Machine Learning Supervised Learning Algorithms Analysis

The Datasets

Student Alcohol Consumption

Description

This data set takes a number of factors and compares it against the amount of alcohol a student consumes. The population of the data set consists of students in one of two schools attributed in the data taking either a Portuguese or Mathematics course. The data was initially separated by subjects in two separate csv files and contained course final grades in the subject. Since my primary interest was to gauge correlation of other variables to alcohol consumption I formatted the data to exclude course grades. I also combined the two files and added an extra variable indicating the course a student was taking. Two factors were used to assess alcohol consumption: workday alcohol consumption and weekend alcohol consumption, both on a scale from 1 to 5. I combined both these factors, and created a categorical variable, High. If total alcohol consumption was greater than or equal to six, alcohol consumption was considered high. There were two reasons for doing this. The first being that a one to ten scale is too broad and would likely have caused large errors in many of the algorithms. The second reason is that establishing a clear boundary ensures that the analysis is more meaningful in explaining what the algorithm predicts.

Why this data set?

This data while only containing 1044 data points had 30 categorical variables. I considered this to be a more practical use of machine learning algorithms. A parent field of machine learning is statistics. In statistics calculating the influence of multiple variables to a single variable is achieved through multiple correlation against a linear function. I wanted to see how useful the algorithms are in comparison to traditional statistics.

Poker Hands

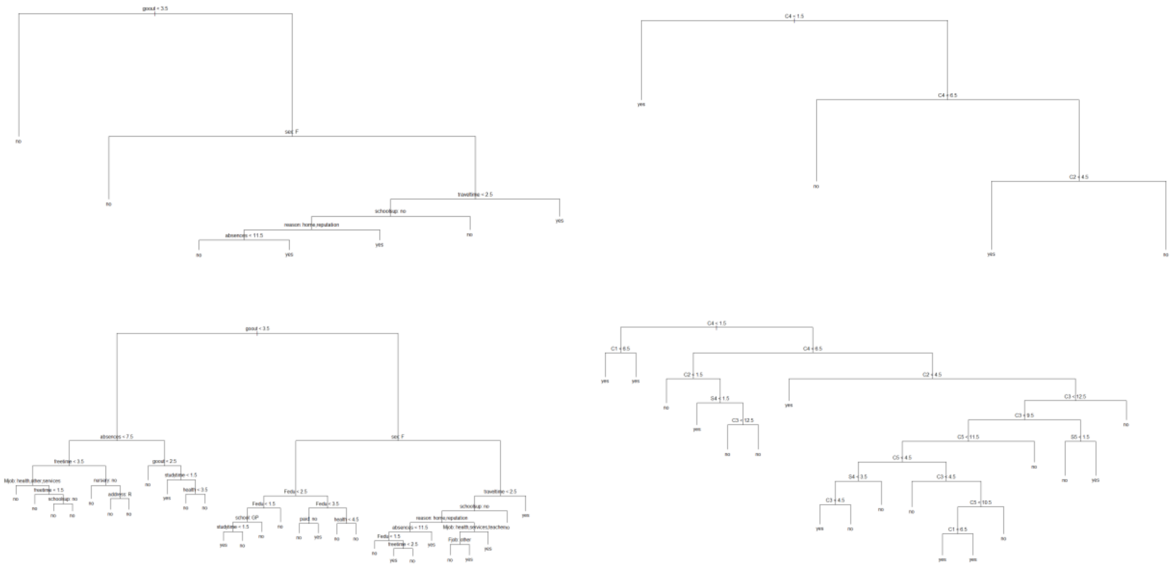
Description

This data set contains values of all five cards and their respective suits matched with the outcome of the hand as a whole. Initially this data set contained the outcome of the hand as a whole ranked from 0-9, 0 being nothing and a 9 being a royal flush. I chose to convert the outcomes to a binary variable called Something. If the hand contained something then Something was a yes, else something was a no. The reason for this was that roughly 50% of the hands in both sets contained nothing and roughly 42% of the hands contained one pair, leaving the remaining 8% to be distributed over the remaining outcomes, a one to one correspondence between these outcomes and their percentages appeared logarithmic, meaning that the higher outcomes would rarely occur; every combination above three-of-a-kind occurred less than 1% of the time combined.

Why this data set?

The number of combinations of a deck of hands is 52!. The number of winning poker hand combinations is upwards of three hundred million. I find this data set fascinating because the set winning combinations is very small compared to all possible hands. It is possible to program the rules of poker; I was interested to see how well a machine learning algorithm performs in comparison.

