Statistical Learning, Fall 2014-15

Class Competition: Netflix Data

The Netflix Prize

The Netflix Prize was a competition to predict user ratings of movies. Netflix provided ratings of 17770 movie titles by 480189 users, along with the date of each rating. The task was to predict ratings for 282000 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 judged performance by root mean squared error on the test set and offered a $1,000,000 reward to the first team to improve the performance of their current system by more than 10%. The prize was won in 2009. Details of the Netflix Prize are available at:
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 provide 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 users 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 **root mean squared error** (RMSE) on the test set.

Data Sets

The data for the competition are available as tab-delimited text files on the class home directory http://www.tau.ac.il/~saharon/StatsLearn2014/ , they include:

- **train_ratings_all.dat**: The ratings that the users in the training data set gave to each of the 99 movies.
- **test_ratings_all.dat**: Same info for the test set.
- **train_dates_all.dat**: The date at which each of the ratings above were made.
- **test_dates_all.dat**: Same info for the test set.
- **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_y_date.dat**: Same info for the test set.
- **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 are 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 of up to three. Each student may be on one team only. Send me an email with group & member names.
2. You may use any modeling technique you like, either parametric or nonparametric.
3. **You may not include information outside the data provided on the class web-site** (no reverse engineering of Netflix data please).
4. 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 me by email. The file name must be: *<your group's name>_<date in format mmdd)>.txt*. (for example: *saharon_0319.txt*) **Submissions in incorrect format will not be evaluated** (or get a lousy score)**.**
5. **You may make a submission at most once per week** (from one Friday at noon to the next). Submissions will be evaluated on Friday after 12 noon. I will reply to your email with the RMSE of predictions on the test set.
6. The current best performance on the test set will be posted on the class homepage.
7. The competition will end on Friday, 16 January at noon. Each team's best score of the (at most) 10 they submitted will be their competition score.
8. The team with the best performance will give a brief (~15min) description of their methods in class on Monday, 19 January.

Grading

The competition is optional, and can only give a boost in your grade. There are three types of boost you can get:

- Any team that beats the simple linear regression performance on the test set of RMSE=0.7795 by more than 0.0095 and goes below RMSE=0.77 will get a boost of 5 points in grade.
- The winning team, after giving a brief but informative talk in class, will receive an additional 5 points boost in grade.
- Any team that beats a much tougher threshold, tentatively set at RMSE=0.755, will get another 5 points boost.

Some Points to Ponder

- Can we gain from treating the rankings as categorical? Ordinal?
- How should missing rankings be handled? Dummy variables?
- What would be a good distance measure for k-NN and related methods?
- Efficient k-NN approaches for this big data?
- How should dates be used? Can we use the release year of the movie?

Good luck!