Machine Learning in Microsoft’s Online Services: TrueSkill, AdPredictor, and Matchbox

Thore Graepel
Online Services and Advertising Research
Microsoft Research Cambridge, UK

Joint work with Thomas Borchert, Joaquin Quiñonero Candela, Ralf Herbrich, David Stern, and Tom Minka

Industry Day at ECML/PKDD 2010 - Barcelona
Outline

TrueSkill™: Ranking and Matchmaking

• Ranking and Matchmaking Task
• TrueSkill Model and Inference

AdPredictor™: Click-Through Rate Prediction

• Paid Search Advertising and Click-Through Rate prediction
• AdPredictor Model and Inference, and results

Matchbox: Large-Scale Recommendations

• Recommendation with Meta-Data
• Matchbox model and inference, and results
TrueSkill(TM): A Bayesian Skill Rating System,
• Pierre Dangauthier, Ralf Herbrich, Tom Minka, and Thore Graepel,
TrueSkill Through Time: Revisiting the History of Chess,

Thore Graepel, Ralf Herbrich, Tom Minka, and our friends from Xbox Live
Ranking and Matchmaking

• Competition is central to our lives
  – Evolutionary principle
  – Driving principle of many sports

• Chess Rating for fair competition
  – ELO: Developed in 1960 by Árpád Imre Élő
  – Matchmaking system for tournaments

• Challenges of online gaming
  – Learn from few match outcomes efficiently
  – Support multiple teams and multiple players per team
The Skill Rating Problem

- Given:
  - Match outcomes: Orderings among $k$ teams consisting of $n$ matches respectively.

- Questions:
  - Skill $s_i$ for each team.
  - Global skill.
  - Fairness of the results.
Multiple Team Match Outcome Model

- Observed outcome $(1,2,3)$, use transitivity!
- Skill posterior $P(s_i | y = (1,2,3)) \approx N(\mu_i, \sigma_i^2)$

$$y_{12} = (1,2) \quad y_{23} = (2,3)$$
Efficient Approximate Inference

Gaussian Prior Factors

Ranking Likelihood Factors
Applications to Online Gaming

• Leaderboard
  – Global ranking of all players
  – Rank conservatively according to $\mu_i - 3 \cdot \sigma_i$

• Matchmaking
  – For gamers: Most uncertain outcome
  – For inference: Most informative
  – Equivalent!
Convergence Speed

- char (TrueSkill™)
- SQLWildman (TrueSkill™)
- char (Halo 2 rank)
- SQLWildman (Halo 2 rank)
Xbox 360 & Halo 3

• **Xbox 360 Live**
  – Launched in September 2005
  – Every game uses TrueSkill™ to match players
  – > 20 million players
  – > 2 million matches per day
  – > 2 billion hours of game play

• **Halo 3**
  – Launched on 25th September 2007
  – Largest entertainment launch in history
  – > 200,000 player concurrently (peak: 1,000,000)
Halo 3 in Action
Halo 3 Analysis: Fair Matches?
Great stuff, guys, but how about online advertising? Are the ads not competing for the attention of the users just like players in a game?!

Why are they wasting their time on computer games?
Thore Graepel, Joaquin Quiñonero Candela, Thomas Borchert, and Ralf Herbrich, and Ralf Herbrich,
Web-Scale Bayesian Click-Through Rate Prediction for Sponsored Search Advertising in Microsoft’s Bing Search Engine, in ICML 2010, Invited Applications Track, June 2010
Importance of accurate probability estimates:

- Increase user satisfaction by better targeting
- Provide better deal to advertisers
- Increase revenue by showing ads with high click-thru rate
AdPredictor: Bayesian Probit Regression

