Click Prediction with adPredictor at Microsoft Advertising

Joaquin Quiñonero Candela
Thore Graepel
Ralf Herbrich
Thomas Borchert

Microsoft Research & Microsoft adCenter
Microsoft + Yahoo! = 1/3 US search market

adPredictor predicts probability of click on ads for Microsoft Bing and Yahoo! search engines
flowers
Results for "flowers" on Bing:

FTD® - flowers

Flowers $19.99+ Free Vase

Free Vase & Delivery Inc.
Flowers.ms/flowers - Delivery Included. FTD Florist Member Satisfaction Guaranteed

Flowers, Roses, Gift Baskets, Same Day Florist | 1-800-FLOWERS.COM
Order flowers, roses, gift baskets and more. Get same-day flower delivery for birthdays, anniversaries, and all other occasions. Find fresh flowers at 1800Flowers.com.
www.1800flowers.com - Cached page

FTD.COM - Flowers Online | Roses, Fresh Flowers, Plants and Gift...
Order flowers online for same day delivery. Shop for flowers and gifts by occasion, season or get beautiful flower bouquets delivered same day by local FTD florists.
www.ftd.com - Cached page

Send Flowers and Gifts with ProFlowers – Fresh Flowers Delivered
Order flowers online with ProFlowers, offering nationwide flower delivery. Choose from a variety of beautiful arrangements including Christmas flowers.
www.proflowers.com - Cached page

Flower - Wikipedia, the free encyclopedia
Flower specialization ... · Morphology · Development · Pollination
In those species that have more than one flower on an axis—so-called composite flowers—the collection of flowers is termed an inflorescence; this term can also refer to the ... en.wikipedia.org/wiki/Flower - Cached page
Importance of accurate probability estimates

- Efficient use of ad space
- Increased user satisfaction by better targeting
- Increased revenue by showing ads with high click-thru rate

Over-simplified ranking function: this is not what is used in practice
Impression Level Predictions

- Sparse binary input features (many 10s of them)
- Some high cardinality (~100M), some low (<10)
## Sparse Linear Probit Regression

<table>
<thead>
<tr>
<th>Ad ID</th>
<th>1341201</th>
<th>1570165</th>
<th>2213187</th>
<th>9215433</th>
</tr>
</thead>
<tbody>
<tr>
<td>Match Type</td>
<td>Exact Match</td>
<td>Broad Match</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Position</td>
<td>ML-1</td>
<td>SB-1</td>
<td>SB-2</td>
<td></td>
</tr>
</tbody>
</table>

\[
p_{\text{Click}} = \text{ML-1} + \text{SB-1} + \text{SB-2} + \text{Exact Match} + \text{Broad Match}
\]
Uncertainty: A Bayesian Treatment

<table>
<thead>
<tr>
<th>Ad ID</th>
<th>Match Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>1341201</td>
<td>Exact Match</td>
</tr>
<tr>
<td>1570165</td>
<td>Exact Match</td>
</tr>
<tr>
<td>2213187</td>
<td>Broad Match</td>
</tr>
<tr>
<td>9215433</td>
<td>Broad Match</td>
</tr>
</tbody>
</table>

\[ p(p_{\text{Click}}) \]
A Linear Probit Model

• Notation
  \[ y = 1 \text{ if click} \]
  \[ y = -1 \text{ if non-click} \]
  \[ w \] is the vector of all weights
  \[ x \] is a sparse binary input vector

• Generalised linear model with weights vector \( w \):
  \[ p(y|x, w) = \Phi \left( \frac{y \cdot w^T x}{\beta} \right) \]

• Inverse link function is the probit function:
  \[ \Phi(t) := \int_{-\infty}^{t} \mathcal{N}(z; 0,1) \, dz \]

\( \beta \) controls the steepness: it corresponds to the standard deviation of additive zero mean noise.
Think of $x$ as indicator variables that select weights: we will soon remove $x$ from the notation
Example = $x = [1; 0; 0; 0; 1; 0; ...; 0; 1]$
Uncertainty About the Weights
A Bayesian Treatment

• Factorizing Gaussian prior over the weights:

\[
p(w) = \prod_{i=1}^{N} \mathcal{N}(w_i; \mu_i, \sigma_i^2)
\]

• Given \( p(y|x, w) \) the posterior is given by:

\[
p(w|x, y) = \frac{p(y|x, w) \cdot p(w)}{\int p(y|x, w) \cdot p(w) \cdot dw}
\]

**Problem:** This posterior cannot be represented compactly nor calculated in closed form
Desiderata and Approximations

We want

– The posterior to remain a factorized Gaussian
– Incremental online learning rather than batch

This is how it is done

– Approximate inference with latent variables
– Single pass approximate (online) schedule
Predicting Average Probability of Click

Now that our posterior over the weights is a factorizing Gaussian...

\[ p(y|x) = \Phi \left( \frac{y \cdot \sum_{i=1}^{N} \mu_i}{\sqrt{\beta^2 + \sum_{i=1}^{N} \sigma_i^2}} \right) \]
Principled Exploration

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

average: 30% (30 clicks out of 100 impressions)
Approximate Inference with Latent Variables

- Prior: \( f_i(w_i) = \mathcal{N}(w_i; \mu_i, \sigma_i^2) \)
- Sum of active weights:
  \[
  u(s, \{w_i\}) = \delta(s - \sum_{i=1}^{N} w_i)
  \]
- Noisy version thereof:
  \[
  v(s, t) = \mathcal{N}(t; s, \beta^2)
  \]
- The sign of \( t \) determines click:
  \[
  q(t, y) = \delta(y - \text{sign}(t))
  \]
Approximating $p(t)$ and $m_{q\rightarrow t}(t)$

$m_{v\rightarrow t}(t) \ast m_{q\rightarrow t}(t) = p(t)$

$\hat{m}_{v\rightarrow t}(t) \ast \hat{m}_{q\rightarrow t}(t) = \hat{p}(t)$
Updating the Posterior

\[ w_1 + w_2 \]

\[ y \]

Prediction
Training/Update
Posterior Updates for the Click Event

$$\mu_i \leftarrow \mu_i + \frac{\sigma^2_i s}{s} \cdot h$$

$$\sigma^2_i \leftarrow \sigma^2_i \left(1 - \frac{\sigma^2_i s^2}{s} \cdot g\right)$$

$$s^2 = \beta^2 + \sum_{j=1}^{d} \sigma^2_j$$

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

$g(t) = h(t) \cdot [h(t) + t]$
The importance of joint updates

[Graph: Comparison of actual and predicted CTR for adPredictor and Naive Bayes]
Calibration by Isotonic Regression

Calibrated adPredictor

Calibrated Naive Bayes

Actual CTR vs. Predicted CTR graphs for adPredictor and Naive Bayes.
Calibration Can’t Improve the ROC

![ROC Curve]

**True Positives** vs **False Positives**

- Blue line: Naïve Bayes
- Red line: adPredictor
adPredictor Wrap Up

Automatic learning rate

Calibrated: 2% prediction means 2% clicks

Use of very many features, even if correlated

Modelling the uncertainty explicitly

Natural exploration mode
Discussion (For Later)

- Sample selection bias and exploration
- Dynamics: forgetting with time
- Pruning uninformative weights
- Approximate parallel inference
- Hierarchical priors
- Input features... the secret sauce

Some of this is detailed in the ICML 2010 paper:

*Web-Scale Bayesian Click-Through Rate Prediction for Sponsored Search Advertising in Microsoft’s Bing Search Engine*

We are hiring! Please contact me if you are interested.
Thank you!

joaquinc@microsoft.com

We are hiring! Please contact me if you are interested.