Cornell Data Science

Project 1

Regression and Classification

March 27th, 2016

Overview

How to submit this: Submit your code (.R or .ipynb), your predictions (.csv) and paragraph write up (.pdf) through CMS by 11:59pm April 11th.

Objectives: Understand and learn how to implement various prediction algorithms.

The Dataset

In 2013, 1,010 people aged 15-30 from Bratislava, Slovakia filled out a 150 question survey inquiring about their music preferences, movie preferences, hobbies & interests, phobias, health habits, personality traits, views on life, opinions, and spending habits. Most of the questions were yes or no, and people answered how much they agree on a scale of 1 to 5. You can read more about the dataset here and download the question description and answers **here.**

Deliverables

By the time this project is due, you will have learned about Linear Regression, Logistic Regression, Decision Trees, and K Nearest Neighbors. **You are expected to predict the values for one specific column in responses.csv, using 3 of the 4 prediction tools mentioned above**. You choose which 3 algorithms to implement, and you can choose to predict any column you want. Simply remove the original data, and generate a csv with your predicted values for every row in that column. **Include 5 columns in your predicted values csv to show that predictions made in 5 different train-test splits meet the requirements.**

After you're done with all of your predictions, please write a brief paragraph summarizing which columns and algorithms you chose, and why. You can also describe what challenges you faced or discovered, and how they influenced what you decisions you made. If you made visualizations to help choose which column to predict, you can include that too.

Scoring

To ensure that you have a decent understanding of each prediction algorithm, **all 3 of your predictions must have an accuracy at least 10% higher than the baseline, with 1 algorithm being at least 15% higher**. The baseline is different for each column, and is equivalent to the percentage of its most frequent value for a classification problem. If you choose to do a regression problem, the baseline and metric to use will be RMSLE or RMSE(either works). In this case, you must achieve a 15% **lower** RMSLE/RMSE than the RMSLE/RMSE obtained when you simply predict the average value for that column for all the test data points. **It will be easier to reach the required score if you pick columns with low baselines.**

Classification examples:

- For example, if I have a column where 5% of values are 1, 30% of values are 2, 10% of values are 3, 40% of values are 4, and 15% of values are 5, then the baseline would be 40%, since 40% of the values are 4.
- If there's a column where 5% of values are 1, 80% of values are 2, 5% of values are 3, 5% of values are 4, and 5% of values are 5, then the baseline would be 80%, since 80% of the values are 2.
- Column X has a baseline of 30%, so 40% of my predictions for this column must be correct. Column Y has a baseline of 80%, so 90% of my predictions for this column must be correct. Column Z has a baseline of 40%, so 50% of my predictions for this column must be correct. Each of these scores is 10% higher than the baseline. However, 1 of my predictions must be 15% higher.

Regression examples:

- If a column has an average of 10, then the baseline RMSE/RMSLE must be calculated using 10 as the predicted value for all test data points
- If the RMSLE of one column being predicted is 10, then you must obtain 8.5 and lower to pass the 15% accuracy increase