Offline Evaluation and Optimization for Interactive Systems

Lihong Li
Microsoft Research

http://research.microsoft.com/en-us/people/lihongli

Tutorial URL
User Interaction
Big Trap

Correlation ≠ Causation
Somewhat Toy-ish Example

- Studies show... people who search their names in search engines tend to have higher income

- Decision making:
WWII Example

- Statistics collected during WWII...
  - Bullet holes on bomber planes that came back from mission

- Decision making:
  - Where to armor?
  - Abraham Wald: the opposite!
Outline

• Introduction
• Contextual bandits
• **Basic offline evaluation**
• **Enhanced techniques**
• Practical issues
• Concluding remarks
Introduction
News Recommendation

• Recommend 2 news articles {sport, movie} to users
• To maximize CTR (click-through rate)

<table>
<thead>
<tr>
<th></th>
<th>Overall CTR</th>
<th>Male</th>
<th>Female</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sport</td>
<td>0.5</td>
<td>0.4</td>
<td>0.8</td>
</tr>
<tr>
<td>Movie</td>
<td>0.6</td>
<td>0.3</td>
<td>0.7</td>
</tr>
</tbody>
</table>

• Known as Simpson’s Paradox
  • Observed in medical research, student administration, ...
  • More data does not help (because of “confounding”)
  • More features do not reliably address the problem

Correlation ≠ Causation!
Correlation vs. Causation

Can I predict click well assuming fixed RecSys?

Metrics
Precision, Recall, MSE, NDCG, ...

Can I increase CTR if I change RecSys?

“causal effect”
“manipulation”

Metrics
CTR, revenue, ...

Similar in Web search, advertising, ...
Controlled Experiments to Identify Causality

Everyday practice of scientist, doctors, ...
See survey of Web applications [KLSH'09]

Also known as
A/B tests, randomized clinical trials, ...

<table>
<thead>
<tr>
<th></th>
<th>Overall</th>
<th>Male</th>
<th>Female</th>
<th>EXP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sport</td>
<td>0.5</td>
<td>0.4</td>
<td>0.8</td>
<td>0.6</td>
</tr>
<tr>
<td>Movie</td>
<td>0.6</td>
<td>0.3</td>
<td>0.7</td>
<td>0.5</td>
</tr>
</tbody>
</table>
## Offline vs. Online Gap in Practice

<table>
<thead>
<tr>
<th></th>
<th>Correlation</th>
<th>Causation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Offline</td>
<td>ML to improve prec/recall, MSE, NDCG, ...</td>
<td>This tutorial</td>
</tr>
<tr>
<td>Online</td>
<td>Verify CTR/$$$$ lift by controlled experiments</td>
<td></td>
</tr>
</tbody>
</table>

### Common practice
- “guess and check”

### Limitations
- Online experiments are expensive
- Online experiments take a long time
- Often correlation $\Rightarrow$ causation

*Offline/online: whether to run a new system on live users to collect new data*
Related Areas

- (Stats/Econ) Estimating causal effects from observational data
  - Neyman-Rubin causal model [R’74] [H’86]
  - Heckman correction [H’79]
  - “Causality” [P’09]

- (AI) Off-policy reinforcement learning [PSS’00]

- (ML/Stats) Covariate shift [CSSL’08]
Recap

• Correlation ⟷ causation
  • E.g., lower MSE ⟷ CTR/revenue lift

• Controlled experiments measure causal effects (e.g., CTR lift)
  • but are expensive

• This tutorial: how to use historical data to estimate causal effects without running new online experiments

Note: Offline experiments cannot fully replace online experiments!
Contextual Bandits
Contextual Bandit [BA85, LZ08]

Observe $K$ “actions” $A_t$ and “context” $x_t$

Follow “policy” $\pi$ to choose $a_t \in A_t$

Receive “reward” $r_t \in [0,1]$

$t \leftarrow t + 1$

Stochastic assumption: $x_t \sim D_x(\cdot)$, $r_t \sim D_r(\cdot \mid x_t, a_t)$

Goal is to maximize “value”: $V(\pi, T) = \mathbb{E}\left[\frac{1}{T}(r_1 + r_2 + \cdots + r_T)\right]$

Stationary policy: $a_t = \pi(x_t)$

Non-stationary policy: $a_t = \pi(x_1, a_1, r_1, \ldots, x_{t-1}, a_{t-1}, r_{t-1}, x_t)$

(e.g., online learning algorithms)

historical data up to time $t$
Contextual Bandit Applications

• Clinical trials
• Resource allocation
• Queuing & scheduling
• ...
• Web (more recently)
  • Recommendation
  • Advertising
  • Search
• Intelligent assistant (Office)
• Adaptive user interface
Example: Personalized News Recommendation

$$x_t:$$ user features (age, gender, location, ...)
$$A_t:$$ available articles at time $$t$$
$$a_t:$$ recommended article
$$r_t:$$ 1 for click, 0 for no-click

Policy value $$V(\pi)$$ is click-through rate (CTR)
Example: Online Advertising

Context: query, user info, ...
Action: displayed ads
Reward: revenue
Example: Web Search Ranking

Search as a bandit (naive formulation):
- Context: query
- Action: ranked list
- Reward: search success-or-not
Policy Optimization

• Given data $D = \{(x_i, a_i, r_i)\}_{i=1,2,...,L}$ collected in the past, find $\pi^* = \arg\max_{\pi} V(\pi)$

• Examples: use log data to optimize...
  • recommender model to maximize CTR
  • ad ranking system to maximize revenue
  • search engine’s query suggestion model to maximize user satisfaction
  • personal treatment plan to maximize survival rate
  • ...

