STA208: Midterm Exam

Prof. James Sharpnack

Due Monday 5/2 at Noon 12pm on smartsite

Download the dataset listed in the exam section of the website. You should work alone on this and will need to submit two things. The instructions of this exam are intentionally open ended. In any job that you get, you will never get a dataset and be told to implement a linear SVM, so I am hoping that this will be more similar to a real world example.

- An executive summary, which is 1-2 pages long that describes the data, any preprocessing, the methods that you used, and your conclusions regarding the performance of your method.
- Two code files, one should be titled “training.py” or “training.R” which should contain all of the code that you wrote to complete your executive summary. The second file should be called “testing.py” or “testing.R”.

The problem. You work for a company and you are told that you need to write code that can predict if a customer will buy your product or not. There are a four features that your boss tells you that you can use to predict this. They have been recording if a customer has bought the product or not and the four corresponding features. The first column of the dataset is the response variable, and the remaining columns are the features.

You should first tune a classification algorithm to solve this problem based on the provided training data. This should include transformations, algorithm selection, and tuning parameter selection. You should provide diagnostics for this selection and describe how well the method does. Your executive summary should have three sections: (1) a description of the data and the task that we need to perform with any visualizations and plots that you think is helpful, a (2) description of the methods, features, and tuning that you tried, and (3) the evaluation and conclusions of the study.

The testing code should contain only the code necessary to make a prediction for a new feature vector. There needs to be a function (def in python) that takes a new 4 dimensional feature vector and returns a 1 or 0 depending on which prediction you make. You will be evaluated based on how thoroughly and creatively you formulate methods to solve this problem. This includes any preprocessing, plots, or initial transformations. Your testing code will be evaluated for its ease of use and if its results are consistent with your conclusions.