Lab Course Machine Learning
Exercise Sheet 10

Prof. Dr. Dr. Lars Schmidt-Thieme, Shayan Jawed
Information Systems and Machine Learning Lab
University of Hildesheim

January 14th, 2020
Submission on January 20th, 2020 at 12 noon, (on learnweb, course code 3116)

Instructions

Please following these instructions for solving and submitting the exercise sheet.

1. You should submit a jupyter notebook detailing your solution.
2. Please set the seed(s) to 3116.
3. Please explain your approach i.e. how you solved a given problem and present your results in form of graphs and tables.
4. Please submit your jupyter notebook to learnweb before the deadline. Please refrain from emailing the solutions except in case of emergencies.
5. Unless explicitly noted, you are not allowed to use scikit, sklearn or any other library for solving any part.
6. Please refrain from plagiarism.

Exercise 1: Exploring Movie Recommendation Dataset (4 Points)

Perform the statistical analysis of the two datasets given. Your analysis should provide as much information as possible. You must use all the related information of users and movies for the analysis i.e. rating, user (age group, zipcode etc) and item(genre, title, release date etc). The grading of this task depends on the useful information extracted from the given dataset, which can help in the learning process. The dataset:

- movielens 100k dataset, Rating prediction dataset (rating scale 1-5).
  Available at: http://grouplens.org/datasets/movielens/100k/

The tasks are as follows:

1) Showcase how the ratings vary across users, as an example consider whether the plot is able to tell if most ratings are only from a handful of users.
2) Showcase how the ratings vary across items.
3) Are there genres that are more highly rated than others?
4) What age groups prefer what genres based on ratings? You can bin respective ages to your preference.

Exercise 2: Implementing basic matrix factorization (MF) technique for recommender systems (8 Points)

In this task you are required to implement a matrix factorization (MF) technique for recommender systems. You are given a rating matrix $R^{n \times m}$ and you have to learn latent matrices $P^{n \times k}$ and $Q^{k \times m}$, where n is the number of users, m is the number of items and k the latent dimensions. You can solve the MF problem by implementing Stochastic Gradient Descent (SGD) learning algorithms (Algorithm LearnLatentFactors on slide 29). Measure the prediction quality (the RMSE score) on the validation and test dataset. You can set 10%/10% of ratings aside for validation and testing.

- normalize your data
- optimize the hyper-parameters i.e. λ regularization constant, α learning rate, k latent dimensions.
- Compute the validation RMSE.
- Compute the test RMSE.

The slides for this exercise are here: `https://www.ismll.uni-hildesheim.de/lehre/ba-18w/script/6_recommender-systems.pdf`

Exercise 3: Recommender Systems using matrix factorization sckit-learn (8 Points)

In this task you are required to use off-the-shelf libraries such as libmf or sckit-learn. You have to learn a matrix factorization model using coordinate descent method. Optimize the hyper parameters and perform a 3-fold cross validation. Compare your results with the results in task 1. List in detail which/how you used these libraries?, what it solves?, and why it is selected?. Present your results in form of plots and tables.