Policy Evaluation

• Given \( D \) and \( \pi \), estimate \( V(\pi) \) or \( V(\pi, T) = \mathbb{E} \left[ \frac{1}{T} (r_1 + r_2 + \cdots + r_T) \right] \)

• Example: use log data to estimate...
  • daily CTR of a news recommendation system
  • click lift of a new user feature in ad ranking
  • reduction of time for user to find a relevant URL on SERP
  • ...

• Why care evaluation
  • An important question on its own
  • Optimization can be reduced to evaluation: \( \pi^* = \arg\max_{\pi} V(\pi) \)
Online vs. Offline Evaluation of $V(\pi, T)$

- Online evaluation
  - Controlled experiments (AB tests)
  - Wait for days/weeks/months and compute average reward
  - Reliable but expensive

- Offline evaluation
  - Use historical data $D = \{(x, a, r_a)\}$
  - Cheap, fast, and risk-free
  - Counterfactuality of rewards: do not observe $r_{\pi(x)}$ if $\pi(x) \neq a$
Recap

• Contextual bandit as natural model for many interactive ML problems
• Policy evaluation vs. optimization
• Online vs. offline policy evaluation
Direct Method (aka Regression Estimator)

Data
\[
\{(x_1, a_1, r_1), \ldots, (x_L, a_L, r_L)\}
\]

Reward simulator:
\[
\hat{r}(x, a) \approx \mathbb{E}[r_a | x]
\]

this (difficult) step is often biased

unreliable evaluation

\[
\hat{V}_{dm}(\pi) = \frac{1}{L} \sum_i \hat{r}(x_i, \pi(x_i))
\]
Biases of Direct Method

• Sampling/selection bias
  • From production systems
  • Simpson’s paradox

• Modeling bias
  • Insufficient features to fully represent $r(x, a)$

Neither issue goes away even with infinite data!
Usually difficult to quantify modeling bias!
Randomized Data Collection

Randomized data collection: at step $t$,

- Observe current context $x$
- Randomly chooses $a \in A$ according to $(p_1, p_2, \ldots, p_K)$ and receives $r_a$

**End result:** “exploration data” $D = \{(x, a, r_a, p_a)\}$

Will use it to evaluate both stationary and nonstationary policies.
Randomized Data Collection: An Example
Randomized Data Collection: An Example
Inverse Propensity Score: Stationary Policy

\[ \hat{V}_{ips}(\pi) = \frac{1}{L} \sum_{(x,a,p_a,r_a) \in D} \frac{r_a \cdot 1(\pi(x) = a)}{p_a} \]

**Theorem:** \( \hat{V}_{ips}(\pi) \) is unbiased

**Proof:**

\[
E[\hat{V}(\pi)] = E \left[ \frac{r_a \cdot 1(\pi(x) = a)}{p_a} \right]
\]

\[
= E \left[ \Sigma_a \left( p_a \times \frac{r_a}{p_a} \right) 1(\pi(x) = a) \right]
\]

\[
= E \left[ \Sigma_a \left( r_a \times 1(\pi(x) = a) \right) \right]
\]

\[
= E_x \left[ r_{\pi(x)} \right] = V(\pi)
\]

Indicator function: 1 if TRUE, 0 if FALSE

“propensity score”
Confidence Interval Estimation for IPS

\[ \hat{V}_{\text{ips}}(\pi) = \frac{1}{L} \sum_{(x,a,p_a,r_a) \in D} \frac{r_a \cdot 1(\pi(x) = a)}{p_a} \]

- Consistency: if \( p_a \) is not too small, \( \hat{V}_{\text{ips}} \) converges to \( V(\pi) \) as \( L \to \infty \)

- Variance: \( \text{Var}[\hat{V}_{\text{ips}}(\pi)] = \frac{1}{L} \text{Var} \left[ \frac{r_a \cdot 1(\pi(x)=a)}{p_a} \right] \)

- 95% confidence interval
  \[ \hat{V}_{\text{ips}}(\pi) \pm \left( 1.96 \times \frac{\hat{\sigma}}{\sqrt{L}} \right) \]

- Generally, width of confidence interval shrinks to 0 at rate \( O\left(1/\sqrt{L}\right) \)
An Illustration