Decision Trees



Decision Trees Pruned and Unpruned. Student Alcohol Consumption Left, Poker Hands Right

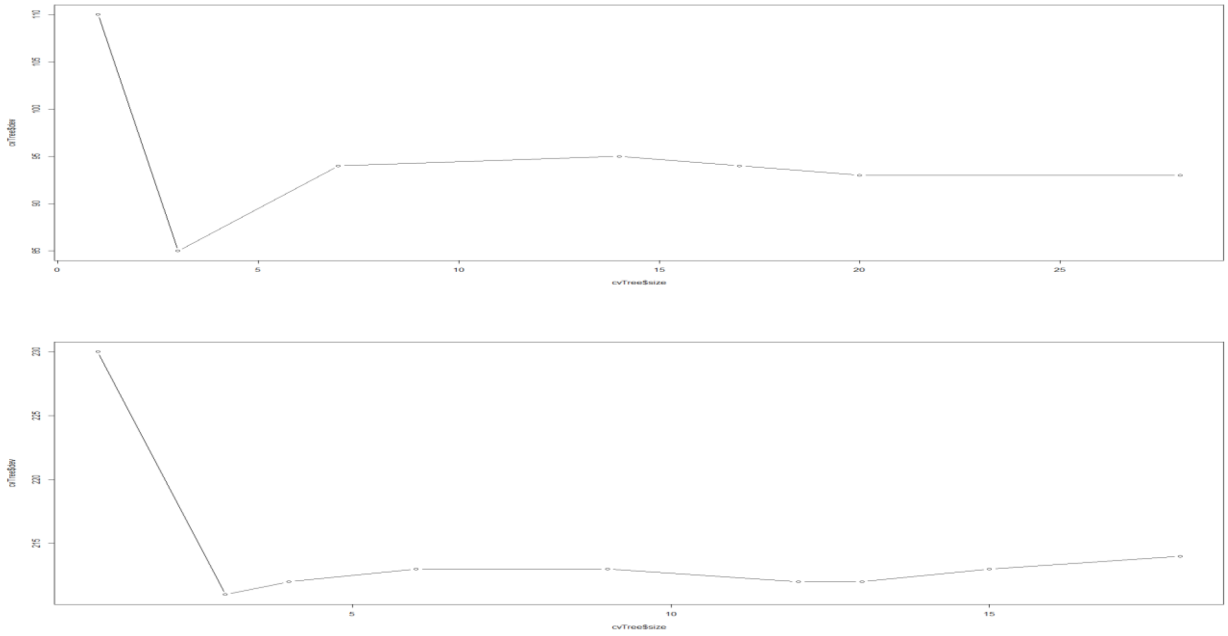
Description

The R algorithm for decision trees uses a variation of the ID3 algorithm that uses the GINI index instead of Information Gain to split the data set and returns an unpruned tree. I Cross-validated the trees and pruning based on the cross-validated tree with the lowest deviation. R was not able to process the poker hands dataset in its entirety and I took a sample of 417 data points from the training data to test against.

Analysis

Both of the algorithms computed fairly quickly. As I had expected the Student Alcohol Consumption set performed far better having an unpruned error of 0.1609195 and a pruned error of 0.1475096 on average which indicated that the data had overfit initially.

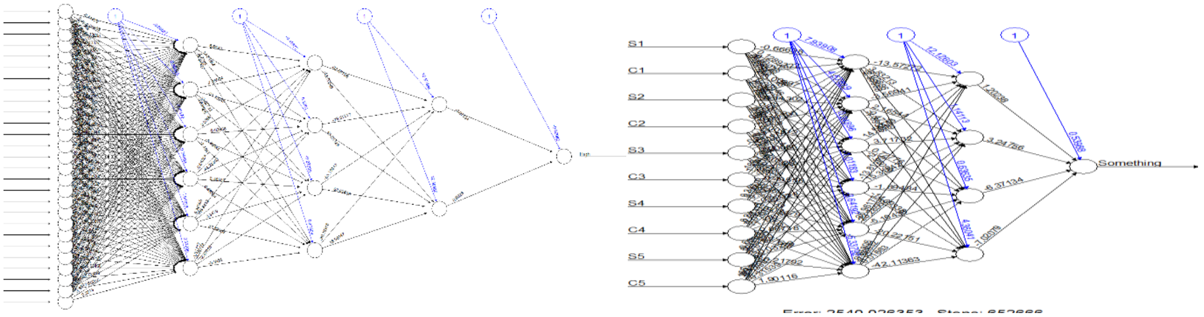
In comparison the unpruned error of the Poker Hands tree was 0.490002 on average, which still satisfies the criteria for a weak classifier, but the pruned error was 0.510575 on average. Such large errors in both cases were fairly unsurprising since the extrapolation of the outcome was from 417 hands and we were predicting 1 million hands.



Cross Validation Tree Deviation vs. Tree Size. Student Alcohol Consumption Top, Poker Hands Bottom.

Taking a look at the plotted decision trees, the pruned trees are far more elegant and in the case of student alcohol consumption reveal more information.

