

Netflix Prize DataSet Creation

Contributed by Jason Davis
Monday, 10 September 2007

The Netflix Prize competition has been going on for almost a year now. While it has been a great publicity event for Netflix, the competition is not very realistic. Netflix is trying to build a system to recommend movies to their users. These systems work by looking through users' past movie ratings to predict their future movie ratings. This notion of past and present is fundamentally missing in the Netflix competition.

The competition supplies three sets of recommendations: the training set, the probe set, and the test set. The goal of the competition is to train the recommendation system on the training set and then test the system on the test set. The ratings are not provided for the test set - the idea is that you submit your predictions to Netflix and then they can score your the quality of your system. The probe set is provided as a set that is similar to the test set yet also has movie rating scores - it allows participants to evaluate their ratings at home without having to submit their results first.

So how are these sets generated? One would expect that Netflix would have picked a date so all reviews before the date (past reviews) were put into the training set and reviews after the date (future reviews) were put into the test set (or the probe set). Unfortunately, this is not the case.

After realizing this, the first thing I did was figure out how Netflix created their data sets. The steps follow:

- Determine the 9 most recent movie ratings for each author and put them in the test set.
 - Randomly choose 1/3 of the ratings in the test set and move them to the probe set.
 - Place the remaining ratings (those not chosen in step 1) into the training set.
- So what is the significance of this? Well, a user's last 9 movie ratings can span anywhere from the last week to the last 6 years. Consequently, there are ratings in the training set from 2005 and ratings from 1999 in the test set. I talked with Charles Elkan, professor at UCSD and prize judge at NIPS last year, and this approach has some benefits from a research point of view. Nevertheless, training a recommendation system using reviews from 2005 and then evaluating on reviews from 1999 doesn't make much practical sense to me.