Homework 4

MSCS 6520 Business Analytics
Spring 2018
Assigned: February 26, 2018
Due: March 5, 2017 (by beginning of class)

Readings

Introduction to *Recommender Systems* (on D2L)

The Million-Dollar Programming Challenge

Netflix Never Used its $1 Million Algorithm Due to Engineering Costs

Netflix Recommendations: Beyond the Stars (Part 1)

Netflix Recommendations: Beyond the Stars (Part 2)

Exercises

For the homework, we are going to explore using the average ratings for movies to predict the ratings given by individual users in the MovieLens data set. You will be replicating the analysis from class but using a different version of the MovieLen data set.

1. Download either the MovieLens + IMDb / Rotten Tomatoes from

   https://grouplens.org/datasets/hetrec-2011/

2. Read in the user_ratedmovies.dat and movies.dat files. These are tab-separated value (TSV) files so you will need to use read_tsv() function.
3. To join the two tibbles, we need to have a common column name. Use the mutate() and select() functions to rename the id column of the movies tibble to movieId so that it matches the movieId column of the user_ratedmovies tibble. Join the user_ratingsmovies and movies tibbles.
4. Compute the average and number of ratings by the MovieLens users.
5. Plot the distributions of the average MovieLens ratings as a histogram. What seem to be the most common ratings given to movies?
6. Plot the number of MovieLens ratings per movie (log scale) as densities. Does the number of ratings per movie seem uniformly distributed (each movie gets the same number of ratings) or skewed (some movies are rated more times than others)?
7. Plot the number of ratings (log scale) per movie vs the average rating per movie as a scatter plot. Is there a relationship between the two variables like we saw in the lecture? How do you interpret this relationship?
8. Build 4 models and evaluate their predictions using RMSE like we did in class:

   Baseline: use the overall average rating across all movies to predict the ratings by each user

Model 1: use the average rating per movie to predict the MovieLens ratings by each user

Models 2 – 4: use the average rating per movie, filtered by the number of ratings, to predict the MovieLens ratings by each user

Which model was best?

Prepare a document containing the answers to the above questions and plots. Submit the document as a PDF to D2L. You may work in pairs; in which case, you should only submit one PDF per group.