Client IP:
- 102.34.12.201
- 15.70.165.9
- 221.98.2.187
- 92.154.3.86

Match Type:
- Exact Match
- Broad Match

Position:
- ML-1
- SB-1
- SB-2

Impression Level Click-Through Rate Prediction

\[ p(p\text{Click}) \]
Closed Form Updates for “Click”

\[
\mu_i \leftarrow \mu_i + \frac{\sigma_i^2}{s} \cdot h \left[ \sum_{j=1}^{d} \mu_j \right] \\
\sigma_i^2 \leftarrow \sigma_i^2 \left( 1 - \frac{\sigma_i^2}{s^2} \right) \cdot g \left[ \sum_{j=1}^{d} \mu_j \right]
\]

\[
s^2 = \beta^2 + \sum_{j=1}^{d} \sigma_j^2
\]

\[
h(t) = \frac{\mathcal{N}(t; 0, 1)}{\Phi(t)}
\]

\[
g(t) = h(t) \cdot [h(t) + t]
\]
Weight Parameters from Production

- Means and variances of the weights for display position
- Mainline ML, sidebar SB
- High mean $\rightarrow$ High CTR
- Low variance $\rightarrow$ Shown often

- Means and variances of weights for user ID feature.
- Initialisation at zero mean unit variance (top)
- Very high mean $\rightarrow$ fraud/bots
Comparison with Naïve Bayes

- **Calibration**: Ratio of Empirical CTR / Predicted CTR
- Horizontal line at 1 is perfect calibration

### RIG: Relative Information Gain

\[
RIG := \frac{CE + H(\bar{p})}{H(\bar{p})}
\]

- **AUC**: Area under the Receiver-Operator Curve

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>RIG</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>adPredictor</td>
<td>61.42%</td>
<td>95.6%</td>
</tr>
<tr>
<td>adPredictor (calibrated)</td>
<td>61.35%</td>
<td>95.6%</td>
</tr>
<tr>
<td>Naïve Bayes</td>
<td>-41.54%</td>
<td>89.4%</td>
</tr>
<tr>
<td>Naïve Bayes (calibrated)</td>
<td>33.86%</td>
<td>89.3%</td>
</tr>
</tbody>
</table>
Web Scale Implementation

AdPredictor drives almost 100% of Bing’s Sponsored Search traffic

<table>
<thead>
<tr>
<th>Large volume of data $O(10^9)$ examples per week</th>
<th>System’s predictions determine composition of future training sample</th>
<th>Continuous Model Training → Maintain controlled memory footprint by pruning</th>
</tr>
</thead>
</table>
Thore Graepel, David Stern, Ralf Herbrich and the Emporia team

MATCHBOX

• David Stern, Ralf Herbrich, and Thore Graepel,
  Matchbox: Large Scale Bayesian Recommendations,
Collaborative Filtering

Users

A
B
C
D

Items

1
2
3
4
5
6

Metadata?
Map Sparse Features To ‘Trait’ Space

User ID
- 234566
- 456457
- 13456
- 654777

Gender
- Male
- Female

Country
- UK
- USA

Height
- 1.2m

Item ID
- 34
- 345
- 64
- 5474

Movie Genre
- Horror
- Drama
- Comedy
- Documentary
Matchbox With Metadata

User\[s = Ux\]

Item\[t =Vy\]

Rating potential\[\sim \mathcal{N}(s^\top t, \beta^2)\]
Message Passing For Matchbox
Message Passing For Matchbox

Message update functions powered by Infer.net
User/Item Trait Space

-1.5 -1 -0.5 0 0.5 1 1.5

Users
Movies

24: Season 3
Adaptation
24: Season 2
A Clockwork Orange
AI: Artificial Intelligence
A Knights Tale
A Cinderella Story

‘Preference Cone’ for user 145035
Feedback Models
Feedback Models
Feedback Models
### MovieLens – 1,000,000 ratings

