SIGIR 2016 Tutorial

Counterfactual Evaluation and Learning

Adith Swaminathan, Thorsten Joachims

Department of Computer Science & Department of Information Science
Cornell University

Website: http://www.cs.cornell.edu/~adith/CfactSIGIR2016/

Funded in part through NSF Awards IIS-1247637, IIS-1217686, IIS-1513692.
User Interactive Systems

Examples
- Search engines
- Entertainment media
- E-commerce
- Smart homes, robots, etc.

→ Logs of User Behavior for
- Evaluating system performance
- Learning improved systems and gathering knowledge
- Personalization
Interactive System Schematic

Utility: $U(\pi_0)$

Context $x$ → Action $y$ for $x$ → Feedback $\delta(x, y)$
Ad Placement

• Context $x$:
  – User and page
• Action $y$:
  – Ad that is placed
• Feedback $\delta(x, y)$:
  – Click / no-click
News Recommender

• **Context** $x$:
  – User

• **Action** $y$:
  – Portfolio of news articles

• **Feedback** $\delta(x, y)$:
  – Reading time in minutes
• Context \( x \):
  – Query

• Action \( y \):
  – Ranking

• Feedback \( \delta(x, y) \):
  – win/loss against baseline in interleaving
Log Data from Interactive Systems

• Data

\[ S = ( (x_1, y_1, \delta_1), \ldots, (x_n, y_n, \delta_n) ) \]

→ Partial Information (aka “Contextual Bandit”) Feedback

• Properties

- Contexts \( x_i \) drawn i.i.d. from unknown \( P(X) \)
- Actions \( y_i \) selected by existing system \( \pi_0 : X \rightarrow Y \)
- Feedback \( \delta_i \) from unknown function \( \delta : X \times Y \rightarrow \mathbb{R} \)

[Zadrozny et al., 2003] [Langford & Li], [Bottou et al., 2014]
Goals for this Tutorial

• Use interaction log data
  \[ S = ((x_1, y_1, \delta_1), \ldots, (x_n, y_n, \delta_n)) \]
  for
– Evaluation:
  • Estimate online measures of some system \( \pi \) offline.
  • System \( \pi \) is typically different from \( \pi_0 \) that generated log.
– Learning:
  • Find new system \( \pi \) that improves performance over \( \pi_0 \).
  • Do not rely on interactive experiments like in online learning.
PART 1: EVALUATION

SIGIR 2016 Tutorial
Counterfactual Evaluation and Learning
Evaluation: Outline

• Evaluating Online Metrics Offline
  – A/B Testing (on-policy) → Counterfactual estimation from logs (off-policy)
• Approach 1: “Model the world”
  – Estimation via reward prediction
• Approach 2: “Model the bias”
  – Counterfactual Model
  – Inverse propensity scoring (IPS) estimator
• Advanced Estimators
  – Self-normalized IPS estimator
  – Doubly robust estimator
  – Slates estimator
• Case Studies
• Summary & Demonstration with code samples
Online Performance Metrics

Example metrics
- CTR
- Revenue
- Time-to-success
- Interleaving
- Etc.

→ Correct choice depends on application and is not the focus of this tutorial.

This tutorial:
Metric encoded as $\delta(x, y)$ [click/payoff/time for (x,y) pair]
• Definition [Deterministic Policy]: Function

\[ y = \pi(x) \]

that picks action \( y \) for context \( x \).

• Definition [Stochastic Policy]: Distribution

\[ \pi(y| x) \]

that samples action \( y \) given context \( x \).
System Performance

Definition [Utility of Policy]:

The expected reward / utility $U(\pi)$ of policy $\pi$ is

$$U(\pi) = \int \int \delta(x, y)\pi(y|x)P(x) \, dx \, dy$$

- e.g. reading time of user $x$ for portfolio $y$
Online Evaluation: A/B Testing

Given $S = ((x_1, y_1, \delta_1), ..., (x_n, y_n, \delta_n))$ collected under $\pi_0$,

$$\hat{U}(\pi_0) = \frac{1}{n} \sum_{i=1}^{n} \delta_i$$

→ A/B Testing

Deploy $\pi_1$: Draw $x \sim P(X)$, predict $y \sim \pi_1(Y|x)$, get $\delta(x, y)$

Deploy $\pi_2$: Draw $x \sim P(X)$, predict $y \sim \pi_2(Y|x)$, get $\delta(x, y)$

⋮

Deploy $\pi_{|H|}$: Draw $x \sim P(X)$, predict $y \sim \pi_{|H|}(Y|x)$, get $\delta(x, y)$
Pros and Cons of A/B Testing