<table>
<thead>
<tr>
<th>ID</th>
<th>x</th>
<th>α</th>
<th>r_a</th>
<th>p_a</th>
<th>π(x)</th>
<th>π'(x)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Alice</td>
<td>F</td>
<td>1/2</td>
<td>M</td>
<td>F</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Bob</td>
<td>M</td>
<td>0</td>
<td>1/3</td>
<td>S</td>
<td>M</td>
</tr>
<tr>
<td>3</td>
<td>Chuck</td>
<td>S</td>
<td>1/6</td>
<td>S</td>
<td>F</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Diane</td>
<td>M</td>
<td>1/3</td>
<td>M</td>
<td>F</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Eric</td>
<td>F</td>
<td>0</td>
<td>1/2</td>
<td>S</td>
<td>M</td>
</tr>
<tr>
<td>6</td>
<td>Frank</td>
<td>F</td>
<td>0</td>
<td>1/2</td>
<td>S</td>
<td>F</td>
</tr>
<tr>
<td>7</td>
<td>Gordon</td>
<td>M</td>
<td>1/3</td>
<td>S</td>
<td>S</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>Henry</td>
<td>S</td>
<td>0</td>
<td>1/6</td>
<td>S</td>
<td>F</td>
</tr>
<tr>
<td>9</td>
<td>Irene</td>
<td>F</td>
<td>0</td>
<td>1/2</td>
<td>M</td>
<td>F</td>
</tr>
<tr>
<td>10</td>
<td>Jennifer</td>
<td>F</td>
<td>1</td>
<td>1/2</td>
<td>M</td>
<td>S</td>
</tr>
</tbody>
</table>

\[ A = \{ \text{Finance, Movie, Sport} \} \]

\[ p = \begin{pmatrix} 1 & 1 & 1 \\ 2 & 3 & 6 \end{pmatrix} \]

\[ \hat{V}_{ips}(\pi) = \frac{1}{|D|} \sum_{(x,a,p_a,r_a) \in D} \frac{r_a \cdot 1(\pi(x) = a)}{p_a} \]

\[ = \frac{1}{10} \left( \frac{1}{1/6} + \frac{1}{1/3} + \frac{0}{1/6} + 0 + \ldots + 0 \right) \]

\[ = \frac{9}{10} \]

\[ \hat{\sigma}^2_{ips} = \hat{\sigma}^2 \left( \frac{1}{1/6}, \frac{1}{1/3}, \frac{0}{1/6}, \ldots, 0 \right) \]

Seven 0s
Case Study 1: News Recommendation [LCLW’11]

- Experiments run in 2009
  - 40M impressions over 10 days in exploration data
  - $p_a = \frac{1}{K}$ (uniform random exploration)
- Fixed an news-selection policy $\pi$
- Online experiment with $\pi$ to measure CTR
  - The online ground truth
- Use exploration data to offline-evaluate $\pi$
  - The offline estimate

$A_t$: available articles at time $t$
$x_t$: user features (age, gender, interests, ...)
$a_t$: the displayed article at time $t$
$r_{t,a_t}$: 1 for click, 0 for no-click
Unbiasedness: Article CTR

The offline estimate is indeed unbiased!
Unbiasedness: Daily Overall CTR

The offline estimate is indeed unbiased!

Recorded Online nCTR

Estimated nCTR

Ten Days in November 2009
Recall the theoretical error bound [LCLW’11]

\[ |V(\pi) - \hat{V}_{ips}(\pi)| = O\left(\frac{1}{\sqrt{L}}\right) \]
Case Study 2: Bing Speller

What Speller does:
- Corrects typos
- May produce multiple candidates (with search results blended later)

Popular approach:
- Obtain human labels for \((q_0, q'_c, \text{label})\)
- Apply ML to rank candidates
- But...
Bing Speller: A Harder Example

ccn: popular and similar query (excellent reformulation candidate)

or

community cable network

ccn international

cement chemist notation

::
A user-oriented solution: use click to measure success

Standard solution is A/B test... but expensive

Click metrics are hard to work with offline (b/c counterfactual nature)
Speller as Contextual Bandit

A round-by-round interaction between **Speller** and **User**

At each round,

- **U** issues query $q_0$ ("context")
- **S** calculates a small set of promising candidates $Q = \{q_1, \ldots, q_L\}$
  - Note: $Q$ is assumed given (from other ML models)
- **S** then chooses an "action" $a \subset Q$
- **S** finally observes the **reward** (some click metric) $r_a$ for $a$
- Repeat

Goal of Speller is to maximize average per-round reward.
Exploration Data Collection \[\text{[LCKG'14]}\]

$q_0 = \{\text{shanghai pass}\}$

$Pr(q_i \text{ is sent}) = \frac{1}{1 + \exp(\lambda_1(\text{score}(q_1) - \text{score}(q_i)) + \lambda_2)}$

$\lambda_1$ and $\lambda_2$ control exploration aggressiveness

Candidates:
$q_1 = \{\text{shanghai pass}\}$
$q_2 = \{\text{shanghai pass}\}$
$q_3 = \{\text{shanghai past}\}$

Randomize the subset

Rest of Bing (retrieval, ranking, merging, ...)

