Developing Web-scale Machine Learning at LinkedIn - from Soup to Nuts

Kun Liu

December 13, 2014
Online Recommendations @ LinkedIn

- **Organic update**
  - Todd Beaupré likes this
  - Stagnant wages linger as a problem in the United States. A look at where wages have grown the most - and least - in the past 10 years.

- **Sponsored content**
  - Mashery shared:
    - Sponsored
    - Find success at your enterprise with private API programs. More about deploying and leveraging an API management strategy that aligns with your company’s business goals.

- **Recommended news**
  - Stories you can't miss today on LinkedIn Pulse
    - When The Disruption Hit: New Republic
    - Reid Hoffman Interview: Bill Gurley
    - Like + Comment + Share
    - Google Will Acquire a Logistics Company, And Other Predictions for 2015
    - Adrian Gonzalez President at Adelante SCM & Founder/Host...
    - Like + Comment + Share

- **Recommended jobs**
  - Jobs you may be interested in
  - Senior Relevance/Machine
    - Twitter — San Francisco Bay Area
  - Data Scientist - Adobe Digital Marketing
    - Adobe — San Francisco Bay Area
  - Data Scientist
    - Adobe — San Jose, CA

- **Recommended companies & groups**
  - Companies You May Want To Follow
    - IBM
    - FischerCunnane
    - BRAYN Consulting LLC
  - Groups You May Like
    - CPA Tech Connect
      - Join - Corporate Group
    - CPAs
      - Join - Professional Group
    - BRAYN Consulting LLC
      - Join - Networking Group
Many Practical Challenges, but

- Fast iteration is desired
  - large-scale machine-learning framework

- Models should adapt to large dynamic data
  - cold start model + warm start model + explore/exploit

- Offline metrics do not always reflect online
  - online A/B test

- Real-time feedback is important
  - near real-time event stream
The 30,000-foot Overview

Data Pipeline
- feature extraction
- feature transformation
- user modeling

Model Fitting Pipeline (Hadoop)
- offline modeling fitting (cold-start model)
- nearline modeling fitting (warm-start model)
  - daily/weekly
  - hourly

Online Serving System
- multi-pass rankers
- candidates generation
- real-time feedback
- online A/B test model evaluation
Fast Iteration is Desired
Fast Iteration is Desired

- Every machine learning task is different
  - still the steps to solving the problem are often quite similar

- Fast iteration is desired
  - successful systems require lots of tuning & experimentation
  - reusable modules and easy-to-config workflows dramatically improve productivity
  - free engineers to concentrate on unique aspects of the project
Large-scale Machine-learning Framework at LinkedIn

scheduled workflow on Azkaban

feature, target sources
feature, target join
snapshots
partition
training data
test data
sampling
sampled data
model training
scoring
model evaluation

previous model
best model
model deployment

model training
model scoring
model evaluation
model deployment

scheduled workflow on Azkaban
Models Should Adapt to Large Dynamic Data
Ads Click Prediction Problem As an Example

- A member comes to LinkedIn
- Ads platform prepares a list of eligible advertising campaigns
- The statistical model predicts CTR for \(<member, ad, context>\)
- Rank all ads by \((\text{predicted CTR} \times \text{bid})\), and show the top-\(k\) results to the member
Models Should Adapt to Large Dynamic Data

- Very Large-scale data
  - Billions of records, large feature space (~100k covariates)

- Low positive rates (e.g. CTR)
  - Sparsity issue

- Data is dynamic
  - New ads come into the system any time

- Models have to be adaptive
  - Otherwise bad experience, $ loss
Cold-start & Warm-start Model

- Cold-start & Warm-start
  \[ p(\text{click} | \text{member, ads, context}) = \frac{1}{1 + \exp \left( -b - \theta^T_{\text{cold}} X_{\text{cold}} - \theta^T_{\text{warm}} X_{\text{warm}} \right)} \]

- Cold-start component \( \theta_{\text{cold}} \) relatively stable
  - Less frequent update
  - Large-scale model fitting using large amount of historical data

- Warm-start component \( \theta_{\text{warm}} \) more dynamic
  - Update as frequently as possible
  - Trained using fresh data to explain the residual of cold start only model

**Large-scale logistic regression**

**Per-item logistic regression**
Cold-start Model Fitting

- Alternating Direction Method of Multipliers (ADMM)
  - Stephen Boyd et al. 2011
  - Constraint that each partition’s coefficient $\beta_i = \text{global consensus } \beta$

$$\min_{\beta_i} \sum_{i=1}^{N} l_i(y_i, X_i^T \beta_i) + r(\beta)$$

subject to $\beta_i = \beta$
Large Scale Logistic Regression via ADMM

Iteration 1

BIG DATA

Partition 1
  Logistic Regression

Partition 2
  Logistic Regression

Partition 3
  Logistic Regression

... (K-1 partitions)

Partition K
  Logistic Regression

Consensus Computation
Large Scale Logistic Regression via ADMM

Iteration 1
Large Scale Logistic Regression via ADMM

Iteration 2
Warm-start Model Fitting

- Warm-start $\theta_{\text{warm}}$ trained with fresh data to explain the residual of cold start only model

\[
\ell(\theta, D) = \prod_{i \in D} \frac{1}{1 + \exp \left( -y_i \left( \theta^T x_i + \text{cold} \right) \right)}
\]

- Select small set of item-dependent features for fast training
- Train as frequently as possible
Explore/Exploit

- Model serving and data feedback loop
  - Ad A: true CTR 5%, 500 clicks, 10000 views
  - Ad B: true CTR 10%, 0 click, 10 views

- We ends up always serving A!

