Probabilistic Machine Learning in Computational Advertising

Thore Graepel, Thomas Borchert, Ralf Herbrich and Joaquin Quiñonero Candela

Online Services and Advertising
Microsoft Research Cambridge, UK

NIPS 2009 – December 2009
Outline

• Online Advertising and Paid Search
• AdPredictor™: Predicting User Clicks on Ads
  [Appendix]
• Model shrinking
• Parallel training
ONLINE ADVERTISING AND PAID SEARCH
Advertising Industry Business: Size

Annual Expenditure (in billion USD)

Year

2001 2002 2003 2004 2005 2006

Outdoor Cinema Radio TV Print Online

GDP Denmark (2006)

Microsoft Revenue (2008)

Advertising Industry Business: Growth

[Chart showing growth percentages for different advertising mediums from 2001 to 2006.]

Display to users (expected bid):

\[ b_1 \cdot p_1 \geq b_2 \cdot p_2 \geq \cdots \]

Charge advertisers (per click):

\[ c_i = \frac{b_{i+1} \cdot p_{i+1}}{p_i} \]

Importance of accurate probability estimates:

- Increase user satisfaction by better targeting
- Fairer charges to advertisers
- Increase revenue by showing ads with high click-thru rate
The Scale of Things

• **Realistic training set for proof of concept:**
  
  7,000,000,000 impressions

• **2 weeks of CPU time during training:**
  
  2 wks × 7 days × 86,400 sec/day = 1,209,600 seconds

• **Learning algorithm speed requirement:**
  
  – 5,787 impression updates / sec
  – 172.8 μs per impression update
Bayesian Linear Probit Regression

ADPREDICTOR
Impression Level Predictions
One Weight per Feature Value

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

pClick
Click Potential

Linear: 
\[ \text{click potential} = \text{sum of feature click contributions} \]
Gaussian Noise

Probit: area under Gaussian tail as a function of click potential
**Probit**: area under Gaussian tail as a function of click potential

\[ P(\text{click}) = \Phi(\text{click potential}) \]
Modelling Uncertainty

ClientIP = 98.0.101.23

ListingId = 798831

PageNumber/DisplayPosition/ReturnedAds = 0/ML-1/2
Uncertainty about the Potential
Probability of Click

\[ P(\text{click}) = \Phi \left( \frac{\sum_{j=1}^{d} \mu_j}{\sqrt{\beta^2 + \sum_{j=1}^{d} \sigma_j^2}} \right) \]
Principled Exploration

average: 25% (3 clicks out of 12 impressions)

average: 30% (30 clicks out of 100 impressions)
Training Algorithm in Action

No Click

Prediction

Training/Update
Posterior Updates for the Click Event

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

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

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

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

\[
g(t) = h(t) \cdot [h(t) + t]
\]
Client IP: Mean & Variance
Calibrated Predictions

adPredictor Predicted vs. Empirical CTR
(Flight 153 from Nov 11 to Nov 17 2008)
Joint Updates vs. Independent Aggregation

Actual CTR vs. Predicted CTR

- adPredictor (blue)
- Naive Bayes (red)

The graph shows the comparison between actual and predicted click-through rates (CTR) for Joint Updates and Independent Aggregation methods.
adPredictor Wrap Up

- Automatic learning rate
- Calibrated: 2% prediction means 2% clicks
- Use of many features, even if correlated
- Modelling the uncertainty explicitly
- Natural exploration mode
- Parallelizable with approximate inference
Thank you!

thoreg@microsoft.com
Dealing with Millions of Variables

- Observation 1: Large variable bags follow a power-law w.r.t. frequency of items
- Observation 2: Weight posteriors of rare items are close to their prior

Idea:
1. Initially, the belief of each new item is compactly represented by one (and the same) prior
2. After observing an item for the first time, the posterior is allocated
3. At regular intervals, all weight posteriors with a small deviation from the prior are removed
Naïve Approach – Shared Memory

- Does not scale
  - Constant contention for locks
  - Some features are very frequent
  - Synchronization issues
Proposal: Approximate Learning

Train Node 1

Impression A

10.0.0.1
USA
MSNH1

(etc)

10.0.1.2
Canada

Update

Merge Deltas

Final Model File

Train Node 2

Impression B

MSNH1

Canada

10.0.1.2

(etc)

Update

Update

Update

Update

Update