COMPSCI 753

Algorithms for Massive Data

Assignment 4 / Semester 2, 2023

Recommender Systems

General instructions and data

Recommender systems are widely used in entertainment. **Goodreads** is the world's largest site for readers and book recommendations. In this assignment, we will explore one of Goodreads' datasets 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 dataset of History & Biography book reviews[1] (containing 2,066,193 reviews). We have split the dataset on **training data** (1,239,715 reviews), **validation data** (413,239 reviews) and **test data** (413,239 reviews). The corresponding files can be found on the assignment page. The training ("goodreads_reviews_historybio_train.json"), validation ("goodreads_reviews_historybio_val.json") and test ("goodreads_reviews_historybio_test.json") files are of the same format. Each line includes a user id, book id, review id, rating and date of the review.

Submission

Please submit (1) a file (.pdf or .html) that reports the requested answers of each task, and (2) a source code file containing detailed comments. Submit this on Canvas by 23:59 NZST, Sunday 15 October. The files must contain your student ID, UPI 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:

- **23:59 NZST, Sunday 15 October – No penalty**
- **23:59 NZST, Monday 16 October – 25% penalty**
- **23:59 NZST, Tuesday 17 October – 50% penalty**

[1] `https://mengtingwan.github.io/data/goodreads.html`

Tasks (100 points)

This assignment is composed of three tasks. Some considerations you may want to follow:

1. Data is provided in json files. Some help on reading json files:
   `https://www.geeksforgeeks.org/read-json-file-using-python/`

2. When developing your solution, it is recommended that you **test your code on a small sample of the data and make sure it doesn't have bugs before running on the whole dataset**. This will help fasten your development process.

Task 1 [10 points]: Explore biases

Calculate the global bias b_g, user specific bias $b_i^{(user)}$ and item specific bias $b_j^{(item)}$ on the **training data**. Report:
(A) [4 points] The global b_g bias
(B) [3 points] The user specific bias of user_id= "3913f3be1e8fadc1de34dc49dab06381"
(C) [3 points] The item specific bias of book_id = "16130".

Task 2 [45 points]: Implement the regularized latent factor model without bias using SGD

(A) [30 points] Implement the regularized latent factor model without considering the bias. The optimization problem that needs to be solved is (see slide 8 of W9.2 lecture notes):

$$\min_{\mathbf{P},\mathbf{Q}} \sum_{r_{ij} \in R} (r_{ij} - \mathbf{q}_i^T \cdot \mathbf{p}_j)^2 + \lambda_1 \sum_{i \in U} ||\mathbf{q}_i||_2^2 + \lambda_2 \sum_{j \in P} ||\mathbf{p}_j||_2^2$$

.

The initialization of $\mathbf{P}$ and $\mathbf{Q}$ should be random, from a normal distribution. Set the number of latent factors to $k = 8$. Use Stochastic Gradient Descent (SGD) to solve the optimization problem on the **training data** (see slide 9 of W9.2 lecture notes). Run SGD for 10 iterations (also called epoches), with a fixed learning rate $\eta = 0.01$ and regularization hyperparameters $\lambda_1 = \lambda_2 = 0.3$. Remember that the regularization terms involve the L2-norms of the $\mathbf{q}_i$ and $\mathbf{p}_j$ vectors for each user i and item j respectively.

Report the RMSE on the training data for each epoch, by using the RMSE formula (see slide 36 of W8 lecture notes):

$$RMSE = \sqrt{\frac{1}{|R|}\sum_{i,j \in R}(r_{ij} - \hat{r}_{ij})^2}$$

(B) [15 points] Use SGD to train the latent factor model on the **training data** for different values of k in $\{4, 8, 16\}$. For each value of k, train the model for 10 epoches/iterations. Report the RMSE for each value of k on the **validation data**. Pick the model that results in the best RMSE on the validation set and report its RMSE on the **test data**.

Task 3 [45 points]: Implement the regularized latent factor model with bias using SGD

(A) [30 points] Incorporate the bias terms b_g, $b_i^{(user)}$ and $b_j^{(item)}$ to the latent factor model. The optimization problem that needs to be solved is (see slide 11 of W9.2 lecture notes):

$$\min_{\mathbf{P},\mathbf{Q},b_i,b_j} \sum_{r_{ij} \in R} (r_{ij} - \mathbf{q}_i^T \cdot \mathbf{p}_j - b_{ij})^2 + \lambda_1 \sum_{i \in U} ||\mathbf{q}_i||_2^2 + \lambda_2 \sum_{j \in P} ||\mathbf{p}_j||_2^2 + \lambda_3 \sum_{i \in U} (b_i^{(user)})^2 + \lambda_4 \sum_{j \in P} (b_j^{(item)})^2$$

.

The initialization of $\mathbf{P}$ and $\mathbf{Q}$ should be random, from a normal distribution. 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 hyperparameters $\lambda_1 = \lambda_2 = \lambda_3 = \lambda_4 = 0.3$. 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= "3913f3be1e8fadc1de34dc49dab06381" , and the learned item-specific bias of the book with book_id = "16130".

(B) [15 points] Similar to Task 2 (B), find the best k in $\{4, 8, 16\}$ for the model you developed in Task 3 (A) on the validation set, by using RMSE to compare across these models, and apply the best of these models to the test data. Compare the resulting test RMSE with Task 2 (B). Analyse and explain your findings.

Note: In this case, you may have users and/or books in the validation or test set that are not in the training set (i.e. you may experience the *cold start* problem). Therefore, you will not have information about the bias of these users or items. For those users or items, use a bias of 0 in your calculations.