Applied Machine Learning
Exercise 8

Prof. Dr. Amr Alanwar

December 2024

Recommender Datasets:

You can use any of the datasets (or optionally, both datasets).

1. **Movielens 100k dataset** D_1: Rating prediction dataset (rating scale 1-5).
   Movielens 100k Dataset
2. **Wine Quality dataset** D_2: (use `winequality-red.csv`).
   Wine Quality Dataset

The RMSE score for rating prediction is available on the Mymedialite website:
Mymedialite Datasets

Exercise 1: Recommender Dataset (2 Points)

Perform a statistical analysis of the two datasets provided. Your analysis should extract as much useful information as possible. You must use all the related information of users and movies for the analysis, such as ratings, user attributes (age group, zipcode, etc.), and item attributes (genre, title, release date, etc.). The grading of this task depends on the useful information extracted from the datasets, which can aid in the learning process. Use tables and graphs to represent your findings.

Exercise 2: Implement Basic Matrix Factorization (MF) Technique for Recommender Systems (4 Points)

In this task, you are required to implement a matrix factorization (MF) technique for recommender systems (see Annex 1). You are given a rating matrix $R_{n\times m}$ and must 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 number of latent dimensions. You can solve the MF problem by implementing Stochastic Gradient Descent (SGD), Alternating Least Squares (ALS), or Coordinate Descent (CD) learning algorithms. Follow a 3-fold cross-validation protocol with train, validation, and test data splits. Measure the prediction quality using the RMSE score on the test dataset.

- Normalize your data.
- Optimize the hyperparameters, i.e., λ (regularization constant), α (learning rate), and k (latent dimensions).
- Compute the test RMSE (averaged across the 3 folds).

Exercise 3: Recommender Systems Using Matrix Factorization Libraries (4 Points)

In this task, you are required to use off-the-shelf libraries such as `libmf` or `scikit-learn`. Learn a matrix factorization model using the coordinate descent method. Optimize the hyperparameters and perform a 3-fold cross-validation. Compare your results with those obtained in Exercise 2. List in detail which libraries were used, what they solve, and why they were selected. Present your results in the form of plots and tables.

Annex

You can use libraries:

- Scikit-Learn User Guide
- sklearn.metrics Documentation
- sklearn.model_selection
- sklearn.linear_model
- sklearn.preprocessing
- `matplotlib` for plotting.
- **Matrix Factorization Technique for Recommender Systems:** by Y. Koren Recommender Systems Paper