Assignment 1 Report

November 15, 2019

Student ID: A53308934
Student Name: Deng Zhang
Kaggle User Name: amazingme

1 Task One - Read Prediction

In this task, we are asked to predict given a (user,book) pair from 'pairs Read.txt' whether the user would read the book (0 or 1).

1.1 Store the data

```
We need to read the file, and store it in list.
```

- For the training set:

I create a user-book dictionary, with the key represents one user, and the value set represents all the books he/she reads. Likewise, I create a book-User dictionary.
I also create userRating & bookRating list, which store the average score of each user/book.

- For the validation set:

Since the given data only consists of positive sample, we should sample a negative entry by randomly choosing a book that user hasn't read.

1.2 Populairity Predict

```
Create an identifier to predict whether a book would be read by a user or not
based on book's popularity.
```

just rank which books are popular and which are not, and return '1' if a book is among the top-ranked

1.3 Jaccard Similarity

```
Create an identifier to predict whether a book would be read by a user or not
based on Jaccrad Similarity.
```

get the max Jaccard similairy value of all the books that this user has read, return '1' if this value exceeds a threshold

1.4 Combine two identifiers

`Determine the best thersholds for each identifier, and combine it`

- Threshold:

Using loops to find out the thresholds for these two identifiers.

- Combination:

Get the prediction result of the two identifiers. If the result of these two identifiers are the same, make it the final prediction. If the result are not the same, predict it by Jaccard Similiarity using another thershold.

1.5 Result

Accuracy: 0.67983

2 Task Two - Rating Prediction

Predict people's star ratings as accurately as possible, for those (user,item) pairs in 'pairs Rating.txt'. Accuracy will be measured in terms of the mean-squared error (MSE).

2.1 Initialization

`Create the initial value of alpha, beta_u, beta_b`

- Alpha:

My alpha is the average rating of dataset

- BetaUsers:

BetaUsers is a dictionary. The key is userID, value is the average rating of this user - alpha

- BetaBooks:

BetaBooks is a dictionary. The key is bookID, value is the average rating of this book - alpha

2.2 Iteration

`Update the value of alpha, betaUsers, betaBooks by iteration until no specific change happen.`

Above steps based on the formula of the slides.
(I don't know how to type mathematical formula here :-(You can see it in the slides of Lecture 8!)

2.3 Optimization

`Find a best lambda that has the lowest MSE`

2.4 Result

MSE on Kaggle: 1.13702