Lessons from the Netflix Prize: Going beyond the algorithms

Yehuda Koren

Yahoo! Research
Haifa
We Know What You Ought To Be Watching This Summer
Welcome!

The Netflix Prize seeks to substantially improve the accuracy of predictions about how much someone is going to love a movie based on their movie preferences. Improve it enough and you win one (or more) Prizes. Winning the Netflix Prize improves our ability to connect people to the movies they love.

Read the Rules to see what is required to win the Prizes. If you are interested in joining the quest, you should register a team.

You should also read the frequently-asked questions about the Prize. And check out how various teams are doing on the Leaderboard.

Good luck and thanks for helping!
“We’re quite curious, really. To the tune of one million dollars.” – Netflix Prize rules

- Goal to improve on Netflix’ existing movie recommendation technology, Cinematch
- Criterion: reduction in root mean squared error (RMSE)
- Oct’06: Contest began
- Oct’07: $50K progress prize for 8.43% improvement
- Oct’08: $50K progress prize for 9.44% improvement
- Sept’09: $1 million grand prize for 10.06% improvement
### Movie rating data

#### Training data
- **100 million** ratings
- **480,000** users
- **17,770** movies
- **6** years of data: 2000-2005

<table>
<thead>
<tr>
<th>user</th>
<th>movie</th>
<th>score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>21</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>213</td>
<td>5</td>
</tr>
<tr>
<td>2</td>
<td>345</td>
<td>4</td>
</tr>
<tr>
<td>2</td>
<td>123</td>
<td>4</td>
</tr>
<tr>
<td>2</td>
<td>768</td>
<td>3</td>
</tr>
<tr>
<td>3</td>
<td>76</td>
<td>5</td>
</tr>
<tr>
<td>4</td>
<td>45</td>
<td>4</td>
</tr>
<tr>
<td>5</td>
<td>568</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>342</td>
<td>2</td>
</tr>
<tr>
<td>5</td>
<td>234</td>
<td>2</td>
</tr>
<tr>
<td>6</td>
<td>76</td>
<td>5</td>
</tr>
<tr>
<td>6</td>
<td>56</td>
<td>4</td>
</tr>
</tbody>
</table>

#### Test data
- Last few ratings of each user (**2.8 million**)

<table>
<thead>
<tr>
<th>user</th>
<th>movie</th>
<th>score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>62</td>
<td>?</td>
</tr>
<tr>
<td>1</td>
<td>96</td>
<td>?</td>
</tr>
<tr>
<td>2</td>
<td>7</td>
<td>?</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>?</td>
</tr>
<tr>
<td>3</td>
<td>47</td>
<td>?</td>
</tr>
<tr>
<td>3</td>
<td>15</td>
<td>?</td>
</tr>
<tr>
<td>4</td>
<td>41</td>
<td>?</td>
</tr>
<tr>
<td>4</td>
<td>28</td>
<td>?</td>
</tr>
<tr>
<td>5</td>
<td>93</td>
<td>?</td>
</tr>
<tr>
<td>5</td>
<td>74</td>
<td>?</td>
</tr>
<tr>
<td>6</td>
<td>69</td>
<td>?</td>
</tr>
<tr>
<td>6</td>
<td>83</td>
<td>?</td>
</tr>
</tbody>
</table>

Dates of ratings are given.
Data >> Models

- Very limited feature set
  - User, movie, date
  - Places focus on models/algorithms

- Major steps forward associated with incorporating new data features
  - Temporal effects
  - Selection bias:
    - What movies a user rated
    - Daily rating counts
Multiple sources of temporal dynamics

- **Item-side effects:**
  - Product perception and popularity are constantly changing
  - Seasonal patterns influence items’ popularity

- **User-side effects:**
  - Customers ever redefine their taste
  - Transient, short-term bias; anchoring
  - Drifting rating scale
  - Change of rater within household
Something Happened in Early 2004…
Are movies getting better with time?
Temporal dynamics - challenges

• **Multiple effects:** Both items and users are changing over time
  ➔ Scarce data per target

• **Inter-related targets:** Signal needs to be shared among users – foundation of **collaborative** filtering
  ➔ cannot isolate multiple problems

➔ Common “concept drift” methodologies won’t hold. E.g., **underweighting older instances is unappealing**
Effect of daily rating counts

• Number of ratings user gave on the same day is an important indicator
• It affects different movies differently

Credit to:
Martin Piotte and Martin Chabbert
Memento vs Patch Adams

Memento (127318 samples)

Patch Adams (121769 samples)

Credit to:
Martin Piotte and Martin Chabbert
Why daily rating counts

- Number of user ratings on a date is a proxy for how long ago the movie was seen
  - Some movies age better than others
- Also, two rating tasks:
  - Seed Netflix recommendations
  - Rate movies as you see them
- Related to selection bias?
Biases matter!

Components of a rating predictor

- User bias
- Movie bias
- User-movie interaction

Baseline predictor
- Separates users and movies
- Often overlooked
- Benefits from insights into users’ behavior
- Among the main practical contributions of the competition

User-movie interaction
- Characterizes the matching between users and movies
- Attracts most research in the field
- Benefits from algorithmic and mathematical innovations
A baseline predictor

- We have expectations on the rating by user $u$ to movie $i$, even without estimating $u$’s attitude towards movies like $i$

- Rating scale of user $u$
- Values of other ratings user gave recently (day-specific mood, anchoring, multi-user accounts)
- (Recent) popularity of movie $i$
- Selection bias; related to number of ratings user gave on the same day
Sources of Variance in Netflix data

- Unexplained: 57%
- Biases: 33%
- Personalization: 10%

Unexplained: 0.732
Biases: 0.415
Personalization: 0.129

Total Variance: 1.276
What drives user preferences?

• Do they like certain genre, actors, director, keywords, etc.?
• Well, some do, but this is far from a complete characterization!
• E.g., a recent paper is titled:
  – “Recommending new movies: even a few ratings are more valuable than metadata” [Pilaszy and Tikk, 09]
• User motives are latent, barely interpretable in human language
• Can be captured when data is abundant
Wishful perception

The Color Purple
Sense and Sensibility
The Princess Diaries
The Lion King
Amadeus
Ocean’s 11
Independence Day
Dumb and Dumber
Lethal Weapon
Braveheart

Geared towards females
Geared towards males
serious
escapist
Complex reality...
Ratings are not given at random!

**Distribution of ratings**

- Netflix ratings
- Yahoo! music ratings
- Yahoo! survey answers

Marlin, Zemel, Roweis, Slaney, “Collaborative Filtering and the Missing at Random Assumption” UAI 2007
Which movies users rate?

- A powerful source of information:
  Characterize **users** by **which** movies they rated, rather than **how** they rated
- ➔ A dense binary representation of the data:

\[
R = \{ r_{ui} \}_{u,i}
\]

\[
B = \{ b_{ui} \}_{u,i}
\]
Ensembles are Valuable for Prediction

- Our final solution was a linear blend of over 700 prediction sets
  - Some of the 700 were blends
- Difficult, or impossible, to build a grand unified model
- Blending techniques: linear regression, neural network, gradient boosted decision trees, and more…
- Mega blends are not needed in practice
  - A handful of simple models achieves 90% of the improvement of the full blend
Yehuda Koren
Yahoo! Research
yehuda@yahoo-inc.com