Collected \(~15\text{M}\) search impressions at Bing
Accuracy of Offline Evaluator

Position-specific click-through rate

Daily click-through rate
Normality of Offline Estimates

Bootstrapping $B = 10000$
Reasonable to use normal approx.
Quantifying Uncertainty in Offline Evaluation

[Graph showing the relationship between online metric and key speller parameter]
Offline Optimization for Speller

• 70% exploration data to learn
  \[ \Pr(\text{GoodResult} \mid \text{Query, CorrectionCandidate}) \]
• 30% exploration data to offline-compare new and old Spellers

• Tends to be better if more are included
• But limited by capacity \(\rightarrow\) threshold needed
• Use unbiased IPS offline evaluation to set a threshold
Offline Optimization for Speller

• Tune Speller parameters to optimize offline estimate of $V(\pi)$
• Online-test one of most promising models
  ✓ showing statistically significant gain

• Some winning examples
  “umecka and zinc” $\rightarrow$ “umcka and zinc” (treatments for cold symptoms)
  “catalina left attorney” $\rightarrow$ “catalina leff attorney” (right correction)
  “acer e1-5726870” $\rightarrow$ “acer e1-572□6870” (correct word breaking)
umecka and zinc vs. umecka

Can Zinc Lozenges and Nasal Sprays Remedy Your Cold?
www.webmd.com › Cold, Flu, & Cough Health Center › Cold Guide
Can zinc prevent or reduce the duration of cold symptoms? Learn more about zinc’s benefits as a cold remedy from the experts at WebMD.

Zinc, umcka & elderberry for cold season | Pharmaca
www.pharmaca.com/projectwellness/2014/10/10/my-3-favorite-natural... Dr. Terraona Low Dog talks about her medicine cabinet must-haves during cold and flu season, including zinc, umcka loabo and elderberry.

ZINC: Uses, Side Effects, Interactions and Warnings - WebMD
www.webmd.com › WebMD Home › Vitamins & Supplements
Find patient medical information for ZINC on WebMD including its uses, effectiveness, side effects and safety, interactions, user ratings and products that have it.

Zinc — Health Professional Fact Sheet - Office of...
ods.od.nih.gov/factsheets/Zinc-HealthProfessional
Zinc is an essential mineral that is naturally present in some foods, added to others, and available as a dietary supplement. Zinc is also found in many cold lozenges ...

Umcka® - Get back to life faster with all natural Umcka...
www.umcka.com
Umcka® - Get back to life with Umcka® Coldcare and Cold+Flu! Recover from the cold and flu faster with Umcka natural cold and flu products including liquids ...

Jolanta Umecka - IMDb
www.imdb.com/name/nm0880840
Jolanta Umecka, Actress: Nóż w wodzie. Jolanta Umecka is an actress, known for Knife in the Water (1962), Panna zázrascnica (1967) and Echo … News › Biography › Awards › Films

Related searches for umeka
Umcka Cold Remedy  Umcka Drops
Umckaloabo Walgreens  Where to Buy Umcka
Umcka Cold  Umcka Walgreens

Knife in the Water - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/Knife_in_the_Water
Knife in the Water is a 1962 Polish drama film co-written and directed by Roman Polanski, which was nominated for Academy Award for Best Foreign Language Film. It … Plot › Cast › Production › Critical reception › Home video
Evaluating Nonstationary Policies

- To estimate: \( V(\pi, T) = \mathbb{E} \left[ \frac{1}{T} (r_1 + r_2 + \cdots + r_T) \right] \)
  where \( a_t = \pi(x_1, a_1, r_1, \ldots, x_{t-1}, a_{t-1}, r_{t-1}, x_t) \)
- Examples: all explore-exploit learning algorithms
- Simple inverse propensity score does not work
- Need to simulate the trajectory
The Replay Method [LCLS’10, LCLW’11]

Key requirement for data collection: $p_a \equiv \frac{1}{K}$

For $i = 1, 2, \ldots, L$

Data

\[
\left\{ \left( x_1, a_1, r_1, \frac{1}{K} \right) \right\} \\
\vdots \\
\left\{ \left( x_L, a_L, r_L, \frac{1}{K} \right) \right\}
\]

reveal $x_i$

choose $\hat{a}_i = \pi(x_i)$

reveal $r_i$ only if $\hat{a}_i = a_i$ ("match")

Nonstationary policy $\pi$

Finally output

\[
\hat{V} \left( \pi, \frac{L}{K} \right) = \frac{K}{L} \times \sum_{i=1}^{L} \left( r_i \cdot 1(\hat{a}_i = a_i) \right)
\]
Unbiasedness of Replay

- **Theorem**: if $L$ is large enough to generate $T$ matches in replay, then
  \[ E[\hat{V}(\pi, T)] = V(\pi, T) \]
- Unfortunately, cannot use $L$ or $T$ to estimate confidence intervals
- Can use bootstrapping instead

- How large $L$ do we need to have $T$ matches?
  - On average, $L = KT$
  - With high probability, need $L \approx 2KT$

- More discussions later
Replay with Non-uniform Exploration

- Data $D = \{(x, a, r_a, p_a)\}$ where $p_a \neq \frac{1}{K}$
- Can apply rejection sampling to obtain a subset of uniform $p_a$

\[ \left\{ (x_1, a_1, r_1, p_1) \right\}, \ldots, \left\{ (x_L, a_L, r_L, p_L) \right\} \]

