STAT 6838
Statistical research II

Thanh Doan

Overview about the problem

- Netflix is a movie rental business

- To help their customers find movies, they developed a movie recommendation system: **Cinematch**SM

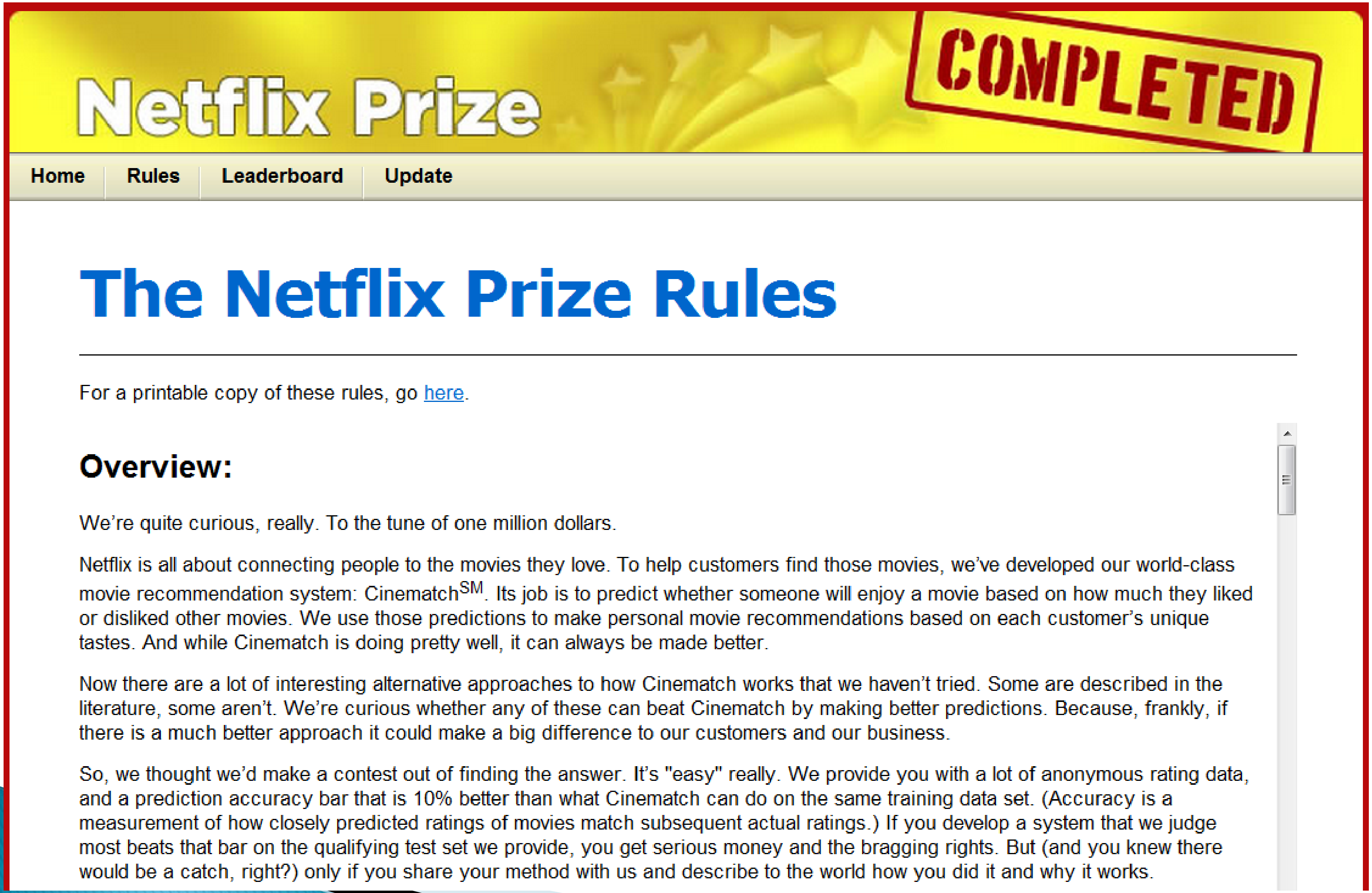
- Its job is to predict whether someone will enjoy a movie based on how much they liked or disliked other movies.

