

BSCB 694: Statistical Machine Learning  
**Project**

**The Netflix Prize**

The Netflix Prize has been announced as a general competition to predict user ratings of movies. Netflix has provided ratings of 17770 movie titles by 480189 users, along with the date of each rating. The task is to predict ratings for 282,000 user-movie-date triples that are not in the training set; all the users and movies in this test set appear in the training set. Netflix judges performance by mean squared error on the test set and has offered a \$1,000,000 reward for the first team that is able to improve the performance of their current system by more than 10%. Details of the Netflix Prize are available at

[www.netflixprize.com](http://www.netflixprize.com)

**Class Competition**

Because the Netflix Prize involves a very large data set and a non-standard problem (you could be asked to predict for any movie), the class competition will simplify the problem considerably. The training data (described below) provides the ratings of 10,000 users for 99 movies, along with the dates at which the ratings were made. The first 14 of these movies were rated by all users; the remaining 85 may have missing values. The outcome is the rating that each user gave to a further movie ("Miss Congeniality",2000); you are also given the date that this rating was made.

The task is to predict the rating for this movie by a further 2931 user in the test set. As with the training set, all users in the test set rated the first 14 movies, while the remaining 85 have missing values. The test set provides the same information as the training set – the dates and ratings of these 99 movies along with the date of the rating for "Miss Congeniality". As with the Netflix Prize, performance will be measured by squared error on the test set.

**Data Sets**

The data for the competition are available as tab-delimited text files on the class web-site. In particular, there you will find:

**train\_ratings\_all.dat** The ratings that the users in the training data set gave to each of the 99 movies.

**train\_dates\_all.dat** The date at which each of the ratings above were made.

**train\_ratings\_nomiss.dat** The training-set user ratings for the first 14 movies – ie, where there are no missing values.

**train\_dates\_nomiss.dat** The corresponding dates for **train\_dates\_nomiss.dat**.

**train\_y\_rating.dat** The ratings that the users in the training set gave to "Miss Congeniality".

**train\_y\_date.dat** The dates at which the training set users rated "Miss Congeniality".

**test\_ratings\_all.dat** The ratings that the users in the test data set gave to each of the 99 movies.

**test\_dates\_all.dat** The date at which each of the ratings above were made.

**test\_ratings\_nomiss.dat** The test-set user ratings for the first 14 movies – ie, where there are no missing values.

**test\_dates\_nomiss.dat** The corresponding dates for **test\_dates\_nomiss.dat**.

**test\_y\_date.dat** The dates at which the testing set users rated "Miss Congeniality".

**movie\_list.dat** Names and release dates for the 99 movies, given in the same order as the columns in the data above.

Some notes

- Ratings were from 1 to 5. A value of 0 indicates a missing entry.
- For convenience, dates are given as number of days from January 1, 1997.
- Missing dates are labeled '0000'.

### Rules and Procedures

1. You may work in groups. However, group work **MUST** be submitted with the names of all members of the group.
2. You may use any modeling technique you like, either parametric or non-parametric. You may not include information outside the data provided on the class web-site.
3. In order to make a submission send a text file containing a single column of predictions *in the same order as the test set* to [gjh27@cornell.edu](mailto:gjh27@cornell.edu). The names of the members of your group should be in the body of the e-mail.
4. You may make a submission at most once per week. Submissions will be evaluated on Friday after 12 noon. If you submit more than one set of predictions, only your last submission will be considered. I will respond with the mean squared error of your predictions on the test set.
5. The current best performance on the test set will be posted on the web-site. To start off, predicting the mean rating of "Miss Congeniality" on the training set yields a mean squared error of 0.9403.

6. Final results will be evaluated on the most recent submission as of class on Tuesday, November 27. You will be notified of your results by that evening.
7. The team with the best performance must give a brief (15min) description of their methods in class on Thursday, November 30.

### **Grades**

The project is worth 5% of the final grade. Your grade will be calculated from the following formula:

$$\frac{\text{squared error of best submission from the class}}{\text{squared error of your best submission}}$$

Good luck!