PSTAT 131 Progress Report

Nathan Fritter

12/1/2016

Question 1: Description of your Dataset

Part A: What are some of the attributes in this dataset? How many observations?

For this project I am using the grades in the Portugese class. Some of the attributes of this data set are: - Age - School - Gender - Mother's/Father's Income - Mother's/Father's Education - Free time - Travel time to school - Number of failures - Daily Alcohol Consumption - Weekly Alcohol Consumption - Number of Absences - Grades (G1, G2 and G3; G3 is the final grade) There are 649 observations in this data set.

Part B: Do you think all attributes to be useful in your analysis? Why or why not?

No I do not think all attributes will be useful in my analysis. Unsurprisingly, the different grades are highly correlated, and at least one of them needs to be removed. I chose to remove G1 because G2 was more important to the initial model. Also, certain variables such as age, school, gender, romantic, address, and others would likely have no correlation with the final grade variable, and could be removed.

Part C: How would you rate overall quality of the data? Justify your response.

The teacher and I have talked about this, and we realized that the paper that involved this data set was not very well written, along with the fact that it had never been published before. Coupled with the fact that this data set did not help with encoding this variables (I had to order most of them to make them ordinal), I would say that the quality of this data set was not very good. I was able to transform the data, so this improved the quality of the data.

Question 2: Description of your Research Question

Part A: What is the research question your project will address? (Summarize into one or two sentences.)

The research question my project will address is whether or not various student information (as alluded to above) can be a good predictor of final grades in a class.

Part B: Can you think of any other questions that may potentially be answered using the same dataset? (Think creatively for this thought experiment.)

For my project, I believe that some of the other variables could be predicted with the rest (like alcohol consumption as a function of the other variables, perhaps excluding grades). In fact, the paper that was based off the data set was titled as predicting student alcohol consumption.

Also, an unsupervised analysis would likely provide interesting insight into the relationship between explanatory variables (perhaps travel time to school versus alcohol use, parent's education versus amount of failures, etc.). Being able to make claims like increased alcohol use is correlated with increased travel time to school would be a nice side benefit of this project. However, most of the data is either ordinal or non ordinal factors, so unsupervised learning would be a tough assignment for this project.

Question 3: Description of your Analysis

Part A: What methods will you use to build models (at least two) for comparison?

I will utilize a decision tree with a random forest to build on the decision tree for improved accuracy. I may also add in a bootstrapping method to further increase accuracy, as well as lower the bias and variance of the model.

I will also compare this method to a Support Vector Machine; this algorithm has the ability to create non-linear decision boundaries between clusters of points with different class labels, so this seemed like a good candidate to compare with.

Part B: How will you choose each of your models? (model selection) State any relevant validation metrics, resampling methods, "rules-of-thumb", etc you are using.

First, I will look at the correlation between the numeric variables (mainly the grade variables) and see if any are worth removing from the model (I did end up removing G1 from the data set for analysis).

Next, I will couple a decision tree with a random forest; the random forest method cross validates decision trees with different numbers of nodes (branches), and I will be able to pick the tree with the number of nodes that gives the best results. In addition, I will add in a bootstrap method to resample the data when creating the decision trees.

To compare, I will create a Support Vector Machine. To find the best model parameters, I will use the train function in the "caret" package and add in a method for 10 fold cross validation to reveal the best model parameters (cost & sigma) and then create a SVM object using these model parameters and check for accuracy using the test error rate and confusion matrix.

Part C: How will you compare the selected models?

I will compare the models using the test error rates from the best random forest (plus bootstrapping) model and the best Support Vector Machine model.

Question 4: Description of your Progress

Part A: Summarize what has been completed and what remains to do.

I have created the random forest and Support Vector Machine with the best parameters, and compared them using the test error rates.

I still have to add in the bootstrap method for resampling the decision trees, but so far both models have very good test error rates (below 10%) and I have a feeling that bootstrapping the random forest will only marginally improve the model.