ECE 3502 / CS 3501: Foundations of Data Analysis **Due Tu 2/17**

Homework 1: Getting Started with Probability

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. The UVA men's and women's basketball teams are both playing Virginia Tech on the same day. There is a 60% chance that the UVA men win and a 60% chance that the UVA women win. In addition, there is a 75% chance that UVA wins at least one of the games (that is, they win the men's game, or the women's game, or both). Let M denote the event that the UVA men win, and let W denote the event that the UVA women win.

   (a) What does $P(M \cup W)$ mean in English? What is its value?

   (b) What is the probability that UVA wins both games?

   (c) What does $P(M \mid W)$ mean in English? What is its value?

   (d) What is the probability that UVA wins **exactly** one game (men or women, but not both)?

   (e) If the UVA women lose, what is the probability that the UVA men also lose?

2. A man is accused of murder, and the prosecutor has fingerprint evidence from the crime scene that matches the defendant. There is only a 1% chance of a fingerprint match if the defendant were not guilty. However, there are 101 people in the town where the murder was committed, that is, there is a 1/101 chance that the man is guilty when the evidence is not taken into account. What is the probability that the defendant is guilty, given the fingerprint evidence?

3. According to the American Lung Association, there is a 0.13% chance to develop lung cancer. Of the people who have lung cancer, 90% of them are smokers. In the population of people who do not have lung cancer, 16.9% are smokers.

   (a) What percentage of the total population are smokers?

   (b) If you are a smoker, what is your probability to develop lung cancer?

   (c) If you are not a smoker, what is your probability to develop lung cancer?

4. In this exercise we will be using data from the OASIS brain database, a publicly-available resource:

   `http://www.oasis-brains.org`

   You will be classifying dementia from the volume of the hippocampus, a brain structure that is critical to memory. The data you will use is in the spreadsheet `OASIS-hippocampus.csv`, which you can download from the class website. The data consists of the hippocampal volume, derived from MRI, for elderly subjects, including healthy control subjects and those with mild to moderate dementia. Model the right hippocampal volume (`RightHippoVol`) as a normal random variable X_1 and the left hippocampal volume (`LeftHippoVol`) as a normal random variable X_2. Then model the diagnosis (`Dementia`) as a binary random variable Y ($Y = 0$: healthy control, $Y = 1$: dementia). Use the training subset of the data (`TrainData = 1`) to learn the mean and variance parameters for a naïve Bayes classifier. Finally, apply your naïve Bayes classifier to get a probabilistic diagnosis of the test subset of the data (`TrainData = 0`).

   Do the following:

   - Plot the data as a 2D scatterplot (right and left hippocampal volume as the two axes). Use two different colors for the two classes (healthy/dementia). Do you think there is separation between the two classes?
   - Plot two density plots for the left and right hippocampus volumes. Again, plot a different density for the two classes (with the same colors as your scatterplot).
   - Run your classifier on the testing data. For each data point, if your classifier probability is greater than 0.5, predict that it is a dementia patient. Then compare with the actual label to see if your classifier is correct. Report your classifier accuracy on the testing data (number of correct classifications divided by size of the test data).

   Note: There are Python machine learning libraries that include a method for naïve Bayes. You can't use these to solve this problem! You have to write your own code to implement naïve Bayes. However, it's okay to try out a library's version to check if it gives a similar answer to your code.