CSE258 Assignment 1 Report

Tianyu Zhuang A53101494
Kaggle user name: mud G John

Task 1. Helpfulness Prediction

Data set:

With the given 200 thousand data set, as what we did in homework assignment 3, I used half of them as training data and the rest half as validation data. While using training set and validation set with 100 thousand data for each to modify the model and choose features, I used all the 200 thousand data for training to get the parameters θ before testing on the test set. Because I found the score to be more satisfying with more training data.

Features & Model:

Before training, I first filtered the training data by selecting a threshold for 'out of', that's the total votes. Because, generally speaking, data with larger 'out of' can be more useful than smaller 'outOf'. Data like {'nHelpful', 1; 'out of', 1} may do little work on our model. A review with large 'outOf' will appear on the top and tends to have more chance to get more votes. What's more, for our MAE score, the accuracy of larger 'out of' may have more influence on our final score since it may get larger absolute difference.

So 'outOf' is really an important feature. What I did first was trying different threshold for 'outOf' from 5 to 50 to get a minimum MAE on validation data. Finally I chose a threshold of 'outOf' = 41. But the result came to be over fitting a lot. A threshold around 15 could be better.

Here I used a simple model as linear model, the features I used and the reasons are shown below:

1. **Bias term;**

Because a rating deviating from the average rating, that's the common idea for most of person, may get less votes:

2. **Difference between rating and global average rating for all data;**
3. **Absolute difference between rating and item average rating for such item;**

Some words and punctuation may indicate a useful analysis:

4. **Whether a word like 'But' and 'However' appears in the 'reviewText';**
5. **The number of punctuations like '?' and '!' in 'reviewText';**

6. **The ratio of 'reviewText' length over 'rating'**
7. **One hot feature for each category including {'Men', 'Women', 'Boys', 'Girls', 'Baby'}**

Result:

The MAE on validation set is 0.17435**.** I ranked top 24% on seen data with

0.15929 but finally dropped to 51% with 0.17914 because of the overfitting, which may be caused by the inappropriate threshold.

Task 2. Rating Prediction

Data set:

Once again, I found using whole data for training performed better. So I used half of them as training data and the rest half as validation data for training to modify the model and chose parameters. I also modified the termination condition to get the minimum for loss function. Then I used whole data for training to get the parameters for testing.

Features & Model:

The model I chose is just the same as what I did in homework assignment 3:

$$f(u,i) = \alpha + \beta_u + \beta_i$$

Then iteratively update the parameters. After each update, I calculated the loss function, that's:

$$\frac{1}{N}\sum_{u,i}(\alpha + \beta_u + \beta_i - R_{u,i})^2 + \frac{\lambda}{N}\left[\sum_u {\beta_u}^2 + \sum_i {\beta_i}^2\right]$$

After the loss function increasing 5 times, I terminated the update and return the 6th parameters from the end, that's the α, β with minimum loss function.

Here, I tried different λ and finally chose $\lambda = 5.6$

Because someone may tend to give good rating to all items while some else can be picky. Thinking of the habit of rating for each user may have relatively greater influence on rating compared with the average rating for such an item, I tried to put more weight on β_u and less on β_i.Putting weights into update and loss function, I finally chose $w_u = 1.15, w_i = 0.85$ and more satisfying result.

Result:

The iterative update converged after 213 times of update. The loss function got a minimum as 1.6052786685809095. I ranked top 32% with 1.13353 and finally went up to top 10% with 1.08187.