Lab Course Machine Learning
Exercise Sheet 9

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

January 12th, 2017
Submission on January 18th, 2017 at 11:55pm, (on moodle, course code 3112)

Instructions

Please read the lab related instructions, i.e. submission, report format and policies, at `https://www.ismll.uni-hildesheim.de/lehre/prakAIML-16w/exercises/ml_lab_instructions.pdf`

Datasets

1. **Recommender Datasets:** You can use one of the two datasets (or optionally, both datasets).
   (a) movielens 100k dataset D_1: Rating prediction dataset (rating scale 1-5). `http://grouplens.org/datasets/movielens/100k/`
   (b) movielens 1m D_2: Rating prediction dataset (rating scale 1-5).`http://grouplens.org/datasets/movielens/1m/`
2. The RMSE score for rating prediction is available at **Mymedialite website** `http://www.mymedialite.net/examples/datasets.html`

Exercise 1: Recommender 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). You can enrich the movielens datasets using the DBpedia datasets. The grading of this task depends on the useful information extracted from the given (and enriched) datasets, which can help in the learning process. Use tables, graphs to represent your information. [Hint: First complete your analysis without enriching the datasets, then proceed with enriching and analyzing.]

Exercise 2: Implement 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 (see Annex **??**). You are given a rating matrix $R^{(n \times m)}$ and you have to learn latent matrices $P^{(n \times k)}$ and $Q^{(m \times k)}$, 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) or Alternating Least

Square(ALS) or Coordinate Descent(CD) learning algorithms. You will follow a 3-fold cross validation protocol with train, validation and test data splits. Measure the prediction quality (the RMSE score) on the test dataset.

- normalize your data
- optimize the hyper-parameters i.e. λ regularization constant, α learning rate, k latent dimensions.
- Compute the test RMSE (averaged across the 3-folds).

Exercise 3: Recommender Systems using matrix factorization *libmf / 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.

Annex

1. Matrix Factorization Technique for Recommender System: by Y. Koren, `https://datajobs.com/data-science-repo/Recommender-Systems-%5BNetflix%5D.pdf`
2. mymedia light `http://www.mymedialite.net/examples/datasets.html`