Final Project: Data and Feature Extraction

CSCI 297: Introduction to Machine Learning

Due: Tuesday November 10th @ 11:59 PM

Professor Watson

In this assignment you will be extracting and preparing your dataset for your final project. The dataset you synthesize has two potential possibilities: the first is that you use a previously extracted dataset from a research paper, the second is that your group extracts and prepares a dataset manually. A manual extraction should consist of not less than 150 examples of data with at least 15 features per example.

After extraction, you should provide some exploratory data analysis and go through at least some feature selection. You are required to extract at least 15 features; this means that some of those features will likely not benefit your ML model. You should apply a feature selection process to determine if there are any features you do not want to include. You are not required to spin up or create any classifiers to do this analysis. Therefore, you do not have to complete your EDA and feature selection process in this step, but you do need to start it and show evidence of both.

I would also like you to perform some feature transformation and dimensionality reduction of your dataset. It is likely that some of the features you extract will not originally be in the correct format. You need to demonstrate that you can transform these features into a format that would be acceptable to your model. Additionally, you can apply some dimensionality reduction to decrease the size of your feature space.

Most importantly, you need to document, document, document. The more you can explain about your process the better. I am looking for critical thinking steps and reasoning behind your applications of these techniques. Demonstrate you know how, why and when to apply these techniques to your features. It is possible that explaining why you did not do something could be as useful as explaining why you did do something else.

You should turn in a zip file with the dataset, any code you use in EDA, feature selection, dimensionality reduction or feature transformation, and your documentation explaining your process.

Good Luck!