Assignment Instructions: User-User Collaborative Filtering

Overview

In this assignment, you will implement user-user collaborative filtering using a spreadsheet. (We are using only basic spreadsheet operations that would work with Google's free Drive-based spreadsheet, Excel, or any other common spreadsheet program. We'll help you find the correct operations.)

Instructions

Part 1 - Without Normalization

First, you will implement user-user collaborative filtering without normalization.

1. Start by downloading the starting spreadsheet. This is a 25 user x 100 movie matrix of ratings selected from the class data set. The spreadsheet has three sheets in it (this is not supposed to be an exercise in spreadsheet tricks; as a result, we've already given you a significant start). 1) The first sheet is a ratings matrix with movies as rows and users as columns, 2) The second sheet is a ratings matrix with movies as columns and users as rows, and 3) The third sheet is the start of your correlations matrix.
2. Open the sample matrix in your favorite spreadsheet program. Note that the matrix contains a significant number of missing values -- do not replace these with zeroes, they are correctly missing.
3. Complete the user-by-user correlations matrix. To check your math, note that the correlation between users 1648 and 5136 is 0.40298, and the correlation between users 918 and 2824 is -0.31706. All correlations should be between -1 and 1, and the diagonal should be all 1's (since they are self-correlations).
4. Identify the top 5 neighbors (the users with the 5 largest, positive correlations) for users 3867 and 89. For example, if the target user were #3712, the closest neighbors are 2824 (corr: 0.46291), 3867 (corr: 0.400275), 5062 (corr: 0.247693), 442 (corr: 0.22713), and 3853 (corr: 0.19366). Don't forget to exclude the target user (corr: 1.0) from your possible selections.
5. Create a new worksheet in your spreadsheet, and use it to compute the predictions for each movie for users 3867 and 89 by taking the correlation-weighted average of the ratings of the top-five neighbors (for each target user) for each movie. The formal formula for correlation-weighted average is $\frac{\sum_{n=1}^{5} r_n w_n}{\sum^{5}}$. Remember, you will