Counterfactual Learning with General Logging Policies

Yusuke Narita [1] Kyohei Okumura [2] Akihiro Shimizu [3] Kohei Yata [4]

[1]Yale University [2]Northwestern University [3]Mercari, Inc. [4]University of Wisconsin-Madison

Summary

Off-policy evaluation (OPE)

- Predicts the performance of counterfactual policies using log data from a different policy.
- Advantages over A/B test: **fast, cheap,** and **safe**.

Problem

- Can we conduct OPE with general logging policies including deficient support logging policies?

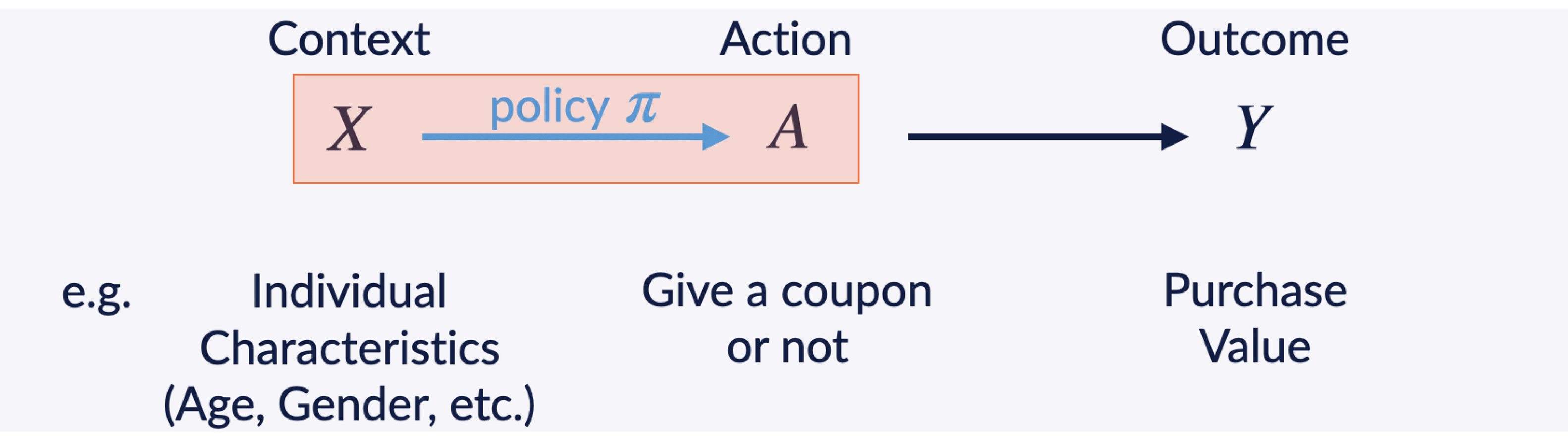
Method

- A new OPE estimator based on a modification of the Propensity Score, *Approximate Propensity Score (APS)*.
- Our estimator converges to the true performance of a counterfactual policy as the sample size increases.

Real-World Application

- Apply our method to data from Mercari, Inc. to evaluate coupon targeting policies

Framework: Off-Policy Evaluation



- Logging policy ML generates log data $(Y_i, X_i, A_i)_{i=1}^n$:
  1. $(Y_i(\cdot), X_i)$ is i.i.d.-drawn from an unknown distrib.
  2. Action $A_i \in \{1, ..., m\}$ is chosen w.p. $ML(A_i \mid X_i)$
  3. Reward $Y_i := Y_i(A_i)$ is observed.
- ML can be of deficient support, i.e., it is possible that $ML(a \mid x) \in \{0, 1\}$ for some action a and context x.

Goal: Estimate the performance $V(\pi)$ of a counterfactual policy π using the log data:

$$V(\pi) := E\left[\sum_{a=1}^{m} Y(a)\pi(a \mid X)\right]$$

Key: Approximate Propensity Score (APS)

Given logging policy ML and a bandwidth $\delta > 0$,

$$p_\delta^{ML}(a|X_i) := \frac{\int_{B(X_i,\delta)} ML(a|x^*)dx^*}{\int_{B(X_i,\delta)} dx^*},$$

where $B(X_i, \delta)$ is a p-dimensional ball with radius δ centered at $X_i \in \mathbb{R}^p$.

- APS is the the average probability that the logging policy chooses action a in a neighborhood around X_i.

