Fall 2019: STAT 445 / CSCI 555 Statistical Learning and Data Mining

Graduate Project. To be completed in either Matlab or Python

Deadline: Thursday November 28th, 2019, In class (that is the last lecture!).

In the course assignments you were instructed to access a dataset compatible with supervised machine learning based *classification*. You will either reuse your dataset from the earlier assignments, or if you want something more interesting, challenging, or just want more experience, you can access a *new* dataset. In this assignment, graduate students are expected to demonstrate creativity in the application of the pattern recognition techniques taught in this course as well as techniques that build on the concepts taught in this course. Many example graduate level projects have been discussed in class lectures. You will be graded based on the quantity and quality (correctness, challenge etc.) of the techniques you implement in your project analyzing the dataset you've collected.

Question 1: Provide a point-by-point summary of very brief (1 line) statements that outline what you've completed as part of your graduate project. *Examples*: (you can choose any task under the sun, it doesn't have to be these ones and bonus points for being creative and performing challenging coding tasks)

a) Obtained astronomical dataset for supervised machine learning (SL)
b) Validation of the random forest, the Artificial Neural Network etc. performed
c) Combination of above techniques with PCA feature reduction investigated
d) Detailed analysis of random forest parameter variability effect on performance
e) Application of K-Means unsupervised learning with comparison to SL
f) Implemented (from scratch) the code for an existing learning algorithm and validated its performance on this dataset.

Question 2: ***This is identical to Assignment 1, Question 2. If you are using the same dataset, just reuse your previous answer (paste it here), if using a new dataset, describe it here.*** Describe the dataset you have collected: total number of samples, total number of measurements, brief description of the measurements included, nature of the group of interest and what differentiates it from the other samples, sample counts for your group of interest and sample count for the group not of interest. Write a program that analyzes each measurement individually. For each measurement, compute Cohen's d statistic (the difference between the average value of the group of interest and the average value of the group not of interest, divided by the standard deviation of the joint distribution that includes both groups). Provide a printout of the 10 leading measurements (d statistic furthest from zero), with their respective d statistics, making it clear what those measurements represent in your dataset (these are the measurements with the most obvious potential to inform prediction in any given machine learning algorithm). If your dataset has fewer than 10 measurements, provide the d statistic for all of them. Provide a printout of this code.

Question 3: Provide a detailed description of what you've done for each point from Question 1 (keep them labelled clearly so they can be matched to the list in Question 1). Provide code and sensibly organized results (output) for us to assess what you've done for each bullet point.