COMPSCI 753

Algorithms for Massive Data

Assignment 4 / Semester 2, 2022

Recommender Systems

Kaiqi Zhao

Submission:

Please submit (1) a PDF file reports the requested values and explanation of each task, and (2) a source code file contains detailed comments on CANVAS by 23:59 NZST, Sunday 16 October. If you use Jupyter Notebook, you can also submit an HTML file with detailed comments and explanations to the questions. The answer file must contain your student ID and name.

Penalty Dates:

The assignment will not be accepted after the last penalty date unless there are special circumstances (e.g., sickness with certificate). Penalties will be calculated as follows as a percentage of the mark for the assignment.

- **23:59 NZST, Sunday 16 October – No penalty**
- **23:59 NZST, Monday 17 October – 25% penalty**
- **23:59 NZST, Tuesday 18 October – 50% penalty**

1 Assignment problem (100 pts)

Recommender systems are widely used in many industrial companies, especially location-based service such as Yelp [1]. Yelp is an app that recommends businesses such as restaurants and stores to users. In this assignment, we will explore this dataset using the recommendation algorithms learned in the lectures. To make the task feasible on most of the laptops and PCs, we have extracted a manageable subset of the datasets, which contains the reviews on businesses in Las Vegas. The dataset could be found on the assignment page. The training ("train.json"), validation ("val.json") and test ("test.json") files are of the same format. They include the review id, business id, user id, stars (ratings) and date of the review.
Note:

1. Reading json files: `https://www.geeksforgeeks.org/read-json-file-using-python/`
2. Some users and businesses may not appear in all these three files. You might want to write codes to scan all records in the three files to obtain the set of unique users and unique businesses if needed.
3. You could test your codes on a small sample of the data and make sure it doesn't have bugs before running on the whole dataset.

1.1 Tasks

This assignment include the following tasks:

1. **Task 1**: Estimate the global bias b_g, user specific bias $b_i^{(user)}$ and item specific bias $b_j^{(item)}$ on the **training data**. Report the global b_g, the user specific bias of the user with user_id= "b4aIMeXOx4cn3bjtdIOo6Q" , item specific bias of the business with business_id = "7VQYoXk3Tc8EZeKuXeixeg". [10 pts]
2. **Task 2**: Implement the basic latent factor model without considering the bias: $r_{ij} = \mathbf{q}_i^T \mathbf{p}_j$. Set the number of latent factors $k = 8$. Run Stochastic Gradient Descent (SGD) for 10 epoches with a fixed learning rate $\eta = 0.01$ and regularization hyperparameter $\lambda_1 = \lambda_2 = 0.3$. Report the RMSE on the training data for each epoch. [30 pts]
3. **Task 3**: Use SGD to train the latent factor model with different values of k in $\{4, 8, 16\}$ and stop after 10 epoches. Report the RMSE for each value of k on the **validation data** ("val.json"). Pick the model that results in the best RMSE on the validation set and report its RMSE on the **test data** ("test.json"). [15 pts]

[1] `https://www.yelp.com/`

4. **Task 4**: Incorporate the bias terms b_g, $b_i^{(user)}$ and $b_j^{(item)}$ to the latent factor model: $r_{ij} = b_g + b_i^{(user)} + b_j^{(item)} + \mathbf{q}_i^T \mathbf{p}_j$ and learn the user bias and business bias from data. Initialize the user bias $b_i^{(user)}$ and item bias terms $b_j^{(item)}$ using the values computed in Task 1. Set the number of latent factors $k = 8$. Run SGD for 10 epoches with a fixed learning rate $\eta = 0.01$ and regularization hyperparameter $\lambda_1 = \lambda_2 = \lambda_3 = \lambda_4 = 0.3$. Report the RMSE on the training data for each epoch. After finishing all epoches, report the learned user-specific bias of the user with user_id= "b4aIMeXOx4cn3bjtdIOo6Q" , and the learned item-specific bias of the business with business_id = "7VQYoXk3Tc8EZeKuXeixeg". [30 pts]

5. **Task 5**: Similar to Task 3, find the best k for the model you developed in Task 4 on the validation set and apply the corresponding model to the test data. Compare the resulting test RMSE with Task 3. [15 pts]