CSE 158 Assignment 1

Sunny Manoj Solanki

TOTAL POINTS

20.5 / 25

QUESTION 1

1 Task 1 7.5 / 10

+ **0 pts** Correct

+ **7.5** **Point adjustment**

QUESTION 2

2 Task 2 8 / 10

+ **0 pts** Correct

+ **8** **Point adjustment**

QUESTION 3

3 Report 5 / 5

- **0.5 pts** Helpfulness: Did not beat baseline

- **0.5 pts** Categorization: Did not beat baseline

✓ - **0 pts** **Beat all baselines/ tried multiple things**

Sunny Solanki

CSE 158

Professor Julian Mcaualey

February 28, 2017

Assignment # 1 Written Report

Task # 1

In this task, I implemented a linear regression model to predict whether a user's review of an item will be considered helpful. The function was defined such that given the userID, itemID, and out Of reviews, the prediction how to be predicted based on the relevant data. My linear regression model that I used was prediction/outOf = θ_0 + θ_1 [,len(datum['reviewText'].split())] + θ_2 [datum['rating']] + θ_3 [len(datum['reviewText'].split())*datum['rating']] + θ_4 [datum['rating']**2] + θ_5 [len(datum['reviewText'].split())*datum['rating']] + θ_6 [datum['rating']*len(datum['summary'])]. Using these features allowed me to achieve a relatively high accuracy rate on the Kaggle competition. Using the length of the review Text, the rating for each product, and the the length of the summary aided in developing a reasonable linear regression model that was subject to minimal overfitting. Since the linear regression model outputs the result as prediction/outOf for each data item in the database, the result had to be multiplied by out Of to get the number of predictions.To increase the accuracy rate from the baseline, the training data was filtered so that only entries where outOf > 0 and nHelpful/outOf > 0.5. This allowed only reviews that were most likely to be helpful to be used in training the linear regression model.

Task # 2

In this task, I implemented a classifier using 5 SVM's to predict the category of the item from a review and product metadata. Five categories were used mainly for this task. A binary SVM classifier was first used on the training/validation/test data to predict if a particular item belonged to either the Men's category or Women's category. Initially, the algorithm went through all the data entry reviews to find the 500 most common words. To get the Xtraining data, a for loop was used to go through every data record to count how many of the common words appeared in the reivew. Several regularization constants of C = 0.01,0.1,1,10, and 100 were used to find the best one that best fits the training data. Then it was used on the validation data and test data to predict the categories. Afterwards, five SVM's were used to predict one of the five categories using the top 1300 words across all the reviews.

1 Task 1 **7.5** / **10**

+ **0 pts** Correct

+ **7.5** **Point adjustment**

Sunny Solanki

CSE 158

Professor Julian Mcaualey

February 28, 2017

Assignment # 1 Written Report

Task # 1

In this task, I implemented a linear regression model to predict whether a user's review of an item will be considered helpful. The function was defined such that given the userID, itemID, and out Of reviews, the prediction how to be predicted based on the relevant data. My linear regression model that I used was prediction/outOf = θ_0 + θ_1 [,len(datum['reviewText'].split())] + θ_2 [datum['rating']] + θ_3 [len(datum['reviewText'].split())*datum['rating']] + θ_4 [datum['rating']**2] + θ_5 [len(datum['reviewText'].split())*datum['rating']] + θ_6 [datum['rating']*len(datum['summary'])]. Using these features allowed me to achieve a relatively high accuracy rate on the Kaggle competition. Using the length of the review Text, the rating for each product, and the the length of the summary aided in developing a reasonable linear regression model that was subject to minimal overfitting. Since the linear regression model outputs the result as prediction/outOf for each data item in the database, the result had to be multiplied by out Of to get the number of predictions.To increase the accuracy rate from the baseline, the training data was filtered so that only entries where outOf > 0 and nHelpful/outOf > 0.5. This allowed only reviews that were most likely to be helpful to be used in training the linear regression model.

Task # 2

In this task, I implemented a classifier using 5 SVM's to predict the category of the item from a review and product metadata. Five categories were used mainly for this task. A binary SVM classifier was first used on the training/validation/test data to predict if a particular item belonged to either the Men's category or Women's category. Initially, the algorithm went through all the data entry reviews to find the 500 most common words. To get the Xtraining data, a for loop was used to go through every data record to count how many of the common words appeared in the reivew. Several regularization constants of C = 0.01,0.1,1,10, and 100 were used to find the best one that best fits the training data. Then it was used on the validation data and test data to predict the categories. Afterwards, five SVM's were used to predict one of the five categories using the top 1300 words across all the reviews.

2 Task 2 8 / 10

+ 0 pts Correct

+ 8 Point adjustment

Sunny Solanki

CSE 158

Professor Julian Mcaualey

February 28, 2017

Assignment # 1 Written Report

Task # 1

In this task, I implemented a linear regression model to predict whether a user's review of an item will be considered helpful. The function was defined such that given the userID, itemID, and out Of reviews, the prediction how to be predicted based on the relevant data. My linear regression model that I used was prediction/outOf = θ_0 + θ_1 [,len(datum['reviewText'].split())] + θ_2 [datum['rating']] + θ_3 [len(datum['reviewText'].split())*datum['rating']] + θ_4 [datum['rating']**2] + θ_5 [len(datum['reviewText'].split())*datum['rating']] + θ_6 [datum['rating']*len(datum['summary'])]. Using these features allowed me to achieve a relatively high accuracy rate on the Kaggle competition. Using the length of the review Text, the rating for each product, and the the length of the summary aided in developing a reasonable linear regression model that was subject to minimal overfitting. Since the linear regression model outputs the result as prediction/outOf for each data item in the database, the result had to be multiplied by out Of to get the number of predictions.To increase the accuracy rate from the baseline, the training data was filtered so that only entries where outOf > 0 and nHelpful/outOf > 0.5. This allowed only reviews that were most likely to be helpful to be used in training the linear regression model.

Task # 2

In this task, I implemented a classifier using 5 SVM's to predict the category of the item from a review and product metadata. Five categories were used mainly for this task. A binary SVM classifier was first used on the training/validation/test data to predict if a particular item belonged to either the Men's category or Women's category. Initially, the algorithm went through all the data entry reviews to find the 500 most common words. To get the Xtraining data, a for loop was used to go through every data record to count how many of the common words appeared in the reivew. Several regularization constants of C = 0.01,0.1,1,10, and 100 were used to find the best one that best fits the training data. Then it was used on the validation data and test data to predict the categories. Afterwards, five SVM's were used to predict one of the five categories using the top 1300 words across all the reviews.

3 Report 5 / 5

- **0.5 pts** Helpfulness: Did not beat baseline
- **0.5 pts** Categorization: Did not beat baseline
- ✓ **- 0 pts** **Beat all baselines/ tried multiple things**