HOMEWORK-4 REPORT (CS-506)

Name: Muhammad Kasim Patel
BU ID: U75595108
BU Username: kasimp93
kaggle Username: Kasim Patel

INTRODUCTION:

As stated in the homework description, the objective is to build a system that predict the star ratings associated with the reviews in Fine Foods on Amazon using the available features in the Amazon dataset. In this report, the predictive process from the feature extraction to using the correct model for final results is explained.

DATASET:

The dataset consists of reviews of fine foods from Amazon. The data includes 500,000 reviews. Reviews include product and user information, ratings, Product Id, User Id, summary, helpfulness, text review and score. There are 74,257 unique users.

PRE-PROCESSING THE DATA:

A number of pre-processing tasks were done on the text review and summary columns. These steps include:

- Lemmatization
- Removing Punctuation
- Removing Digits
- Converting to lower case
- Tokenizing
- Removing Stop words
- Pruning
- Feature Reduction using PCA

Using sentiment analysis toolkit from sklearn, I analyzed the text of review and made new features which include positive sentiment, negative sentiment, neutral sentiment and combined sentiments.

Feature reduction was done on the bag of words using Truncated SVD selecting 50 components which was giving the lowest MSE on the test set. Before feeding the data in the algorithms, it was normalized

FEATURE SELECTION:

Feature selection is an integral part of any type of predicting task and can increase or decrease accuracy of our algorithm to a great extent. Apart from the given features of the data set as mentioned in the description of the data I created some new features which could help increase the efficiency of my classifier model which include positive sentiment, negative sentiment, neutral sentiment and combined sentiments.

I tried to use different combinations of the features and finally ended up with two approaches. One was using only the bag of words with a 50 SVD components of the word features extracted using PCA for dimensionality reduction.
The other approach which I used was to drop the text features while using the new sentiment features and concatenating them with the features already given with the dataset.

MODELS:

We have learned many models for predicting in our lectures and previous homework, however using the right model is very important for a good result. I tried out different models which include:

- Linear Regression
- Random Forest
- Decision Tree
- Ridge Regression
- SVM
- Neural Network
- Naïve Bayes

PREDICTIVE TASK:

Ridge Regression uses L2-norm as regularization. While searching for optimal parameter, the best regularization strength parameter α was found to be 1.

The SVC approach to train the model also performed well with the linear model and C =1

The best performing model was the one with the least MSE. A two-layer neural network was the best performing model using only the 50 features extracted from summary of the reviews. The neural network had two hidden layers the first one of 10 neurons and the second layer with 5 neurons. Regularization was set to 1 and the maximum iterations to be 20,000. The model performs best using these parameters and using the 50 features extracted after feature reduction from PCA.

OFFLINE EVALUATION:

Offline evaluation was done using RMSE on the training set as well as cross-validation. The scores which had a star rating value greater than 5 were set to 5.

RESULT:

With the best performing model, the offline evaluation RMSE was around 1.21. I tried different methods using cross-validation on the training data to train it. Dividing it into test and train data sets.

CONCLUSION:

The model which performed better than others was picked. There were a lot of factors which could influence the results. Analyzing the data and its features before the fitting and predicting can reduce errors and avoid overfitting. With the best performing model, the offline evaluation RMSE was around 1.21.