Homework 8

Jon Ashbrock

March 22, 2018

For this homework we will be predicting profitability of movies based on various factors. We will use the "HollywoodMovies" dataset in Lock5Data. The column "Profitability" gives the money made by the movie as a percentage of the cost to make the movie. We use this to categorize movies into 3 categories: Those who made less than 80% are called not profitable, between 80 and 120% is called neutral, and greater than 120% is called profitable. As in the previous homework, you should make a list of factor variables containing the classifications of each movie. You should use columns $4, 6, 7, 8$, and 13 of the dataframe to build the decision tree. You should turn in the code, the decision tree graphic and the answers to the following questions. You are reminded at this point that you are only supposed to be working with others for ideas on how to write the code. The interpretations and question answers should be entirely your own work. A final hint, there is missing data in this data set so you should determine how to remove those movies.

1. What does the decision tree do when data is missing?

2. Which variable (that is, which column amongst 4,6,7,8, and 13) would you guess is the best predictor of how profitable a movie is? Why is that your answer?

3. Based on the decision tree generated, is the genre of the movie a good predictor of the profitability?

4. Could I theoretically use this model to predict how profitable a movie is before it is even released? Could I use it to predict how profitable the movies with missing profitability data are?

5. How accurate (as a percentage) is this decision tree at predicting the profitability of the dataset it was trained on?

6. Finally, could I have used an SVM to do this classification? Why or why not?