• Pro
  – User centric measure
  – No need for manual ratings
  – No user/expert mismatch

• Cons
  – Requires interactive experimental control
  – Risk of fielding a bad or buggy $\pi_i$
  – Number of A/B Tests limited
  – Long turnaround time
Evaluating Online Metrics Offline

- **Online**: On-policy A/B Test
  - Draw $S_1$ from $\pi_1 \rightarrow \hat{U}(\pi_1)$
  - Draw $S_2$ from $\pi_2 \rightarrow \hat{U}(\pi_2)$
  - Draw $S_3$ from $\pi_3 \rightarrow \hat{U}(\pi_3)$
  - Draw $S_4$ from $\pi_4 \rightarrow \hat{U}(\pi_4)$
  - Draw $S_5$ from $\pi_5 \rightarrow \hat{U}(\pi_5)$
  - Draw $S_6$ from $\pi_6 \rightarrow \hat{U}(\pi_6)$
  - Draw $S_7$ from $\pi_7 \rightarrow \hat{U}(\pi_7)$

- **Offline**: Off-policy Counterfactual Estimates
  - Draw $S$ from $\pi_0 \rightarrow \hat{U}(\pi_0)$
Evaluation: Outline

- Evaluating Online Metrics Offline
  - A/B Testing (on-policy) → Counterfactual estimation from logs (off-policy)
- Approach 1: “Model the world”
  - Estimation via reward prediction
- Approach 2: “Model the bias”
  - Counterfactual Model
  - Inverse propensity scoring (IPS) estimator
- Advanced Estimators
  - Self-normalized IPS estimator
  - Doubly robust estimator
  - Slates estimator
- Case Studies
- Summary & Demonstration with code samples
Approach 1: Reward Predictor

• Idea:
  – Use $S = ((x_1, y_1, \delta_1), \ldots, (x_n, y_n, \delta_n))$ from $\pi_0$ to estimate reward predictor $\hat{\delta}(x, y)$

• Deterministic $\pi$: Simulated A/B Testing with predicted $\hat{\delta}(x, y)$
  – For actions $y'_i = \pi(x_i)$ from new policy $\pi$, generate predicted log $S' = \left( (x_1, y'_1, \hat{\delta}(x_1, y'_1)), \ldots, (x_n, y'_n, \hat{\delta}(x_n, y'_n)) \right)$
  – Estimate performance of $\pi$ via $\hat{U}_{rp}(\pi) = \frac{1}{n} \sum_{i=1}^{n} \hat{\delta}(x_i, y'_i)$

• Stochastic $\pi$: $\hat{U}_{rp}(\pi) = \frac{1}{n} \sum_{i=1}^{n} \sum_{y} \hat{\delta}(x_i, y) \pi(y|x_i)$
Regression for Reward Prediction

Learn $\hat{\delta}: x \times y \rightarrow \mathbb{R}$

1. Represent via features $\Psi(x, y)$
2. Learn regression based on $\Psi(x, y)$ from $S$ collected under $\pi_0$
3. Predict $\hat{\delta}(x, y')$ for $y' = \pi(x)$ of new policy $\pi$
News Recommender: Exp Setup

• **Context x:** User profile
• **Action y:** Ranking
  – Pick from 7 candidates to place into 3 slots
• **Reward** $\delta$: “Revenue”
  – Complicated hidden function
• **Logging policy** $\pi_0$: Non-uniform randomized logging system
  – Placket-Luce “explore around current production ranker” (see case study)
News Recommender: Results

RP is inaccurate even with more training and logged data
Problems of Reward Predictor

• Modeling bias
  – choice of features and model

• Selection bias
  – $\pi_0$’s actions are over-represented

$\hat{U}_{rp}(\pi) = \frac{1}{n} \sum_i \hat{\delta}(x_i, \pi(x_i))$

Can be unreliable and biased
Evaluation: Outline

• Evaluating Online Metrics Offline
  – A/B Testing (on-policy) → Counterfactual estimation from logs (off-policy)
• Approach 1: “Model the world”
  – Estimation via reward prediction
• Approach 2: “Model the bias”
  – Counterfactual Model
  – Inverse propensity scoring (IPS) estimator
• Advanced Estimators
  – Self-normalized IPS estimator
  – Doubly robust estimator
  – Slates estimator
• Case Studies
• Summary & Demonstration with code samples
Approach “Model the Bias”

• Idea:

Fix the mismatch between the distribution $\pi_0(Y|x)$ that generated the data and the distribution $\pi(Y|x)$ we aim to evaluate.

$$U(\pi_0) = \int \int \delta(x, y)\pi_0(y|x)P(x) \, dx \, dy$$
Counterfactual Model

