the "furthest first" intialization (each run should have a different random initialization for the first center point). Compare this to the uniform random initialization result in part (b). What difference does it make?

   Next, test your K-means clustering with the "furthest first" initialization on the MNIST handwritten digits database (provided as "`mnist.csv`"). These are greyscale, 28×28 images of digits 0 - 9. The table contains one image per row, flattened into a 784-dimensional vector.

   (d) Plot a 10×10 grid of 100 of the input data as images. Reshape each of the 784-dimensional input data vectors as 28×28 arrays and display them as images.

   (e) Report the means found for $K = 5, 10, 15$. Again, reshape your mean vectors as 28×28 arrays and display them as images.

   (f) For each of these, do you see clusters means that look like the actual digits? What is the effect of changing K?

   (g) Now, using your $K = 10$ clustering result, plot a 10×10 grid of the input images, where each row contains 10 images all from the same cluster (i.e., one row for each cluster). Are the images clustered correctly? Which digits worked best? Which worked the worst? Why do you suppose some digits are over- or under-represented?

2. Write a Python function that computes the probability function for a hypergeometric random variable, X. (See the class notes and Wikipedia page for this formula.) Your function should take inputs:

$N =$	number of available bits to select from
$K =$	number of available bits that are 1
$n =$	number of bits drawn at random
$k =$	number of bits drawn that are 1

   Your function should return $P(X = k)$. Using your function, compute the following:

   (a) Recall the "lady drinking tea" example from class. Verify that your function gives the correct values for $k = 2, 3, 4$. (See the notes for the right answers!)

   (b) You are running an internet security firm trying to catch packets sent to a server by hackers. There are 100 packets sent to the server, with 10 of them from hackers, 90 from legitimate traffic. If you sample 50 packets at random, what is the probability that you will capture all 10 packets from the hackers?

   (c) What is the chance that you will capture at least half of the hackers' packets? That is, what is $P(X \geq 5)$? **Hint:** You are going to need to sum probabilities from multiple calls to your function.

3. Here we are going to test a hypotheses about cardiac measurements from a study of cardiac disease contained in the file "`cardiac.csv`".

   To understand what the variables mean, read the description of the data set here: `http://tomfletcher.github.io/FoDA/homeworks/cardiac-explanation.txt`

   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) Using your hypergeometric probability function from the previous question, 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 your hypergeometric probability function, 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 $p \leq 0.05$?