Movielens 10M. 2-rank Matrix factorization

Viacheslav Simonov

12/17/2021

Introduction

Dataset

The dataset is the famous movie rating dataset - Movielens, in particular it's 10M version, that consists of 10 millions ratings provided by users on movies.

Dataset is divided basically for training set - 9M ratings and validation set - 1M ratings.

Goal of the project

The goal of the project is to train the model that will properly predict the ratings for new pairs of users and ratings.

Method

The method used for the model training is collaborative filtering 2-rank matrix factorization. The analysis was performed manually without additional libraries used.

The closed form solution was used for minimizing the objective function. The details on formula provided by this link - https://towardsdatascience.com/evaluating-recommender-systems-root-means-squared-error-or-mean-absolute-error-1744abc2beac.

Firstly, the training set was divided again for regularization, choosing the best learning coefficient - lambda.

After we get the Lambda that works best for our training-validation set, the model was trained on the hole training set with the best Lambda parameter. And validated with the initial validation set.

Results

RMSE on validation set is approximately **0.835**. This result is considered as pretty solid.

Conclusion

Although the results of the given model are pretty good, the model can be improved even more.

Potential improvements: * Try higher rank matrix factorization * Consider Neural Networks to improve results