Subsample with probability $p_i/p_{\text{max}}$

\[ \left\{ (x_1, a_1, r_1, 1/K) \right\}, \ldots, \left\{ (x_L, a_L, r_L, 1/K) \right\} \]

- Not very efficient when $p_i$’s vary a lot
- Adaptive rejection sampling [DELL’12]
Case Study 3: News Recommendation

- Data collected in 2009
  - 40M impressions over 10 days in exploration data
  - \( p_a = \frac{1}{K} \) (uniform random exploration)

- Low variance when evaluating representative nonstationary policies

<table>
<thead>
<tr>
<th>algorithm</th>
<th>mean</th>
<th>std</th>
<th>max</th>
<th>min</th>
</tr>
</thead>
<tbody>
<tr>
<td>\epsilon\text{-greedy}</td>
<td>1.2664</td>
<td>0.0308</td>
<td>1.3079</td>
<td>1.1671</td>
</tr>
<tr>
<td>UCB</td>
<td>1.3278</td>
<td>0.0192</td>
<td>1.3661</td>
<td>1.2812</td>
</tr>
<tr>
<td>LinUCB</td>
<td>1.3867</td>
<td>0.0157</td>
<td>1.4268</td>
<td>1.3491</td>
</tr>
</tbody>
</table>

100 independent runs with different randomization seed

**Conjecture**: Replay has low variance for *reasonable* nonstationary policies
Application of Replay: Personalized Explore-Exploit Algorithms [LCLS’10]
Application of Replay: Effects of Reward Delay [CL’11]
Application of Replay: Multi-objective Optimization [ACEW’11&12]
Recap

• Direct method by estimating $\hat{f}(x, a)$ is inherently biased
• Stationary policies: Inverse propensity Score ensures unbiasedness
  • With easily quantified variance
• Nonstationary policies: Replay method

• Case studies:
  • News recommendation
  • Bing search engine
Enhanced Techniques

**Unknown propensity scores**
Direct policy optimization
Doubly robust estimation
Bootstrapped replay
Unknown Propensity Scores

• So far we have assumed exploration data $D = \{(x, a, r_a, p_a)\}$
• Sometimes $p_a$ is unavailable
  • Data was generated by multiple deterministic policies ($p_a \equiv 1$ in this case) — “natural exploration”
  • Data loss or contamination ($p_a$ not truthful of real action distribution in data)
  • ...

• Not all hope is lost
IPS with Estimated Propensity Scores

• Data $D = \{(x_1, a_1, r_1), (x_2, a_2, r_2), \ldots, (x_L, a_L, r_L)\}$
  where $a_t \sim p_t(\cdot | x_t)$ [\(p_t\) unknown or deterministic]
• **Assumption**: $\pi_t$ independent of $D$
• Define “averaged” distribution $p = \frac{1}{L}(p_1 + p_2 + \cdots + p_L)$
• Estimate $\hat{p}(a|x) \approx p(a|x)$
  • Multinomial logistic regression, neural network, decision trees, ...

$$\hat{V}_{ips}(\pi) = \frac{1}{L} \sum_i r_i \cdot \frac{1(\pi(x_i) = a_i)}{\max\{\hat{p}(a_i|x_i), \tau\}}$$

Avoid division by tiny numbers.
Properties

\[
\hat{V}_{ips}(\pi) = \frac{1}{L} \sum_{i} \frac{r_i \cdot 1(\pi(x_i) = a_i)}{\max\{\hat{p}(a_i|x_i), \tau\}}
\]

- Slightly biased
  - \(\tau\): Under-estimation since it makes ratio smaller
  - \(1/\hat{p}\): Over-estimation

- Variance control
  - \(\tau\) helps stability (preventing division by tiny numbers)

- Combined [SLLK’10]

\[
|E[\hat{V}_{ips}(\pi) - V(\pi)]| \leq E_x \left[ \max_a |p(a|x) - \hat{p}(a|x)|/\tau \right]
\]

if \(p(\pi(x)|x) < \tau\) otherwise
Enhanced Techniques

Unknown propensity scores

**Direct policy optimization**

Doubly robust estimation

Bootstrapped replay
Policy Optimization

• Most often ultimate goal is to find optimal $\pi$ with maximum $V(\pi)$

• Approach 1: guess and check
  • Offline optimization against MSE/NDCG
  • Online experiment to verify gain in CTR/satisfaction/revenue

• Approach 2: direct solution
  • Offline optimization against $\hat{V}(\pi)$
  • Example: Bing Speller
  • Can be substantially generalized
Classification as Contextual Bandit

• Multi-class, multi-label classification

• Example $x$ associated with subset of correct labels $c \subseteq L = \{1, 2, \ldots, K\}$
  • $x$ (“imitation game”) -> $c$ ({historical, thriller})
Multi-label Classification as Contextual Bandit

• Use classification example \((x, c)\) to simulate interaction in bandit
  • \(x\): context
  • \(A = L\): candidate actions
  • \(r_a = 1(a \in c)\)
  • Essentially, \((x, c) \Rightarrow (x; r_1, r_2, ..., r_K)\)

• Policy \(\pi\) is treated as classifier

\[
V(\pi) = E_x[r(x, \pi(x))] = E_x[1(\pi(x) \in c)]
\]

Policy value is classification accuracy!
Policy Optimization as Classification

Contextual bandit $\rightarrow$ weighted multi-class classification

$$(x, a, r_a, p_a) \Rightarrow (x, a, w_a) \quad w_a = r_a/p_a$$

$$E_{x,a}[w_a \cdot 1(\pi(x) = a)] = E_x[r(x, \pi(x))] = V(\pi)$$

Policy value is same as weighted classification accuracy!

