Matrix Fcatorization

Dependencies:

Scikit, Pandas and NumPy.

Installation:

pip install scikit-learn

Pip install numpy

Pip install pandas

Methodology:

Matrix factorization is a collaborative based filtering method in which the input is user-item interaction that can be stored in a matrix with each row represents each user, and each column corresponds to a movie. Since not every user is going rate each item, the user-item interaction matrix will be sparse. Therefore, the goal of matrix factorization method is to predict missing entries.\\

For predict missing entries, the rating matrix is decomposed into the product of two lower dimensional matrices. The user matrix has the same number of rows as the rating matrix and each row belongs to each person, while the movie matrix has a column for each movie. Each of the two matrices expresses the power of association between a user or an item with latent factors or hidden features. The feature vectors are chosen in such a way that the pair-wise inner product matches known ratings in the utility matrix.\\

Non-Negative Matrix Factorization (NMF) is used for the decomposition of rating matrix with 75\% of the data is used for training and 25\% for testing. The optimal parameters that result in the lowest RMSE is found and is chosen as the model. The opted number of latent factors is 15, initialisation is non-negative double singular value decomposition (nndsvd) and the optimiser is coordinate descent algorithm (cd). The similarity measure between the original rating matrix and the product of decomposed matrices is in terms of Frobenius norm. The final model is trained on the whole dataset and the recommendations are based on the prediction of the model with the lowest RMSE. Lastly, for a given user, the 5 unseen movies with the best predicted ratings are displayed.

Base Code:

Code courtesy: https://github.com/nicolasfguillaume/Recommender-Systems-Making-Movies-Recommendation/blob/master/MovieLens%20(NMF)%20v1.ipynb

The changes made to the code are the older version of classes is changed to the ones which are used at present. For instance, there were classes like cross-validation, grid search etc. for K-Fold cross-validation and grid search CV respectively which now come under the class model selection. Also from the optimization techniques, it is found that non-negative double singular value decomposition gives the best result and we have changed the random initialisation to nndsvd.