Assignment 6: Naïve Bayes

CSCI 297: Introduction to Machine Learning

Due: Saturday October 18th @ 11:59 PM

Professor Watson

In this assignment you will be creating a naïve bayes classifier to determine whether an adult makes over 50K per year. There are 14 different attributes and a classification of whether the adult makes over 50K. There are 32,561 examples of adults, so this assignment will also give you a taste of beginning to work with larger datasets. For the attributes, you have age, working class, fnlwgt (number of similar individuals this datapoint represents), education, education number, martial status, occupation, relationship, race, sex, capital gain, capital loss, hours per week, native country and income.

I am not concerned with how you code the naïve bayes classifier. You should perform some exploratory data analysis to provide evidence for feature selection, regularization, scaling, manipulation, or any other thing you do to your dataset. Make sure you also test the variety of naïve bayes classifiers you have access to.

Once you have a good performing classifier, you should perform a series of tests to determine the best hyper parameterization. I would like to see some evidence of this testing; it can be a graph of accuracy given changes in hyperparameters or any other clever way you can demonstrate your testing. This would include the type of NB you choose to implement. There should be some justification as to why you chose this type.

You have complete freedom for this assignment in terms of how you code the assignment (no right answers per say). I am only interested in the results in terms of accuracy and precision. You may use any code from canvas or our book, but you **may not** use the internet.

This will be a paired programming assignment. You must work with someone else in the course. Pairs should not discuss their assignment with other pairs, please just consult one another. A large part of getting good accuracy on this assignment is manipulating the hyperparameters. I am not going to give you a marker for accuracy, you should get the best accuracy you can. Your explanations, process, steps, and analysis should all be documented and turned in along with your code. This well help determine if you took enough steps to increase your accuracy.

Good Luck!