Neural Networks



Neural Networks Graphed. Student Alcohol Consumption Left, Poker Hands Right

Description

The R Neural network uses a resilient backpropagation algorithm and uses the sum of square errors to denote its error. For this algorithm a numeric representation of the alcohol data set was used. Due to a formatting bug, the guardian variable was excluded in the calculation.

Analysis

The algorithm by default uses a 0.01 as a threshold value for errors in partial derivatives as its stopping value and stops calculating after 100000 steps if the threshold is not reached. For training against poker hands I customized these value to 0.1 and 1 million respectively due to time constraints and the difficult nature of the data set. Despite this the algorithm took over 7 hours to compute 3 hidden layers. However, this algorithm had the lowest error for all algorithms training against poker hands 0.34266 on average. It was also the only algorithm that could process all of the training and testing values. In comparison, computing 3 layers for student alcohol consumption took 5108 steps and 11.77 seconds and produced an error of 0.1743295019 on average. This was the highest error produced for student alcohol consumption against all algorithms. This is not surprising as the neural network calculated the values as numeric values than discrete factors.

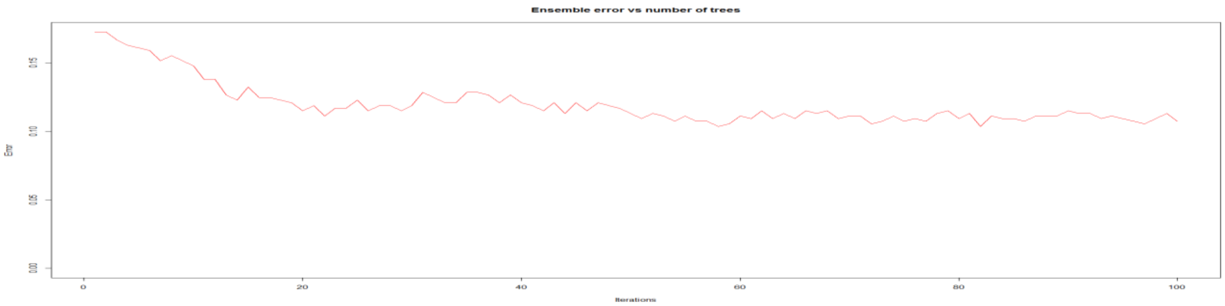
Boosting

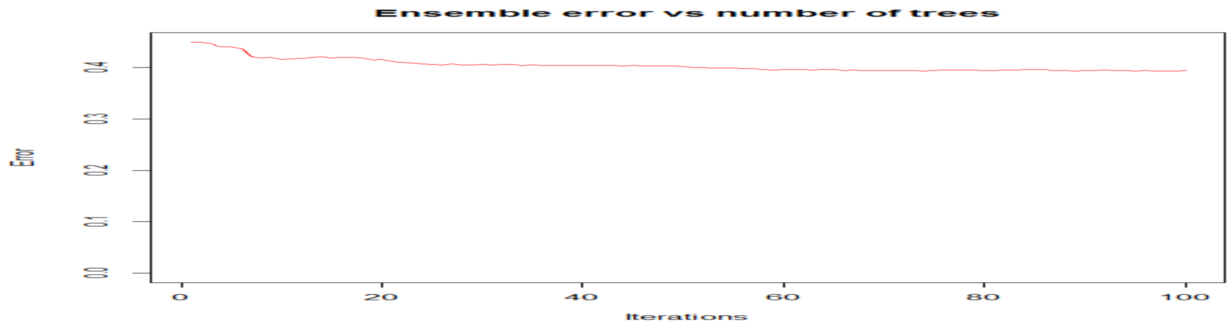
Description

R uses ADA boost and the Breiman coefficient of learning. It also runs 100 iterations on the give data. Since boosting needs weak classifiers the trees did not need pruning, the algorithm produces trees that are gently pruned to avoid under fitting. R was not able to process the entire training data and testing data; I took a sample of 2501 data points and 100000 data points for training and testing respectively.

Analysis

Boosting produced lower error for both data sets than decision trees alone as should be expected. This algorithm produced the lowest error for student alcohol consumption 0.1072797. In comparison the error for poker hands was 0.39349, which was far better than decision trees alone; however, it should be noted that the training sample size was larger and the testing domain was smaller for this algorithm than for the decision tree algorithm.





Ensemble Error vs. Number of Trees. Student Alcohol Consumption Top, Poker Hands Bottom.

