Sean Cordrey

November 14, 2014

Stat 3654

Roberts

Weekly Paper 12

Chapter 6 in our R books discusses memorization methods. Depending on what type of problem we are trying to answer it’ll give us different results but a couple examples include returning the majority response in a categorical problems or average/expected scores for scoring problems.

Starting with the simpler cases, we have single-variable models. For categorical problems, we use a table which is called a pivot table or contingency table by different people. R implements this with the table() function we can use the summary() function to see what our results are. The book also discusses what to do with a calibration set to check for overfitting in the model and using the test set as our final check. For numerical problems, we can apply R’s quantile() and cut() functions. What this does it put our numerical features into well-distributed levels (so that we don’t have to worry about overfitting). We can then plot these onto a double density plot and compare the areas under the curves to draw some conclusions.

Next we can step up into the far more likely cases of having multiple variables. An important part of constructing these models is deciding which variables should be used and treated in the model. One of these models is decision trees. We can construct them with the rpart() function. This function unlike many others in R has its own way of dealing with NA values. However, if there are more NAs than the function was designed for, one may need to begin adding additional arguments to the rpart() function. After carrying out this function, a few more lines of code gives us a graphical representation of our decision tree. Another method is k-nearest neighbors which uses Euclidean distance to compare values to the nearest value, and takes the average to return a score. The final model the chapter discusses is Naïve Bayes which forms its decisions by constructing a large number of single-variable models.