Technische Universiteit Delft
Faculty of Electrical Engineering, Mathematics and Computer Science

Netflix Challenge: Movie Rating Prediction

CSE-2525 Data Mining
Thomas Abeel, Gosia Migut

Prepared by
Yanqing Wu
Student ID 5142571
Kaggle ID yanqingwutudelft
Exchanged Computer Engineering
31 December 2019

Table of Contents

List of Figures iii
List of Tables iv
1 Introduction 1
1.1 Netflix Datasets 1
2 Methodology 1
2.1 Item-Item Collaborative Filtering 2
3 Results 2
4 Discussion 2
5 Conclusions and Future Work 2
References 3
Appendix A Title of First Appendix 4
Appendix B Another Appendix 5

List of Figures

List of Tables

Table 1-1 The Basic Information of Provided Data Sets 1

1 Introduction

The report, entitled ‘‘Netflix Challenge: Movie Rating Prediction’’, is prepared as my Challenge report for the course CSE2525-Data Mining at the Technische Universiteit Delft. The purpose of this report is to develop a recommendation system for predicting movie ratings. The goal of the recommendation system is to achieve Root Mean Square Error (RMSE) as small as possible on an unseen dataset.

1.1 Netflix Datasets

Table 1-1. The Basic Information of Provided Data Sets

Dataset	Features	Mean	Std	Min & Max
users	gender	0.72	0.45	0.00 1.00
	age	30.64	12.90	1.00 56.00
	profession	8.15	6.33	0.00 20.00
movies	year	1985.81	16.91	1919.00 2000.00
	title (string)	-	-	-
ratings	rating	3.58	1.12	1.00 5.00

In users - ‘gender’, ‘0’ and ‘1’ indicates female users and male users, respectively;
In movies - ‘year’, only non-zero entries are considered.

Three datasets are provided for training, as described in Table 1-1. There is a total of 910,190 ratings, which were given by 6,040 users and 3,706 movies. The rest ‘predictions.csv’ file is used for final testing, which contains only ‘userID‘ and ‘movieID’ for each entry.

2 Methodology

Data interpretation was first conducted at the beginning of the Challenge, as demonstrated in Table 1-1. After examining the data and reviewing the course materials, the first version of the recommendation algorithm was decided as Collaborative Filtering (CF). More specifically, Item-Item CF was selected as the first attempt to the Challenge. Item-Item CF is more reliable than User-User CF in practice [1], because items are much less dynamic than users.

2.1 Item-Item Collaborative Filtering

$$r_{xi} = b_{xi} + \frac{\sum_{j \in N(i;x)} S_{ij} \cdot (r_{xj} - b_{xj})}{\sum_{j \in N(i;x)} S_{ij}} \qquad (1)$$
$$b_{xi} = \mu + b_x + b_i$$

Equation (1) shows the core idea of the algorithm. Term r_{xi} represents the rating of user x on movie i. Term s_{ij} is the similarity of movie i and movie j, which is measured by cosine similarity. Term b_{xi} is the baseline estimator for r_{xi}, where μ is the average ratings of all movies, b_x is the rating deviation of user x and b_i is the rating deviation of movie i.

3 Results

Some more text.

4 Discussion

asdfasdfasdf

5 Conclusions and Future Work

From the analysis in the report body, it was concluded that...

References

[1] J. Leskovec, A. Rajaraman, and J. D. Ullman. *Mining of Massive Datasets*. New York: Cambridge University Press, 2009.

Appendix A Title of First Appendix

Use the No Spacing style.

Appendix B Another Appendix

Again, use the no spacing style for appendices.