Homework 1, ST578, 2018
Due date: Feb. 20, 2018

MovieLens data are comprised of movie ratings. You can find different Movielens data sets at http://grouplens.org/datasets/movielens/.

For this homework, we use **MovieLens 100K Dataset**: http://grouplens.org/datasets/movielens/100k/. The data was collected through the MovieLens web site movielens.umn.edu during the seven-month period from September 19th, 1997 through April 22nd, 1998. It 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 README.txt on the website. The 100K MovieLens data consists of 100,000 anonymous ratings on a five-star scale from 1,000 users on 1,700 movies. There are four user-related covariates, including gender, age, occupation and zip code. We will treat age and zip code as continuous variables, and gender and occupation as categorical variables. There are 24 item-related covariates. The last 19 covariates stand for 19 different movie genres that are reparameterized into binary covariates encoding if certain movie belongs to a particular genre. Movies can be in several genres at once.

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

(b) 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: provide visualization tools to illustrate.)

(c) If the goal is to build a recommender system with a high accuracy of prediction, then demonstrate that the predictive performance can be enhanced by incorporating the missing pattern.

Please type your homework, write a summary report based on your finding, and attach your code for your homework. Email your homework to your TA Yubai Yuan: yubaiy2@illinois.edu.