Maximize policy value $V(\pi)$ $\rightarrow$ Maximize weighted classification accuracy $V(\pi)$ $\rightarrow$ Multi-class classification algorithm

Offset tree [BL’09]: a similar and sometimes more effective optimization algorithm
Case Study 4: Advertising [SLLK’10]

- Problem: choose ad $a$ for $x = (\text{user, page})$ to maximize clicks
- Goal: learn from production data a warm-start policy better than random

- Non-exploration data $D = \{(x, a, r_a)\}$
  - 35M impressions for training
  - 19M impressions for test
  - 880K ads
  - 3.4M distinct webpages
  - $r_a \in \{0,1\}$: click or not
Three Algorithms for Comparison

• Random (baseline)

• Naive (supervised learning):
  • Learn scoring function $s(x, a)$ from data $D$
  • Policy $\pi(x) = \arg \max_a s(x, a)$

• Our approach (addressing bias in data):
  • Estimate propensity scores $\hat{p}(a|x)$ from data $D$
  • Learn regressor $f$ to minimize $\frac{(r_{a} - f(x, a))^2}{\max\{\hat{p}(a|x), \tau\}}$
  • Policy $\pi(x) = \arg \max_{a: \hat{p}(a|x) > 0} f(x, a)$
Warm Start Results

<table>
<thead>
<tr>
<th>Method</th>
<th>$\tau$</th>
<th>Estimate</th>
<th>Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>Learned</td>
<td>0.01</td>
<td>0.0193</td>
<td>[0.0187,0.0206]</td>
</tr>
<tr>
<td>Random</td>
<td>0.01</td>
<td>0.0154</td>
<td>[0.0149,0.0166]</td>
</tr>
<tr>
<td>Learned</td>
<td>0.05</td>
<td>0.0132</td>
<td>[0.0129,0.0137]</td>
</tr>
<tr>
<td>Random</td>
<td>0.05</td>
<td>0.0111</td>
<td>[0.0109,0.0116]</td>
</tr>
<tr>
<td>Naive</td>
<td>0.05</td>
<td>0.0</td>
<td>[0,0.0071]</td>
</tr>
</tbody>
</table>

- Ignoring bias in data, naive supervised learning even worse than random!
- Reasonably strong warm-start policies, even learned from non-exploration data
Enhanced Techniques

Unknown propensity scores
Direct policy optimization
**Doubly robust estimation**
Bootstrapped replay
Doubly Robust Estimation

- Direct Method (DM)
  \[ \hat{V}_{dm}(\pi) = \frac{1}{L} \sum \hat{r}(x, \pi(x)) \]

- Inverse Propensity Score (IPS)
  \[ \hat{V}_{ips}(\pi) = \frac{1}{L} \sum r_a \cdot 1(\pi(x) = a) \cdot \frac{1}{\hat{p}_a} \]

- Doubly Robust (DR) [RRZ’94]
  \[ \hat{V}_{dr}(\pi) = \frac{1}{L} \sum_{(x,a,r_a,\hat{p}_a) \in D} \left( \hat{r}(x, \pi(x)) + \frac{(r_a - \hat{r}(x, \pi(x))) \cdot 1(\pi(x) = a)}{\hat{p}_a} \right) \]

Estimate \( \hat{r}(x, a) \approx r(x, a) \)
- Small variance
- Large bias

No or small bias
- Large variance if \( p_a \approx 0 \)
DR: Unbiasedness

\[ \hat{V}_{dr}(\pi) = \frac{1}{L} \sum_i \left( \hat{r}(x, \pi(x)) + \frac{(r_a - \hat{r}(x, \pi(x))) \cdot 1(\pi(x) = a)}{\hat{p}_a} \right) \]

\[ = \frac{1}{L} \sum_i \left( \hat{r}(x, \pi(x)) \left( 1 - \frac{1(\pi(x) = a)}{\hat{p}_a} \right) + \frac{r_a \cdot 1(\pi(x) = a)}{\hat{p}_a} \right) \]

\[ \hat{r} = r \implies E[\hat{V}_{dr}] = V(\pi) \]

\[ \hat{p} = p \implies E[\hat{V}_{dr}] = V(\pi) \]

- Two ways to ensure unbiasedness ("doubly protected")
- Implemented in Vowpal Wabbit (http://hunch.net/~vw)
- Well-known in statistics, but not entirely satisfying
  - Almost impossible to have \( \hat{r} = r \) or \( \hat{p} = p \) in reality
  - Refined analysis for practically relevant situations [DLL’11]
DR: Bias Analysis