OPE Estimator

1. For a small bandwidth δ, compute APS $p_\delta^{ML}(a|X_i)$. Let $q_\delta^{ML}(a|X_i) := \frac{p_\delta^{ML}(a|X_i)}{p_\delta^{ML}(a|X_i)+p_\delta^{ML}(1|X_i)}$.
2. For each $a = 2, ..., m$, minimize the sum of squared errors on the subsample $\mathcal{I}(a;\delta) := \{i : A_i \in \{1, a\}, q_\delta^{ML}(a|X_i) \in (0, 1)\}$:
$$\min_{(\alpha_a,\beta_a,\gamma_a)} \sum_{i\in\mathcal{I}(a;\delta)} \left(Y_i - \alpha_a - \beta_a 1\{A_i = a\} - \gamma_a q_\delta^{ML}(a|X_i)\right)^2,$$
where $1\{\cdot\}$ is the indicator function.
3. Define our OPE estimator for $V(\pi)$ as:
$$\hat{V}(\pi) = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i + \sum_{a=2}^{m}\hat{\beta}_a(\pi(a|X_i) - ML(a|X_i))\right).$$

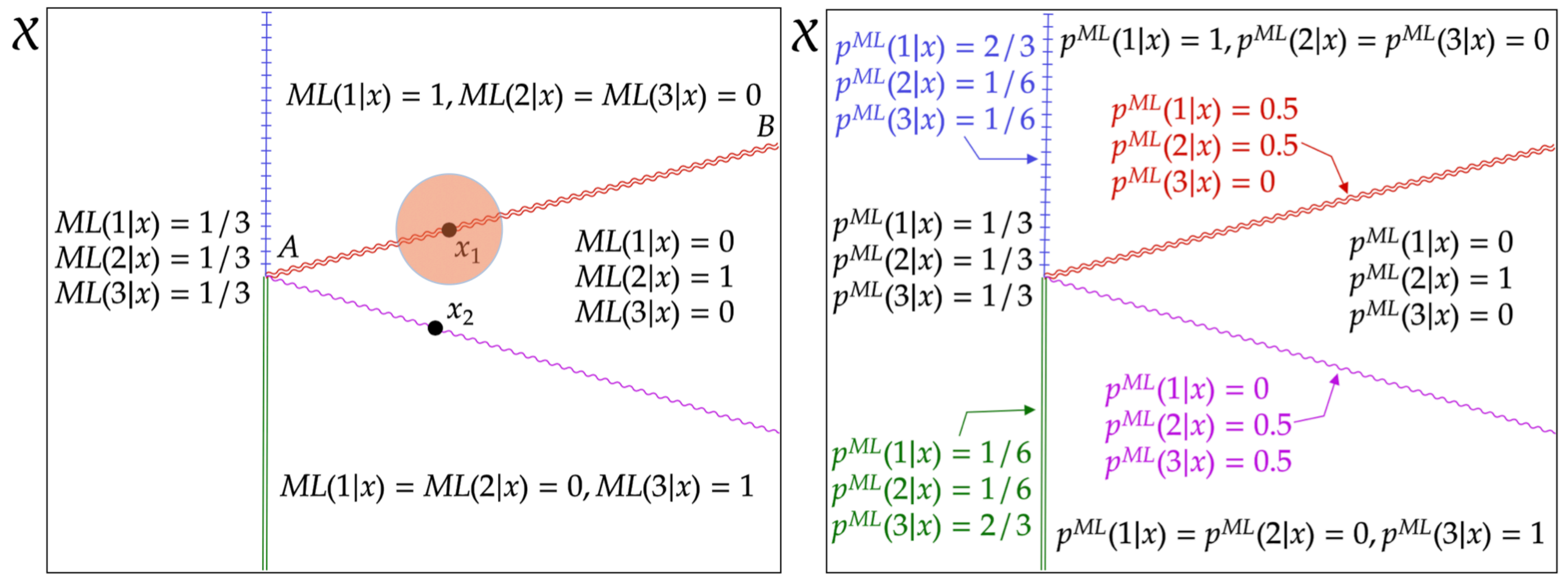
Theorem (Consistency)

Under some assumptions, $\hat{V}(\pi)$ converges in probability to $V(\pi)$ as $n \to \infty$.

Key Assumption

Constant Conditional Mean Reward Differences

For all a, a', $E[Y(a)|X = x] - E[Y(a')|X = x]$ is constant over $x \in \mathcal{X}$.



