Learning to Act and Causality

CS4780/5780 – Machine Learning
Fall 2019

Nika Haghtalab & 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

Music Voice Assistant

- Context $x$:
  - User and speech
- Action $y$:
  - Track that is played
- Feedback $\delta(x, y)$:
  - Listened to the end

Search Engine

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

Log Data from Interactive Systems

- Data
  $S = (x_1, y_1, \delta_1, \ldots, x_n, y_n, \delta_n)$

  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 $\delta: X \times Y \rightarrow \mathbb{R}$

[Zadrozny et al., 2003] [Langford & Li]
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 \( \pi \)

**Goal**

**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

**Example metrics**

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

This lecture:

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

**Online Performance Metrics**

**System Performance**

Definition [Utility of Policy]:

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

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

\( \delta(x, y) \) is reading time of user \( x \) for portfolio \( y \).

**Evaluation: A/B Testing**

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

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

\( \Rightarrow \) 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 \( \delta(x, y) \)

\[\vdots\]

Deploy \( \pi_{|H|} \): Draw \( x \sim P(X) \), predict \( y \sim \pi_{|H|}(Y|x) \), get \( \delta(x, y) \)

**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 \( x \)
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 π₁
  - Number of A/B Tests limited
  - Long turnaround time

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

Approach 1: Reward Predictor

- **Idea:**
  - Use \( S = \{(x_1, y_1, \delta_1), \ldots, (x_n, y_n, \delta_n)\} \) from \( \pi_o \) to estimate reward predictor \( \delta(x, y) \)
  - 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 \( S' = \{(x_1, y'_1, \delta(x_1, y'_1)), \ldots, (x_n, y'_n, \delta(x_n, y'_n))\} \)
    - Estimate performance of \( \pi \) via \( \hat{U}_{\text{rp}}(\pi) = \frac{1}{n} \sum_{i=1}^{n} \delta(x_i, y'_i) \pi(y'_i|x_i) \)
  - Stochastic \( \pi \): \( \hat{U}_{\text{rp}}(\pi) = \frac{1}{n} \sum_{i=1}^{n} \sum_{y \in \mathcal{Y}} \delta(x_i, y) \pi(y|x_i) \)

Regression for Reward Prediction

- **Learn** \( \hat{\delta}: x \times y \rightarrow \Re \)
  - **1. Represent via features** \( \Psi_1(x, y) \)
  - **2. Learn regression based on** \( \Psi_1(x, y') \) from \( S \) collected under \( \pi_o \)
  - **3. Predict** \( \hat{\delta}(x, y') \) for \( y' = \pi(x) \) of new policy \( \pi \)

Problems of Reward Predictor

- **Modeling bias**
  - choice of features and model
- **Selection bias**
  - \( \pi_o \)'s actions are over-represented
  - \( \hat{U}_{\text{rp}}(\pi) = \frac{1}{n} \sum_{i=1}^{n} \delta(x_i, \pi(x_i)) \)
  - Can be unreliable and biased
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 estimation from logs (off-policy)
  - 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.

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

Potential Outcome 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?

Potential Outcome 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?

Treatment Effects

- Average Treatment Effect of Treatment \( y \)
  - \( U(y) = \frac{1}{n} \sum \delta(x_i, y) \)
- Example
  - \( U(\text{bypass}) = \frac{5}{11} \)
  - \( U(\text{stent}) = \frac{7}{11} \)
  - \( U(\text{drugs}) = \frac{3}{11} \)
Assignment Mechanism

- Probabilistic Treatment Assignment
  - For patient $i$, $\pi_i(y_i | x_i)$
  - Selection Bias
- Inverse Propensity Score Estimator
  - $\hat{\theta}_{IPS}(y_i) = \frac{1}{n} \sum \delta(y_i = y | x_i)$
  - Propensity: $p_i = \pi_i(y_i | x_i)$
  - Unbiased: $\hat{\theta}(y_i | x_i)$
  - Estimator $\hat{\theta}_{IPS}(y_i)$
  - Example