• $E[\hat{V}_{dr}] - V(\pi) = E_x[\text{err}_p(x) \cdot \text{err}_r(x)]$

  Error in $\hat{p}$  Error in $\hat{r}$

• $E[\hat{V}_{ips}] - V(\pi) = E_x[\text{err}_p(x) \cdot r(x, \pi(x))]$

• $E[\hat{V}_{dm}] - V(\pi) = E_x[\text{err}_r(x, \pi(x)) \cdot \max_{x,a}\{r(x, a)\}]$

DR has lowest bias with “reasonable” $\hat{p}$ and $\hat{r}$.
DR: Variance Analysis

- $\text{Var}\left[\hat{V}_{dr}\right] \approx \frac{1}{L} E_x \left[ \frac{\text{err}_r(x)^2 \cdot (1 - \text{err}_p(x))^2}{p(\pi(x) | x)} \right]$

- $\text{Var}\left[\hat{V}_{ips}\right] \approx \frac{1}{L} E_x \left[ \frac{r(x, \pi(x))^2 \cdot (1 - \text{err}_p(x))^2}{p(\pi(x) | x)} \right]$

- $\text{Var}\left[\hat{V}_{dm}\right] = \frac{1}{L} \text{Var}_x \left[ \hat{r}(x, \pi(x)) \right]$

DR has lower variance than IPS with “reasonable” $\hat{r}$

DM often has low variance, not affected by $p(a | x)$
Case Study 5: UCI datasets [DLL’11]

<table>
<thead>
<tr>
<th>Dataset</th>
<th>ecoli</th>
<th>glass</th>
<th>letter</th>
<th>optdigits</th>
<th>page-blocks</th>
<th>pendigits</th>
<th>satimage</th>
<th>vehicle</th>
<th>yeast</th>
</tr>
</thead>
<tbody>
<tr>
<td>Classes ($k$)</td>
<td>8</td>
<td>6</td>
<td>26</td>
<td>10</td>
<td>5</td>
<td>10</td>
<td>6</td>
<td>4</td>
<td>10</td>
</tr>
<tr>
<td>Dataset size</td>
<td>336</td>
<td>214</td>
<td>20000</td>
<td>5620</td>
<td>5473</td>
<td>10992</td>
<td>6435</td>
<td>846</td>
<td>1484</td>
</tr>
</tbody>
</table>

Classification to bandit: $(x, c) \Rightarrow (x; r_1, r_2, ..., r_K)$

Bandit to classification: $(x, a, r_a, p_a) \Rightarrow (x, a, w_a) \quad w_a = r_a / p_a$
Policy Evaluation

• 50% data for training (regular classification) to obtain $\pi$

• 50% data for testing with bandit labels
  • For each $x$, randomly pick $a \in \{1, \ldots, K\}$ and reveal $r_a = 1(a = c)$
    [classification to bandit reduction]
  • Only $1/K$ fraction of labels observed
  • Compare DM, IPS, DR
Policy Evaluation
Policy Optimization

• 70% data for training with bandit labels to obtain \( \pi \)
  • For each \( x \), randomly pick \( a \in \{1, \ldots, K\} \) and reveal \( r_a = 1\{a = c\} \)
  • Only \( 1/K \) fraction of labels observed

Optimization algorithms
• Direct loss minimization [MHK’11]
• Filter tree [BLR’08]
• Offset tree [BL’09]: alternative policy optimization algorithm

Generic multi-class classification
(Combined with DM, IPS, DR)

• 30% data for testing accuracy of \( \pi \) (regular classification)
Policy Optimization
Enhanced Techniques

Unknown propensity scores
Direct policy optimization
Doubly robust estimation
Bootstrapped replay
Time Acceleration Problem [NMP’14]

- With $L = |D|$ data and uniform exploration $p_a = 1/K$
  - Expected number of matches is $L/K$
  - Replay can estimate $V(\pi, T)$ up to $T \approx L/K$

Replay cannot evaluate $\pi$ for too large $T$ (from [NMP’14])
BRED [NMP’14] “Bootstrapped Replay on Expanded Data”

Data \( \{ (x_1, a_1, r_1, 1/K) \} \)

Subsample w/ replacement & jittering on \( x_i \)

Data \( \{ (x'_1, a'_1, r'_1, 1/K) \} \)

Data \( \{ (x''_1, a''_1, r''_1, 1/K) \} \)

Data \( \{ (x'''_1, a'''_1, r'''_1, 1/K) \} \)

Replay \( \hat{V}_1(\pi, T_1) \)

Replay \( \hat{V}_{k}(\pi, T_k) \)

Replay \( \hat{V}_{k}(\pi, T_k) \)

Replay \( \hat{V}_{k}(\pi, T_k) \)

\( \hat{V}_{\text{bred}}(\pi, T) \)
BRED Theory

• For stationary policies, confidence intervals are estimated much faster
  • $O(1/T)$ as opposed to $O\left(1/\sqrt{T}\right)$
  • under mild assumptions (similar to the bootstrap theory)

