Movie Rating Prediction based on Netflix Prize Data

P01-Zhangqi Zha zzha@ncsu.edu
Gang Zhang gzhang22@ncsu.edu
Junhua Ma Jma20@ncsu.edu

1 Data Set

We chose the Netflix Prize data as our project data set. The data set gives 100 million records of the form "user X rated movie Y a 4.0 on 2/12/05". The data set is downloaded from kaggle website: https://www.kaggle.com/netflix-inc/netflix-prize-data.

2 Project idea

From the data set, we will try three types of methods: Neighborhood based method, matrix factorization based method and neural network based method.

For Neighborhood based method and matrix factorization based method, we will set standard kNN approach as our baseline (User-user approach, using PCA to perform dimension reduction), and try the SVD-inspired method to improve the performance. We will also try to ensemble the results of different models to achieve better result.

For neural network based method, we will first try a basic approach: feedforward neural network, with concatenated user embedding and movie embedding as input, and predicted rating as output. Then We will try to enhance our model by modifying the network structure or embedding method.

3 Software

Python and related open source libraries: numpy, scikit-learn, matplotlib, Tenforflow.

4 Papers to read

[1] Yehuda Koren, (2009) The BellKor Solution to the Netflix Grand Prize.

[2] Zheng, Yin, et al. A neural autoregressive approach to collaborative filtering

[3] Martin P. & Martin C. (2009) The Pragmatic Theory solution to the Netflix Grand Prize

5 Teammate and work division

Work will be equally among three of the team members. Preliminary analysis suggests that we can implement multiple classifiers. We will start to preprocess data together to study and explore the data, then try to apply various machine learning algorithms separately.

6 Midterm milestone

We will explore the data first, determine what kind of data prepossessing is needed and execute those steps. By midterm milestone, we will have the baseline experimental output(kNN and feedforward neural network) ready for comparison to other models. The Midway Report will be fully structured with most of the sections completed and a few 'place-holders' for the final results.