Counterfactual Model for Learning

CS6780 – Advanced Machine Learning
Spring 2019
Thorsten Joachims
Cornell University

Reading:

Interactive System Schematic
Utility: $U(\pi_0)$

News Recommender
- Context $x$:
  - User
- Action $y$:
  - Portfolio of news articles
- Feedback $\delta(x,y)$:
  - Reading time in minutes

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

Search Engine
- Context $x$:
  - Query
- Action $y$:
  - Ranking
- Feedback $\delta(x,y)$:
  - Click / no-click

Log Data from Interactive Systems
- Data
  $$S = \{(x_1, y_1, \delta_1), ..., (x_n, y_n, \delta_n)\}$$
  - Partial Information (aka “Contextual Bandit”)
- 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 $\delta: X \times Y \rightarrow \mathbb{R}$
Goal

Use interaction log data
\[ S = \{(x_1, y_1, \delta_1), ..., (x_n, y_n, \delta_n)\} \]
- for evaluation of system \( \pi \)
  - Offline estimate of online performance of some system \( \pi \).
  - System \( \pi \) can be different from \( \pi_0 \) that generated log.
- for learning new system \( \pi \)

Evaluation: Outline

- Offline Evaluating of Online Metrics
  - A/B Testing (on-policy)
  - Counterfactual estimation from logs (off-policy)
- Approach 1: "Model the world"
  - Imputation via reward prediction
- Approach 2: "Model the bias"
  - Counterfactual model and selection bias
  - Inverse propensity scoring (IPS) estimator

Online Performance Metrics

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

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

This lecture: Metric encoded as \( \delta(x, y) \) [click/payoff/time for \((x, y)\) pair]

System

- 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
\[ U(\pi) = \int \int \delta(x, y) \pi(y|x)P(x) \, dx \, dy \]

Online Evaluation: A/B Testing

Given \( S = \{(x_1, y_1, \delta_1), ..., (x_n, y_n, \delta_n)\} \) collected under \( \pi_0 \)
\[ \bar{U}(\pi_0) = \frac{1}{n} \sum_{i=1}^{n} \delta_i \]

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 \( \pi \)
  - Number of A/B Tests limited
  - Long turnaround time

**Approach 1: “Model the world”**
- Imputation via reward prediction
- Counterfactual model and selection bias
- Inverse propensity scoring (IPS) estimator

**Approach 2: “Model the bias”**
- Regression for reward prediction
- Offline Evaluating of Online Metrics

**Evaluating Online Metrics Offline**

- Online: On-policy A/B Test
- Offline: Off-policy Counterfactual Estimates

**Approach 1: Reward Predictor**
- Idea:
  - Use \( S = \{ (x_1, y_1, \delta_1), ..., (x_n, y_n, \delta_n) \} \)
  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 \( S^* = \{ (x_1, y_1, \hat{\delta}(x_1, y_1)), ..., (x_n, y_n, \hat{\delta}(x_n, y_n)) \} \)
  - Estimate performance of \( \pi \) via \( \bar{U}_{sp}(\pi) = \frac{1}{n} \sum_{i=1}^{n} \hat{\delta}(x_i, y_i) \)
- Stochastic \( \pi \): \( \bar{U}_{sp}(\pi) = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{d} \hat{\delta}(x_i, y_j) \pi(y_j|x_i) \)

**Regression for Reward Prediction**

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** \( \delta \): “Satisfaction”
  - Complicated hidden function
- **Logging policy** \( \pi_0 \): Non-uniform randomized logging system
  - Placket-Luce “explore around current production ranker”
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
  – π₀'s actions are over-represented

\[ \bar{U}_{rp}(\pi) = \frac{1}{n} \sum \delta(x_i, \pi(x_i)) \]

Evaluation: Outline

• Offline Evaluating of Online Metrics
  – A/B Testing (on-policy)
    \[ \rightarrow \] Counterfactual estimation from logs (off-policy)
• Approach 1: “Model the world”
  – Imputation via reward prediction
• Approach 2: “Model the bias”
  – Counterfactual model and selection bias
    – Inverse propensity scoring (IPS) estimator

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.

\[ U(\pi_0) = \int \int \delta(x,y) \pi_0(y|x) \pi_0(x) \, dx \, dy \]

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
  – Which treatment is best?

Placing Vertical

Click / no Click on SERP
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
    - → Drugs 3/4, Stent 2/3, Bypass 2/4 – really?

Assignment Mechanism

- Probabilistic Treatment Assignment
  - For patient i: \( n_i = y_i(x_i) \)
  - Selection Bias
- Inverse Propensity Score Estimator
  - \( \theta_{ips}(y) = \frac{1}{n_i \pi(x_i)} \)
  - Propensity: \( p_i = n_i \pi(x_i) \)
  - Unbiased: \( \bar{\theta}_{ips}(y) = \frac{\theta_{ips}(y)}{\hat{p}} \)
  - Example
    - \( \bar{\theta}_{drugs} = \frac{\theta_{drugs}}{\hat{p}} = 0.36 < 0.75 \)