Support Vector Machines

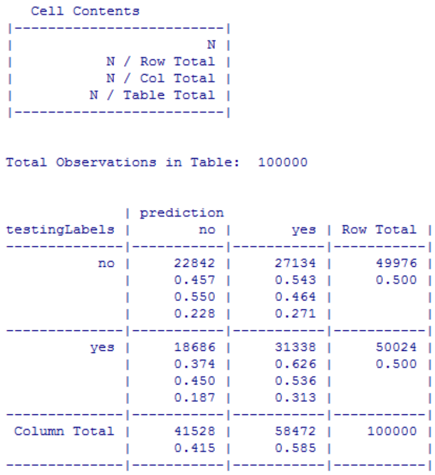
Description

The R support vector machine algorithm uses a radial kernel by default, I also tested against a third degree polynomial. R was not able to process the entire training data and testing data; I took a sample of 2501 data points and 100000 data points for training and testing respectively. In both cases, the algorithm checked the data against multiple variables and the algorithm was finding the margins of separation in dimensions greater than three; therefore, the support vector machine was not plotted.

Analysis

Support vector machines are useful tools when working with datasets that can be visualized in two or three dimensions, since the datasets did not meet this criteria, I did not expect the predictions to be good. For both data sets error predictions fell somewhere in the middle, being 0.43775 for poker hands and 0.137931 for student alcohol consumption on average with a radial kernel. When testing against a polynomial kernel the errors increased for both going to 0.49739 and 0.1551724 respectively.

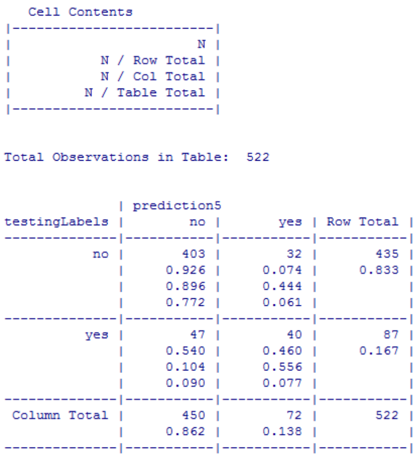
K-Nearest Neighbor



```
   Cell Contents
|-------------------------|
|                       N |
|           N / Row Total |
|           N / Col Total |
|         N / Table Total |
|-------------------------|


Total Observations in Table:  100000


             | prediction
testingLabels |        no |       yes | Row Total |
--------------|-----------|-----------|-----------|
           no |     22842 |     27134 |     49976 |
              |     0.457 |     0.543 |     0.500 |
              |     0.550 |     0.464 |           |
              |     0.228 |     0.271 |           |
--------------|-----------|-----------|-----------|
          yes |     18686 |     31338 |     50024 |
              |     0.374 |     0.626 |     0.500 |
              |     0.450 |     0.536 |           |
              |     0.187 |     0.313 |           |
--------------|-----------|-----------|-----------|
 Column Total |     41528 |     58472 |    100000 |
              |     0.415 |     0.585 |           |
--------------|-----------|-----------|-----------|
```



```
   Cell Contents
|-------------------------|
|                       N |
|           N / Row Total |
|           N / Col Total |
|         N / Table Total |
|-------------------------|


Total Observations in Table:  522


             | prediction5
testingLabels |        no |       yes | Row Total |
--------------|-----------|-----------|-----------|
           no |       403 |        32 |       435 |
              |     0.926 |     0.074 |     0.833 |
              |     0.896 |     0.444 |           |
              |     0.772 |     0.061 |           |
--------------|-----------|-----------|-----------|
          yes |        47 |        40 |        87 |
              |     0.540 |     0.460 |     0.167 |
              |     0.104 |     0.556 |           |
              |     0.090 |     0.077 |           |
--------------|-----------|-----------|-----------|
 Column Total |       450 |        72 |       522 |
              |     0.862 |     0.138 |           |
--------------|-----------|-----------|-----------|
```

Cross comparison data tables for best KNN prediction. Poker Hands left, Student Alcohol Consumption Right

Description

The R k-nearest neighbor algorithm can be customized to include a limit which serves as a minimum vote for a definite decision, for simplicity this variable was not included. The chosen values for k for both data sets were 20, 15, 10, 5, and 1. R was not able to process the entire training data and testing data; I took a sample of 2501 data points and 100000 data points for training and testing respectively. For this algorithm a numeric representation of the alcohol data set was used. Due to a formatting bug, the guardian variable was excluded in the calculation.

Analysis

While at k = 20, error on student alcohol consumption was low at 0.15900383, error spiked at k=15 and exhibited a downward trend. Error was 0.17432950 at k=15 and 0.15134010 at k=1. In the poker hands data set the opposite was apparent where error was lowest at k=20 being 0.45820 and highest at k=1 at 0.47045. There are two reasons why this might be the case. First, the student alcohol consumption set has nearly three times as many variables as the poker hands data set. It is far more likely that a neighboring data point has a similar outcome. In the case of poker hands the data points are clustered in pockets and the distribution of these pockets is not uniform, comparing against a larger set of neighbors is a better guarantee of identification into the right cluster.