- Example: Treating Heart Attacks
  - Treatments: $Y$
    - Bypass / Stent / Drugs
  - Chosen treatment for patient $x_i$: $y_i$
  - Outcomes: $\delta_i$
    - 5-year survival: 0 / 1
  - Which treatment is best?
Counterfactual Model

• Example: Treating Heart Attacks
  – Treatments: $Y$
    • Bypass / Stent / Drugs
  – Chosen treatment for patient $x_i$: $y_i$
  – Outcomes: $\delta_i$
    • 5-year survival: 0 / 1

• Placing Vertical

– Which treatment is best?
Counterfactual Model

• Example: Treating Heart Attacks
  – Treatments: \( Y \)
    • Bypass / Stent / Drugs
  – Chosen treatment for patient \( x_i : y_i \)
  – Outcomes: \( \delta_i \)
    • 5-year survival: 0 / 1
  – Which treatment is best?
    • Everybody Drugs
    • Everybody Stent
    • Everybody Bypass
  \( \rightarrow \) Drugs 3/4, Stent 2/3, Bypass 2/4 – really?
Treatment Effects

• Average Treatment Effect of Treatment $y$
  \[ U(y) = \frac{1}{n} \sum_i \delta(x_i, y) \]

• Example
  \[ - U(bypass) = \frac{5}{11} \]
  \[ - U(stent) = \frac{7}{11} \]
  \[ - U(drugs) = \frac{4}{11} \]
Assignment Mechanism

• Probabilistic Treatment Assignment
  – For patient $i$: $\pi_0(Y_i = y|x_i)$
  – Selection Bias

• Inverse Propensity Score Estimator
  – $\hat{U}_{ips}(y) = \frac{1}{n} \sum_i \frac{\mathbb{I}\{y_i = y\}}{p_i} \delta(x_i, y_i)$
  – Propensity: $p_i = \pi_0(Y_i = y_i|x_i)$
  – Unbiased: $E[\hat{U}(y)] = U(y)$, if $\pi_0(Y_i = y|x_i) > 0$ for all $i$

• Example
  – $\hat{U}(drugs) = \frac{1}{11} \left( \frac{1}{0.8} + \frac{1}{0.7} + \frac{1}{0.8} + \frac{0}{0.1} \right)$
    $= 0.36 < 0.75$

$$
\begin{array}{c|ccc}
\pi_0(Y_i = y|x_i) & 0.3 & 0.6 & 0.1 \\
0.5 & 0.4 & 0.1 \\
0.1 & 0.1 & 0.8 \\
0.6 & 0.3 & 0.1 \\
0.2 & 0.5 & 0.7 \\
0.7 & 0.2 & 0.1 \\
0.1 & 0.1 & 0.8 \\
0.1 & 0.8 & 0.1 \\
0.3 & 0.3 & 0.4 \\
0.3 & 0.6 & 0.1 \\
0.4 & 0.4 & 0.2 \\
\end{array}
$$

<table>
<thead>
<tr>
<th>Patients</th>
<th>Bypass</th>
<th>Stent</th>
<th>Drugs</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>
Experimental vs Observational

- **Controlled Experiment**
  - Assignment Mechanism under our control
  - Propensities \( p_i = \pi_0(Y_i = y_i|x_i) \) are known by design
  - Requirement: \( \forall y: \pi_0(Y_i = y|x_i) > 0 \) (probabilistic)

- **Observational Study**
  - Assignment Mechanism not under our control
  - Propensities \( p_i \) need to be estimated
  - Estimate \( \hat{\pi}_0(Y_i|z_i) = \pi_0(Y_i|x_i) \) based on features \( z_i \)
  - Requirement: \( \hat{\pi}_0(Y_i|z_i) = \hat{\pi}_0(Y_i|\delta_i, z_i) \) (unconfounded)
Conditional Treatment Policies

• Policy (deterministic)
  – Context $x_i$ describing patient
  – Pick treatment $y_i$ based on $x_i$: $y_i = \pi(x_i)$
  – Example policy:
    • $\pi(A) = \text{drugs}, \pi(B) = \text{stent}, \pi(C) = \text{bypass}$

• Average Treatment Effect
  – $U(\pi) = \frac{1}{n} \sum_i \delta(x_i, \pi(x_i))$

• IPS Estimator
  – $\hat{U}_{ips}(\pi) = \frac{1}{n} \sum_i \frac{\mathbb{1}\{y_i = \pi(x_i)\}}{p_i} \delta(x_i, y_i)$

\[
\begin{bmatrix}
0 & 1 & 0 \\
1 & 1 & 0 \\
0 & 0 & 1 \\
0 & 0 & 0 \\
0 & 1 & 0 \\
1 & 0 & 1 \\
0 & 0 & 1 \\
0 & 1 & 0 \\
0 & 1 & 0 \\
1 & 0 & 1 \\
1 & 1 & 1 \\
\end{bmatrix}
\]
Stochastic Treatment Policies

