AM11 2022-23

Final Assessment

You get access to this movie rating dataset at https://grouplens.org/datasets/movielens/25m/. The dataset includes ratings of users to movies (0 to 5) through a long period of time. In addition there is one file that includes a description of each movie (in terms of genre) and another file that each user gives some description (tags) to the movie they rate.

Part 1

Step 1: Exploratory Analysis

(a) Build a histogram of show frequency (of ratings) of all movies
(b) Build a histogram of show frequency (of ratings) of all ~~movies~~ Users

Step 2: Data Engineering

(a) Select movies which have been viewed by at least 'm = 20' users
(b) Select users who have rated at least 'n = 50' movies

Step 3: Model Build

(a) Build three separate recommendation systems - (i) Item Based CF, (ii) User Based CF and (iii) a Model Based CF using Matrix Factorisation
(b) Try different values of ‘m’ and ‘n’ to see how that affects performance of your Recommendation System

Step 4: Report on Performance of model and parameter tuning

(a) Assess performance of the above models for different values of ‘m’ and ‘n’.
Hint: Try m <- c(10, 20, 50, 100, 200); m <- n; modelNames <- c(IBCF, UBCF, LIBMF); expand.grid(m,n,modelNames)
(b) Which model and combination of m & n provides the lowest RMSE?