**6,040 users**

<table>
<thead>
<tr>
<th>User ID</th>
<th>User Job</th>
<th>User Age</th>
<th>User Gender</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Other</td>
<td>&lt;18</td>
<td>Male</td>
</tr>
<tr>
<td></td>
<td>Academic Programmer</td>
<td>18-25</td>
<td>Male</td>
</tr>
<tr>
<td></td>
<td>Artist</td>
<td>25-34</td>
<td>Female</td>
</tr>
<tr>
<td></td>
<td>Admin</td>
<td>35-44</td>
<td>Male</td>
</tr>
<tr>
<td></td>
<td>Student</td>
<td>45-49</td>
<td>Male</td>
</tr>
<tr>
<td></td>
<td>Customer Service</td>
<td>50-55</td>
<td>Male</td>
</tr>
<tr>
<td></td>
<td>Health Care</td>
<td>&gt;55</td>
<td>Male</td>
</tr>
<tr>
<td></td>
<td>Managerial</td>
<td>&lt;18</td>
<td>Female</td>
</tr>
<tr>
<td></td>
<td>Farmer</td>
<td>18-25</td>
<td>Female</td>
</tr>
<tr>
<td></td>
<td>Homemaker</td>
<td>25-34</td>
<td>Female</td>
</tr>
<tr>
<td></td>
<td></td>
<td>35-44</td>
<td>Female</td>
</tr>
<tr>
<td></td>
<td></td>
<td>45-49</td>
<td>Female</td>
</tr>
<tr>
<td></td>
<td></td>
<td>50-55</td>
<td>Female</td>
</tr>
<tr>
<td></td>
<td></td>
<td>&gt;55</td>
<td>Female</td>
</tr>
</tbody>
</table>

**3,900 movies**

<table>
<thead>
<tr>
<th>Movie ID</th>
<th>Movie Genre</th>
</tr>
</thead>
<tbody>
<tr>
<td>Action</td>
<td>Horror</td>
</tr>
<tr>
<td>Adventure</td>
<td>Musical</td>
</tr>
<tr>
<td>Animation</td>
<td>Mystery</td>
</tr>
<tr>
<td>Children’s</td>
<td>Romance</td>
</tr>
<tr>
<td>Comedy</td>
<td>Thriller</td>
</tr>
<tr>
<td>Crime</td>
<td>Sci-Fi</td>
</tr>
<tr>
<td>Documentary</td>
<td>War</td>
</tr>
<tr>
<td>Drama</td>
<td>Western</td>
</tr>
<tr>
<td>Fantasy</td>
<td>Film Noir</td>
</tr>
</tbody>
</table>
MovieLens with Thresholds Model

(ADF), Training Time = 1 Minute

Mean Absolute Error

<table>
<thead>
<tr>
<th>K</th>
<th>MetaData Off</th>
<th>MetaData On</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.67</td>
<td>0.67</td>
</tr>
<tr>
<td>2</td>
<td>0.67</td>
<td>0.67</td>
</tr>
<tr>
<td>5</td>
<td>0.67</td>
<td>0.67</td>
</tr>
<tr>
<td>10</td>
<td>0.67</td>
<td>0.67</td>
</tr>
<tr>
<td>20</td>
<td>0.67</td>
<td>0.67</td>
</tr>
</tbody>
</table>

---

*Lam et al.*
Applications

- Ranking of content on web portals
- Online advertising (Display and Paid Search)
- Personalised web search
- Algorithm portfolio management
- Tweet/News recommendation
- Friends recommendation on social platforms
Welcome to Emporia

Emporia determines the TOP stories being shared NOW on Twitter and categorizes them so you can read, vote, or share them with friends.

1. Click on a story on the left to read more
2. Share a story via email, Facebook or Twitter
3. Vote* and get more relevant top stories to match your taste
4. Click refresh to view the latest stories and apply your votes

* You'll get prompted to sign-in so we can save your votes.
Conclusions

Probabilistic Graphical Models Work at Scale!

- **Single pass learning** benefits from explicit model of uncertainty
- The models make **accurate and calibrated** predictions
- They are **efficient** if we use problem-specific independence structure
- They are **modular** and can be composed into more powerful models

Wide Range of Applications

- Skill estimation in online games → **TrueSkill**
- Click-through rate estimation in Paid Search → **AdPredictor**
- Large scale recommendations of content → **Matchbox / Emporia**
- Hopefully many more to come...
Thank you!

thoreg@microsoft.com