Assignment 1

Le Yang (A59011332)

Team Name Ethannewbee

November 16, 2021

1 Cooking Prediction

Method

In this task, I use Jaccard similarity and popularity of the items to predict whether the use would cook the recipe.

1. Popularity

For the popularity part, we calculate the cooked number for each item(recipe), and sort them. We set a threshhold for the set, then we can get the popular item set.

2. Jaccard similarity

For the Jaccard similarity part, given the definition of Jaccard similarity as below:

$$Jaccard(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

Compared to the homework3, we not only use the similarity between users, but also use the similarity between items. That is, for a given {user,item} in test set, we first calculate the maximum similarity between the item and items that user cooked, and then calculate the maximum similarity between the user and all users that cook this item.

However, there could be some users that do not appear in the train set, for this part, we check if the item is in the popular set; If yes, we predict the user would cook the item, otherwise the user would not cook the item. Also, there could be some items that do not appear in the train set, for this part, we use the ingredients of the item and apply Jaccard similarity to find the most similar items in the train set; Then we can perform prediction on it.

After calculate the two similarities, we can set two thresholds for them.

3. Prediction

We need to incorporate the above two parts to perform prediction. After many experiments, I find that the popularity is a more significant coefficient for the task. In this case, for a given {user,item} in test set I apply the following strategy:

a. If the item is in the popular set, we predict 1 which means the use would cook the item. Otherwise, we go to step b.

b. If the item is not in the train set, we find the most similar one. Then we calculate the two similarities for the given {user,item}, if both sim1 and sim2 is above the threshholds, we predict 1 which means the use would cook the item. Otherwise, we predict 0 which means the use would not cook the item

Result

In this part, I do experiments and utilized the grid search method to find the better threshholds. To find the better threshholds, I also print the calculated similarities during the progress to find the common scope for the threshholds.

The final threshold for popularity is 0.694, for the two similarities is 0.1 and 0.01. And this can get **0.70510**, which is 2.66 % from the strongest baseline.

2 Recipe Rating Prediction

In this task, I apply the latent factor model to perform the rating Prediction task.

Method

We first try the simple latent factor model with the definition below:

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

Then we try the complete latent factor model with the definition below:

$$rating(u, i) = \alpha + \beta_u + \beta_i + \gamma_u \gamma_i$$

Because the use can only rating the item between 0 to 5, then if the rating prediction is above 5, we set it to 5; if the rating prediction is below 0, we set it to 0. For the two latent factor model, we tuning the regularization factor, the number of dimensions for γ_u and γ_i, the iterations.

Result

In this part, I do experiments on the validation set and utilized the grid search method to find the better regularization factor, the number of dimensions for γ_u and γ_i, the iterations. I get the best result **0.79375**, which is 3% improvement from the strongest baseline for regularization factor at 0.0000191, 2 dimensions for γ_u and γ_i and 100 iterations.