- Multi-armed bandit problem
  - $\epsilon$-greedy: random selection for $\epsilon\%$ traffic
  - UCB (Upper Confidence Bound)
  - Gittins Index
  - Thompson sampling
Thompson Sampling

- Instead of using MAP estimation of $\theta$, we sample it from its posterior distribution

- We sample from warm-start model only

- Compute CTR using the sampled coefficients

Assumptions
- Cold start coefficients have 0 variance
- Covariance between warm start coefficients is 0
Explore/Exploit Experiments

- Campaign Warmness Segment
- Lift Percentage of CTR
- LASER

Graph showing:
- Exploration Cost
- Winner’s Curse
- CTR Lift vs. Campaign Warmness Segment

Almost no training data vs. lots of training data

Lines:
- Without E/E
- With E/E
Offline Metrics Do Not Always Reflect Online
Offline Metrics Do Not Always Reflect Online

- **Offline Metrics**
  - ROC, AUC
  - Test log likelihood

- **Online Metrics**
  - eCPC (effective cost per click)
  - eCPM (effective cost per thousand impressions)
  - downstream or sitewide effects are hard to measure offline…

- **How to select best model to ramp**
  - A/B testing in a scientific and controlled manner
A/B Testing in One Slide

80%  
Control

---

20%

Treatment

Get started—it’s free.

First name  Last name

xulongya@gmail.com

......

By clicking Join Now, you agree to LinkedIn’s User Agreement, Privacy Policy and Cookie Policy.

Join now

Get started— it’s free.

First name  Last name

xulongya@gmail.com

......

By clicking Join Now, you agree to LinkedIn’s User Agreement, Privacy Policy and Cookie Policy.

Join now

---

collect results to determine which one is better
Ads CTR Drop

profile top ads

sudden CTR drop
Root Cause

5 Pixels!

- navigation bar
- profile top ads
Growth of Experiments at LinkedIn

- 200+ experiments
- 800+ side-wide and vertical specific metrics
- Billions of experiment events
- All on a daily basis, and counting …
Simple Experiments Management

- Easy design and deployment of experiments with many built-in targeting attributes to select
Automatic Analysis

- Statistical analysis
  - Statistical significance test (p-value, conf. interval)
  - Deep-dive: slicing & dicing capability

- Metrics design and management
  - De-centralized ownership
  - Centralized onboarding and management
Who Moved My Cheese?

Experiments Owners

- Provide visibility, ensure responsible ramping, and encourage communication

Metrics Owners

Most Impactful Experiments

- XLNT

- Most Impactful Experiments

- Metrics I care about

- Results for week of Oct 6

- Owner(s)

- Actions

- Impact

- Test Key

- Test Description

- abronz, egoyal, aijer, chicks, traeing,...

- kanderas, pruev

- kanderas, pruev

- pruev
Real-time Feedback is Important
Real-time Feedback is Important

- Recommendations served every millisecond

- How frequently do users keep seeing the same item?
  - High frequency ➔ user fatigue ➔ low engagement

- Impression discounting to penalize relevance scores by impression counts
Impression Discounting

Real-time impression tracking

<table>
<thead>
<tr>
<th>Member</th>
<th>member:1234567</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time Range</td>
<td>within last 12 hours</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Item Id</th>
<th>Action</th>
<th>Counts</th>
</tr>
</thead>
<tbody>
<tr>
<td>item:10</td>
<td>VIEW</td>
<td>3</td>
</tr>
<tr>
<td>item:15</td>
<td>VIEW</td>
<td>0</td>
</tr>
<tr>
<td>item:20</td>
<td>VIEW</td>
<td>2</td>
</tr>
</tbody>
</table>

Impression discounting

<table>
<thead>
<tr>
<th>Item Id</th>
<th>Relevance Score</th>
<th>Action</th>
<th>Counts</th>
<th>Adjusted Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>item:10</td>
<td>0.9</td>
<td>VIEW</td>
<td>3</td>
<td>0.9 * (0.2231) = 0.20</td>
</tr>
<tr>
<td>item:15</td>
<td>0.8</td>
<td>VIEW</td>
<td>0</td>
<td>0.8</td>
</tr>
<tr>
<td>item:20</td>
<td>0.7</td>
<td>VIEW</td>
<td>2</td>
<td>0.7 * (0.3679) = 0.26</td>
</tr>
</tbody>
</table>

score ← score × exp(-exposure / factor),
where exposure is a function of impression counts.

Sounds a simple idea! But …
How to Achieve (Near) Real-time Tracking

- **Apache Kafka**: a high-throughput distributed messaging system
- **Voldemort**: distributed key-value storage system
Conclusions

- Fast iteration is desired
- Models should adapt to large dynamic data
- Offline metrics do not always reflect online
- Real-time feedback is important
- and there is something else as important

We are hiring! 😊