Real-World Application: Coupon Targeting Policy at Mercari, Inc.

Data

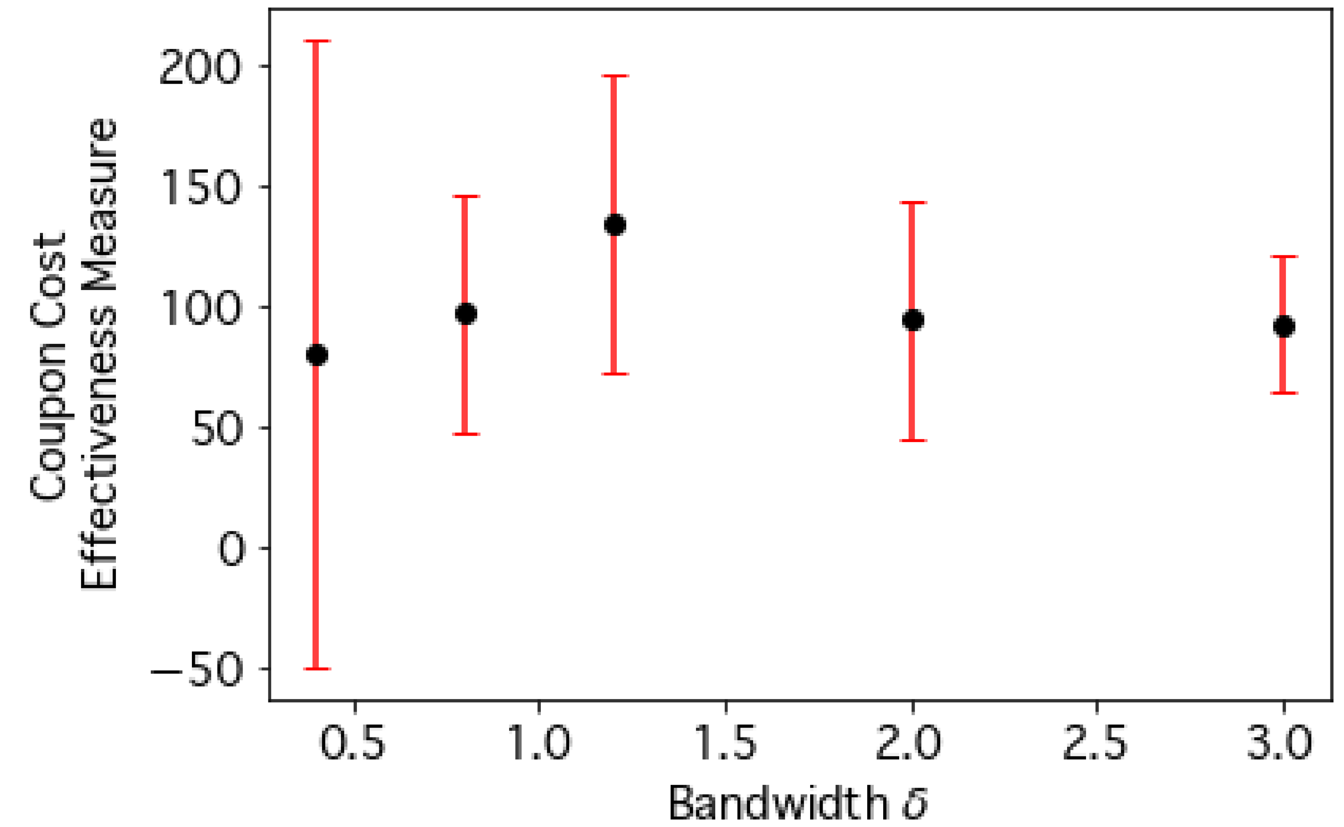
- Y_i: purchase value, # of transaction, or point usage over 18 days after the coupon offer decision.
- X_i: more than 200 features
- $A_i \in \{0, 1\}$: whether or not receive coupon

Deficient Support Logging Policy

1. Train an uplift model τ using data from a past A/B test.
2. Offer a coupon if $\tau(X_i)$ is in the top 80%:
$$ML(1|X_i) := 1\{\tau(X_i) \geq q_{0.2}\}.$$

Coupon Cost Effectiveness Measure:

How much would the total purchase value increase in USD if we increased the cost by 1 USD? — 80–134 USD.



kyohei.okumura@u.northwestern.edu