Overview about the problem

- Netflix uses those predictions to make personal movie recommendations based on each customer’s unique tastes.

- To see whether alternative approaches can beat **Cinematch**^{SM} by making better predictions, Netflix make a contest.

Netflix Prize



The Netflix Prize Rules

For a printable copy of these rules, go here.

Overview:

We’re quite curious, really. To the tune of one million dollars.

Netflix is all about connecting people to the movies they love. To help customers find those movies, we’ve developed our world-class movie recommendation system: Cinematch^{SM}. Its job is to predict whether someone will enjoy a movie based on how much they liked or disliked other movies. We use those predictions to make personal movie recommendations based on each customer’s unique tastes. And while Cinematch is doing pretty well, it can always be made better.

Now there are a lot of interesting alternative approaches to how Cinematch works that we haven’t tried. Some are described in the literature, some aren’t. We’re curious whether any of these can beat Cinematch by making better predictions. Because, frankly, if there is a much better approach it could make a big difference to our customers and our business.

So, we thought we’d make a contest out of finding the answer. It’s "easy" really. We provide you with a lot of anonymous rating data, and a prediction accuracy bar that is 10% better than what Cinematch can do on the same training data set. (Accuracy is a measurement of how closely predicted ratings of movies match subsequent actual ratings.) If you develop a system that we judge most beats that bar on the qualifying test set we provide, you get serious money and the bragging rights. But (and you knew there would be a catch, right?) only if you share your method with us and describe to the world how you did it and why it works.

Netflix Prize

- Netflix provide movie rating data and a prediction accuracy bar that is 10% better than what Cinematch can do on the same training data set
  - Accuracy is a measurement of how closely predicted ratings of movies match subsequent actual ratings.
- If someone develop a system that beats that bar on the qualifying test set, they get $1M Grand Prize
  - Only if they non-exclusively license the solution to Netflix and describe to the world how they did it and why it works

Netflix datasets

- Include a training set and qualifying test sets

- Training data consist of 100,480,507 ratings from 480,189 users on 17,770 movies

- Each training rating is of the form:

  <user, movie, date of rating, rating>

- For each movie, title and year of release are provided in a separate dataset

Netflix datasets

- The qualifying data set contains 2,817,131 triplets of the form

  <user, movie, date of rating>

  - With ratings known only to the jury.

Netflix datasets

- A participating algorithm must predict ratings on the entire qualifying set, but they are only informed of the score for half of the data, the quiz set of 1,408,342 ratings.

- The other half is the test set of 1,408,789, and performance on this is used by the jury to determine potential prize winners.

- This arrangement is intended to make it difficult to hill climb on the test set

Measuring prediction accuracy

- Submitted predictions are scored against the true ratings in terms of root mean squared error (RMSE), and the goal is to reduce this error as much as possible

- The RMSE measure is the square root of the averaged squared difference between ***each prediction*** and the actual rating in each subset, rounded to the nearest .0001

Cinematch^{SM} prediction accuracy bar

- **Cinematch**^{SM} scores an RMSE of 0.9514 on the quiz data. It has a similar performance on the test set, 0.9525
- In order to win the grand prize a participating algorithm had to improve this by 10%
  - To achieve 0.8572 on the test set
  - Such an improvement on the quiz set corresponds to an RMSE of 0.8563

My research objectives (1)

- Learn statistical methods in building recommendation systems
  - Read chapter 9 of the book http://infolab.stanford.edu/~ullman/mmds/ch9.pdf
  - Read papers published by the wining teams of the Netflix Grand Prize
- Write software to implement algorithms published by winners of the Netflix Prize
  - Build scalable recommendation system that run on Map-reduced parallel systems (on top of Hadoop)

My research objectives (2)

- Give it a shot. Try to improve the prediction accuracy further
  - Measuring the confidence of the predictions
  - Measuring the coverage of predictions
    - i.e. find which titles a system can make a good prediction
  - Apply different (or own) methods/algorithms
  - Incorporate new input features into the training set (such as other data about the movies, like genres, actors, directors, countries, etc)

References

- http://www.netflixprize.com/
- http://www2.research.att.com/~volinsky/netflix/bpc.html
- http://en.wikipedia.org/wiki/Netflix_Prize