Stat8056, Spring, 2022 - Homework, due 04/11/2022

Questions 1. (Recommender systems): Define user-item preference matrix: $R = [r_{ui}]_{U\times M}$, with rating score r_{ui}. When R is complete, denote $\Theta_{U\times M} = A_{U\times k}B_{k\times M}$, $\|\cdot\|_F$ is Frobenius norm, and $k \leq min\{U, M\}$. Let $\|\Theta\|_* = \sum_u \sigma_u$ be the nuclear norm and σ_u's be the singular values of Θ. Prove that

$$\arg\min_{A,B} \frac{1}{2}\|R - AB\|_F^2 + \lambda(\|A\|_F^2 + \|B\|_F^2)$$

is equivalent to

$$\arg\min_{\Theta} \frac{1}{2}\|R - \Theta\|_F^2 + \lambda\|\Theta\|_*,$$

where $\lambda > 0$ is a tuning parameter.

Question 2. (Movielens data): Movielens data are comprised of movie ratings and are available at http://grouplens.org/datasets/movielens/. The data was collected through the MovieLens website (movielens.umn.edu) during the seven-month from September 19th, 1997 through April 22nd, 1998. This data has been cleaned up - users who had less than 20 ratings or did not have complete demographic information were removed from this data set. Detailed descriptions of the data file can be found at the end of README.txt on the website. The 100K MovieLens data consists of 1,000,000 anonymous ratings on a five-star scale from 1,000 users on 1,700 movies. Here are four categorical covariates and one continuous covariate, including four user-related covariates, gender, age, occupation, and zip-code, as well as one content-related covariate, genres. For simplicity, we treat gender, age, and occupation as continuous variables, in addition to 19 different movie genres that are reparametrized into 19 binary covariates encoding if a certain movie belongs to a particular genre, where only the first digit is used for the zip-code.

1). Use any method of your choice (SVM, boosting, deep learning) to predict the preference scores of each user. There are five pairs of training and testing data sets for, denoted by (u1.base,u1.test),$\cdots$, (u5.base,u5.test) there. It is natural to use 5 fold cross-validation for training and testing. Compute the root mean square error based on cross-validation.

2). For this dataset, many ratings are not observed or missing. Do you have any evidence to suggest that missing does not occur at random? (Hint: Supply suitable plots to argue.) If the goal is to build a recommender system with high prediction accuracy, then demonstrate that the predictive performance can be enhanced by incorporating the missing pattern.