• For stationary policies, can estimate $V(\pi, T)$ for $T \gg L/K$
  • although the bootstrap theory does not apply

• Practical limitation: computationally expensive
  • fast, approximate bootstrap [OR’01]
  • implemented in Vowpal Wabbit [QPKLL’13]
Replay vs. BRED on Yahoo! News Recommendation

- Policy being evaluated: UCB
- Each point is a segment of data with same pool of candidate articles
- BRED much more accurate
Practical Issues
How to Design Exploration Distributions

• Use of natural exploration (without collecting truly randomized data)
  • Cheap, and potentially useful
  • But risky (by ignoring potential confounding)

• Need to design A properly before collecting data
How to Design Exploration Distributions (2)

• $Var\left(\hat{V}(\pi)\right)$ depends on how much $\pi$ “agree” with $p$
  • Usually $\pi$ not known in advance
  • Choice #1: uniform (best in the worst case) [news recommendation]
  • Choice #2: randomize around current/production policy [Speller]

• More exploration with $p$ causes greater potential risk
  • Negative user satisfaction, monetary loss, ...

• May use inner/outer confidence intervals to guide design [B+13]

Best decisions have to be on a case-by-case level
What Information to Log

• Data $D = \{(x, a, r, p_a)\}$

• Should log $x$ if possible to avoid inconsistency
  • Eg., $x$ has time-sensitive features
  • Eg., $x$ may be missing due to timeouts

• Should log $p_a$ (unless it’s precisely known)

• Should log immediate actions (not final actions)
Detecting Data Quality Issues

Data $D = \{(x, a, r, p)\}$

- Mean tests [LCKG’14]
  
  arithmetic: $\forall a': \sum_D 1(a = a') \approx \sum_D p(a'|x)$
  
  harmonic: $\sum_D \frac{1}{p} \approx L \times K$

- Can log randomization seed in $D$ and check offline to detect bugs
Concluding Remarks
Review

General theme: use historical data to offline-discovery online metrics (estimate causal effects from historical data)

• Policy evaluation/optimization

• Unbiasedness with IPS and Replay
• Variance reduction techniques with DR, etc.

• Case studies in news, search, advertising, and benchmark
More Bing Examples

**Including results for contextual bandit.**
Do you want results only for contextual bandit?

**Contextual Bandits** ≪ Machine Learning (Theory)
hunch.net/?p=298

When you compare contextual bandit to RL in general, contextual bandit is a special case for the most general RL formulations (as essentially everything is).

[A Contextual-Bandit Approach to Personalized ...](www.research.rutgers.edu/~lilong/pub/Li10Contextual.pdf)

A Contextual-Bandit Approach to Personalized News Article Recommendation Lihong Li, Wei Chu, †Yahoo! Labs lihong,chuwei@yahoo-inc.com John Langford‡

**Multi-armed bandit** - Wikipedia, the free encyclopedia
More Bing Examples

---

**Contextual Bandit**

When you compare... for the most general

---

**Multi-armed bandit** - Wikipedia, the free encyclopedia
Many More Applications

- Yahoo!, Google, Microsoft, LinkedIn, Adobe, Criteo, ...
  [LP'07] [LSW’08] [CGGHL'10] [PPBK’11] [ACEW’11] [TRSA’13] [A+’14] ...

- Can be combined with other methods like interleaving [HWR’12&’14]

- WWW 2015 Workshop in May (Florence, Italy)
  http://evalworkshop.com

- Datasets available at Yahoo! Webscope (R6B)
  http://webscope.sandbox.yahoo.com/catalog.php?datatype=r
Limitations and Open Questions

• Many actions
  • Relies on natural exploration and approximate matching [LKZ’15]
  • Use production data to approximate online behavior [YBL’15]
  • Continuous actions [B+’13]

• Cannot model long-term effects
  • Off-policy reinforcement learning
  • Equilibrium analysis [B+’13]

• Relies on stationary assumption

• Statistically more efficient (even optimal) offline estimation
References

• [ACEW'11] Deepak Agarwal, Bee-Chung Chen, Pradheep Elango, Xuanhui Wang: Click shaping to optimize multiple objectives. KDD 2011: 132-140

• [ACEW'12] Deepak Agarwal, Bee-Chung Chen, Pradheep Elango, Xuanhui Wang: Personalized click shaping through lagrangian duality for online recommendation. SIGIR 2012: 485-494


• [CGGHL'10] David Chan, Rong Ge, Ori Gershony, Tim Hesterberg, Diane Lambert: Evaluating online ad campaigns in a pipeline: Causal models at scale. KDD 2010: 7-16


• [HWR'13] Katja Hofmann, Anne Schuth, Shimon Whiteson, Maarten de Rijke: Reusing historical interaction data for faster online learning to rank for IR. WSDM 2013: 183-192


• [LKZ'15] Lihong Li, Jinyoung Kim, Imed Zitouni: Toward predicting the outcome of an A/B experiment for search relevance. WSDM 2015: 37-46

• [LMS'15] Lihong Li, Remi Munos, Csaba Szepesvari. Toward minimax off-policy value estimation. AISTATS 2015