Experimental vs Observational

- Controlled Experiment
  - Assignment Mechanism under our control
    - Propensities \( p_i = n_i \pi(x_i) \) known by design
    - Requirement: \( Y: \pi_0(y_i) = y_i(x_i) > 0 \) (probabilistic)
- Observational Study
  - Assignment Mechanism not under our control
    - Propensities \( p_i \) need to be estimated
    - Estimate \( \theta_{ips}(y_i) = n_i \pi_0(y_i) \) based on features \( x_i \)
    - Requirement: \( \theta_{ips}(y_i) = \theta_{ips}(y_i(x_i)) \) (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:
      - \( \pi(A) = drugs, \pi(B) = stent, \pi(C) = bypass \)
- Average Treatment Effect
  - \( \tau(y) = \sum x_i \delta(x_i, \pi(x_i)) \)
- IPS Estimator
  - \( \theta_{ips}(y) = \frac{1}{n_i} \sum \delta(x_i, \pi(x_i)) / \hat{p} \)

Stochastic Treatment Policies

- Policy (stochastic)
  - Context \( x_i \) describing patient
    - Pick treatment \( y_i \) based on \( x_i: \pi(y_i|x_i) \)
    - Note
      - Assignment Mechanism is a stochastic policy as well!
- Average Treatment Effect
  - \( \tau(y) = \sum x_i \delta(x_i, \pi(y|x_i)) \)
- IPS Estimator
  - \( \theta_{ips}(y) = \frac{1}{n_i} \sum \pi(y|x_i) / \hat{p} \)

Counterfactual Outcome

- Average Treatment Effect of Treatment \( y \)
  - \( \tau(y) = \frac{1}{n_i} \sum \delta(x_i, y) \)
- Example
  - \( \tau(bypasses) = \frac{4}{11} \)
  - \( \tau(stent) = \frac{6}{11} \)
  - \( \tau(drugs) = \frac{3}{11} \)

Counterfactual Outcomes

- Control
  - \( x \)
  - \( Y \)
  - \( \pi(x) \)
  - \( \delta \)

Experimental Outcome

- Counterfactual
  - \( x \)
  - \( Y \)
  - \( \pi(x) \)
  - \( \delta \)
Counterfactual Model = Logs

<table>
<thead>
<tr>
<th>Context $x_i$</th>
<th>Treatment $y_i$</th>
<th>Outcome $\delta_i$</th>
<th>Propensities $p_i$</th>
<th>New Policy $\pi$</th>
</tr>
</thead>
</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

System Evaluation via Inverse Propensity Score Weighting

Definition [IPS Utility Estimator]: Given $S = \{(x_i, y_i, \delta_i) \ldots, (x_n, y_n, \delta_n)\}$ collected under $\pi_o$,

$$\bar{U}_{IPS}(\pi) = \frac{1}{n} \sum_{i=1}^{n} \delta_i \frac{\pi(y_i|x_i)}{\pi_o(y_i|x_i)}$$

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

Note:
- If $\pi = \pi_o$, then online A/B Test with $\bar{U}_{IPS}(\pi_o) = \frac{1}{n} \sum_{i=1}^{n} \delta_i$
- Off-policy vs. On-policy estimation.

IPS Estimator is Unbiased

$$E[\bar{U}_{IPS}(\pi)] = \frac{1}{n} \sum_{i=1}^{n} \sum_{x_{-i}} \sum_{y_{-i}} \sum_{\delta_{-i}} p(x_{-i}, y_{-i}, \delta_{-i}) \bar{U}_{IPS}(\pi)$$

Unbiased:
- if $\forall x, y, \pi(y|x) \pi_o(x) > 0$ then $E[\bar{U}_{IPS}(\pi)] = E(\pi)$

News Recommender: Results

IPS eventually beats RP; variance decays as $O\left(\frac{1}{n}\right)$
Counterfactual Policy Evaluation

- Controlled Experiment Setting:
  - Log data: $D = \{(x_1, y_1, \delta_1, p_1), ..., (x_n, y_n, \delta_n, p_n)\}$
- Observational Setting:
  - Log data: $D = \{(x_1, y_1, \delta_1, z_1), ..., (x_n, y_n, \delta_n, z_n)\}$
  - Estimate propensities: $p_i = P(Y_i|X_i, Z_i)$ based on $x_i$ and other confounders $z_i$

Goal: Estimate average treatment effect of new policy $\pi$.
- IPS Estimator
  $$\psi(\sigma) = \frac{1}{n} \sum_{i=1}^{n} \frac{\pi(Y_i|X_i, Z_i)}{p_i}$$
or many others.

Evaluation: Summary

- Offline Evaluation of Online Metrics
  - A/B Testing (on-policy)
  - Counterfactual estimation from logs (off-policy)
- Approach 1: "Model the world"
  - Estimation via reward prediction
  - Pro: low variance
  - Con: model mismatch can lead to high bias
- Approach 2: "Model the bias"
  - Counterfactual Model
  - Inverse propensity scoring (IPS) estimator
  - Pro: unbiased for known propensities
  - Con: large variance

From Evaluation to Learning

- Naïve "Model the World" Learning:
  - Learn: $\delta: x \times y \rightarrow \mathbb{R}$
  - Derive Policy:
    $$\pi(y|x) = \arg \min_{y'} \delta(x, y')$$
- Naïve "Model the Bias" Learning:
  - Find policy that optimizes IPS training error
    $$\pi = \arg \min_{\pi} \left[ \sum_{i=1}^{n} \frac{\pi(Y_i|X_i)}{\pi(Y_i|X_i)} \delta_i \right]$$