Interventional vs Observational

- Interventional Controlled Experiment
  - Assignment Mechanism under control
  - Propensities $p_i = \pi_i(y_i | x_i)$ known by design
  - Requirement: $\forall y: \pi_i(y_i = y | x_i) > 0$ (probabilistic)
- Observational Study
  - Assignment Mechanism not under control
  - Propensities $p_i$ need to be estimated
  - Estimate $\hat{\theta}_o(y_i | x_i) = \pi_o(y_i | x_i)$ based on features $x_i$
  - Requirement: $\hat{\theta}_o(y_i | x_i) = \pi_o(y_i | x_i)$ (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:
    - $\pi(A) = \text{drugs}, \pi(B) = \text{stent}, \pi(C) = \text{bypass}$
- Average Treatment Effect
  - $U(\pi) = \frac{1}{n} \sum \delta(x_i, \pi(x_i))$
- IPS Estimator
  - $\hat{\theta}_{IPS}(\pi) = \frac{1}{n} \sum \delta(y_i = \pi(x_i))$

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
  - $U(\pi) = \frac{1}{n} \sum \delta(x_i, y) \pi(y | x_i)$
- IPS Estimator
  - $\hat{\theta}_{IPS}(\pi) = \frac{1}{n} \sum \delta(x_i, y) \pi(y | x_i)$

Counterfactual Model = Logs

- 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

Evaluation: Outline

- $\pi_a(y_i = y | x_i)$
  - Counterfactual treatment
- $\delta(x_i, y_i)$
  - Outcome
- $p_i$ (Propensities)
  - New Policy $\pi$
System Evaluation via Inverse Propensity Score Weighting

Definition (IPS Utility Estimator):
Given \( S = \{(x_i, y_i, \delta_i), ..., (x_n, y_n, \delta_n)\} \) collected under \( \pi_0 \),
\[
\hat{U}_{IPS}(\pi) = \frac{1}{n} \sum_{i=1}^{n} \delta_i \frac{\pi(y_i | x_i)}{\pi_0(y_i | x_i)}
\]
\( \Rightarrow \) 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 \( \Rightarrow \) Off-policy vs. On-policy estimation.

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

(Hassan & Thompson, 1982; Rubin, 1983; Zadrozny, et al., 2003)

IPS Estimator is Unbiased

\[ E[\hat{U}_{IPS}(\pi)] = \frac{1}{n} \sum_{i=1}^{n} \left[ \sum_{x,y} \pi(y|x) \pi_0(x | y) \delta_i \cdot \frac{\pi(y|x)}{\pi_0(y|x)} \right] \]
- Independent
- Marginal
- Full support
- Identical x,y

Illustration of IPS

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

Counterfactual Policy Evaluation

- Controlled Experiment Setting:
  - Log data: \( D = \{(x_1, y_1, \delta_1), ..., (x_n, y_n, \delta_n)\} \)
  - observational Setting:
  - Log data: \( D = \{(x_1, y_1, z_1), ..., (x_n, y_n, z_n)\} \)
  - Estimate propensities: \( \pi_i = \pi(y | x, z) \) based on \( x \) and other confounders \( z \)

Goal: Estimate average treatment effect of new policy \( \pi \).
- IPS Estimator
\[
\hat{U}(\pi) = \frac{1}{n} \sum_{i=1}^{n} \delta_i \frac{\pi(y_i | x_i)}{\pi_i}
\]
or many others.

Evaluation: Summary

- Offline Evaluation of Online Metrics
  - A/B Testing (on-policy)
  - \( \Rightarrow \) 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

Setting: Batch Learning from Bandit Feedback (BLBF)

- "Model the World" Learning:
  - Learn: \( \delta: x \times y \rightarrow \mathbb{R} \)
  - Derive Policy:
  \[
  \pi(y | x) = \arg\min_{y'} \delta(x, y')
  \]
- "Model the Bias" Learning:
  - Find policy that optimizes IPS training error
  \[
  \pi = \arg\min_{\pi} \sum_{i=1}^{n} \delta_i \frac{\pi(y_i | x_i)}{\pi_0(y_i | x_i)}
  \]