- Policy (stochastic)
  - Context $x_i$ describing patient
  - Pick treatment $y$ based on $x_i$: $\pi(Y|x_i)$
- Note
  - Assignment Mechanism is a stochastic policy as well!
- Average Treatment Effect
  - $U(\pi) = \frac{1}{n} \sum_i \sum_y \delta(x_i, y) \pi(y|x_i)$
- IPS Estimator
  - $\hat{U}(\pi) = \frac{1}{n} \sum_i \frac{\pi(y_i|x_i)}{p_i} \delta(x_i, y_i)$
# Counterfactual Model = Logs

<table>
<thead>
<tr>
<th>Context $x_i$</th>
<th>Recorded in Log</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treatment $y_i$</td>
<td>Recorded in Log</td>
</tr>
<tr>
<td>Outcome $\delta_i$</td>
<td>Recorded in Log</td>
</tr>
<tr>
<td>Propensities $p_i$</td>
<td>Recorded in Log</td>
</tr>
<tr>
<td>New Policy $\pi$</td>
<td>Recorded in Log</td>
</tr>
<tr>
<td>T-effect $U(\pi)$</td>
<td>Recorded in Log</td>
</tr>
</tbody>
</table>

Average quality of new policy.
Evaluation: Outline

• Evaluating Online Metrics Offline
  – A/B Testing (on-policy) → Counterfactual estimation from logs (off-policy)
• Approach 1: “Model the world”
  – Estimation via reward prediction
• Approach 2: “Model the bias”
  – Counterfactual Model
  – Inverse propensity scoring (IPS) estimator
• Advanced Estimators
  – Self-normalized IPS estimator
  – Doubly robust estimator
  – Slates estimator
• Case Studies
• Summary & Demonstration with code samples
System Evaluation via Inverse Propensity Scoring

Definition [IPS Utility Estimator]:

Given $S = ((x_1, y_1, \delta_1), ..., (x_n, y_n, \delta_n))$ collected under $\pi_0$,

$$\hat{U}_{ips}(\pi) = \frac{1}{n} \sum_{i=1}^{n} \delta_i \frac{\pi(y_i|x_i)}{\pi_0(y_i|x_i)}$$

→ Unbiased estimate of utility for any $\pi$, if propensity nonzero whenever $\pi(y_i|x_i) > 0$.

Note:

If $\pi = \pi_0$, then online A/B Test with $\hat{U}_{ips}(\pi_0) = \frac{1}{n} \sum_{i} \delta_i$

→ Off-policy vs. On-policy estimation.

[Horvitz & Thompson, 1952] [Rubin, 1983] [Zadrozny et al., 2003] [Li et al., 2011]
IPS Estimator:

\[ \hat{U}_{IPS}(\pi) = \frac{1}{n} \sum_{i} \frac{\pi(y_i|x_i)}{\pi_0(y_i|x_i)} \delta_i \]
IPS Estimator is Unbiased

\[ E[\hat{U}(\pi)] = \frac{1}{n} \sum_{x_1, y_1} \cdots \sum_{x_n, y_n} \left[ \sum_i \frac{\pi(y_i | x_i)}{\pi_0(y_i | x_i)} \delta(x_i, y_i) \right] \pi_0(y_1 | x_1) \cdots \pi_0(y_n | x_n) P(x_1) \cdots P(x_n) \]

\[ = \frac{1}{n} \sum_{x_1, y_1} \pi_0(y_1 | x_1) P(x_1) \cdots \sum_{x_n, y_n} \pi_0(y_n | x_n) P(x_n) \left[ \sum_i \frac{\pi(y_i | x_i)}{\pi_0(y_i | x_i)} \delta(x_i, y_i) \right] \]

\[ = \frac{1}{n} \sum_i \sum_{x_i, y_i} \pi_0(y_i | x_i) P(x_i) \left[ \frac{\pi(y_i | x_i)}{\pi_0(y_i | x_i)} \delta(x_i, y_i) \right] \]

\[ = \frac{1}{n} \sum_i \sum_{x_i, y_i} \pi(y_i | x_i) P(x_i) \delta(x_i, y_i) = \frac{1}{n} \sum_i U(\pi) = U(\pi) \]
IPS eventually beats RP; variance decays as $O \left( \frac{1}{\sqrt{n}} \right)$
Adith takes over