Columbia GSB: Machine Learning
Homework 3

Dr. George A. Lentzas

This Homework is due by 11:59pm on Sunday March 11th. Submit your results in .pdf format in Canvas. Good Luck!

This homework is based on the Book-Crossing Dataset containing over 1 million book ratings.

1. Go to the Book-Crossing Database website `http://www2.informatik.uni-freiburg.de/~cziegler/BX/` (click on the hyperlink), download and import the BX-Book-Ratings dataset in R (make sure to acknowledge and reference appropriately in your paper). Use `table()`— and `order()`— to select the 100 "most active" (i.e. those with the most ratings) users. Then subset the full dataset to select the ratings of these most active users only. How many unique items/book ratings are there in this reduced data set? (20 points)

2. Split this reduced dataset into a "training set" consisting of $100,000$ observations and a "test set" using the `sample()`— function. Use `set.seed(1)`— before running the split so that you get the same results as everyone else. (5 points)

3. Convert your data into a ratings matrix. (10 points)

4. Apply the single Singular Value Decomposition (SVD) technique to build a recommendation system. Estimate the test error of your recommendation system using the test set. How well does the SVD recommendation system work? (30 points)

5. Apply the Iterative SVD technique (with 2 iterations) to potentially improve the performance of your recommendation system. Estimate the new test error and discuss your findings. Does the iterative recommendation system perform better? (25 points)

6. Given your analysis briefly discuss how you could improve the performance of the above recommendation system.

   (10 points)