Mihov, Ilko

May 5, 2018

CS 506

**Homework 5**

Fine Foods Amazon Star Rating Prediction

**Kaggle Username:** ilkomihov

1. **Data**

The Amazon dataset provided us with data on thousands of reviews which contained user input about restaurants that they had visited. The dataset contained a unique identifier for the product, for the user, and for the review. Every review also contained a “HelpfulnessNumerator” (which showed us the number of users who found the review helpful) and a “HelpfulnessDenominator” (which showed us the number of users who indicated whether they found the review helpful). Along with that we were also provided with a summary of the review and its full text, a time stamp of when it was posted, and lastly the score that the user gave of the restaurant.

1. **Preprocessing**

Not all of the data contained in the reviews was useful to our model, so I decided to keep only use the columns that signify the “HelpfulnessNumerator”, “HelpfulnessDenominator”, the Score when performing the data analysis. However, this gave me very unconvincing results, so I had to add the Text category which contained the written review that the user had given. There was no way for me to use the actual text in my machine learning algorithms, so I decided to use TextBlob to obtain the sentiment analysis of each textual review and use that instead in order to give a numerical representation of this field. After all of this was done, I removed all of the data that did not contain a Score entry.

1. **Machine Learning**

After extracting all of the important information, I decided to test out a few popular machine learning algorithms on my dataset in order to figure out which one would work best. I ran Linear Regression, Support Vector Machine, and K-Nearest Neighbors from the Scikit-Learn Python library. I used the function cross\_validation.train\_test\_split() in order to scramble the data and divide it into training and testing sets(8:2). The results from my experiment proved that K-Nearest Neighbors is the most suitable algorithm to use for this project as it yielded a 0.74 confidence rate, as opposed to 0.24 for LR, and 0.38 for SVM. To perform cross-validation I used the cross\_val\_score function from the model\_selection module of Skicit-Learn.