Harvard Capstone Movielens

Manoj Bijoor

3/7/2021

Contents

List of Figures

List of Tables

1 Project Overview: MovieLens - A Harvard Capstone Project

A movie recommendation system using the MovieLens dataset.

For this project, I will be creating a movie recommendation system using the MovieLens dataset, provided by GroupLens Research, a research lab in the Department of Computer Science and Engineering at the University of Minnesota, Twin Cities specializing in recommender systems, online communities, mobile and ubiquitous technologies, digital libraries, and local geographic information systems.

GroupLens Research has collected and made available rating data sets from the MovieLens web site. The data sets were collected over various periods of time, depending on the size of the set.

I will use the 10M version of the MovieLens dataset to make the computation a little easier.

First, I will download the MovieLens data and run code provided to generate my datasets.

Second, I will train a machine learning algorithm using the inputs in one subset to predict movie ratings in the validation set.

1.1 Create Train and Final Hold-out Test Sets

I will develop my algorithm using the edx set. For a final test of my final algorithm, I predict movie ratings in the validation set (the final hold-out test set) as if they were unknown. RMSE will be used to evaluate how close my predictions are to the true values in the validation set (the final hold-out test set).

1.1.1 Important: Data sets usage

The validation data (the final hold-out test set) will NOT be used for training, developing, or selecting my algorithm and it will ONLY be used for evaluating the RMSE of my final algorithm. The final hold-out test set will only be used at the end of my project with my final model. It will not be used to test the RMSE of multiple models during model development. I will split the edx data into separate training and test sets to design and test my algorithm.

1.2 Final Product

1.2.1 My submission for this project is three files:

1. My report in Rmd format
2. My report in PDF format (knit from my Rmd file)
3. A script in R format that generates my predicted movie ratings and RMSE score (contains all code and comments for my project)

The report documents the analysis and presents the findings, along with supporting statistics and figures. The report assumes that the reader is not familiar with the project or the data. The report includes the RMSE generated and the following sections:
1. an introduction/overview/executive summary section that describes the dataset and summarizes the goal of the project and key steps that were performed 2. a methods/analysis section that explains the process and techniques used, including data cleaning, data exploration and visualization, insights gained, and my modeling approach 3. a results section that presents the modeling results and discusses the model performance 4. a conclusion section that gives a brief summary of the report, its limitations and future work

---

```
knitr::knit_exit()
```