Recommender Systems
lecture 8: counterfactual evaluation

Alexey Grishanov

Moscow Institute of Physics and Technology

Spring 2023

Today's outline

Motivation

Have you ever wanted to return back and explore different decisions?

Problem

Don't our recommendations change how customers click or purchase?
If customers can only interact with items shown to them, why do we perform offline evaluation on static historical data?

Objective

What would have happened if we show users our new recommendations instead of the existing strategy?

Off-policy evaluation (OPE)

Given:

logged dataset $\mathcal{D}$ obtained from policy $\pi_0 = \pi_0(a|x)$

$$\mathcal{D} = \{(x_i, a_i, r_i)\}_{i=1}^n \sim \prod_{i=1}^n p(x_i)\ \pi_0(a_i|x_i)\ p(r_i|x_i, a_i)$$

- x_i — sample state from states X
- a_i — sample action from π_0 on x_i (what we can control)
- r_i — sample reward when the state is x_i and action is a_i

Estimate:

value of policy π_{test} given data $\mathcal{D}$

$$\hat{V}(\pi_{test}, \mathcal{D})$$

close to true policy value

$$V(\pi_{test}) = \mathbb{E}_{p(x)\ \pi_{test}(a|x)\ p(r|x,a)}[r]$$

A/B testing

Deploy policy π_{test} in production to get an online estimation of performance

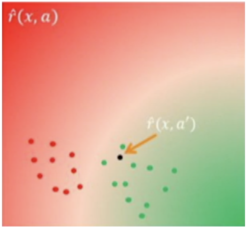
$$\mathcal{D}_{test} \sim \prod_{i=1}^{n} p(x_i)\ \pi_{test}(a_i|x_i)\ p(r_i|x_i, a_i)$$

Estimate

$$\hat{V}_{A/B}(\pi_{test}, \mathcal{D}_{test}) = \frac{1}{n} \sum_{i=1}^{n} r_i$$

Straightforward but takes time and risks

Direct Method (simulators)



Learn reward (response) model using $\{(x_i, a_i, r_i)\}_{i=1}^n$

$$r(x, a) = \mathbb{E}(r|x, a) \approx \hat{r}(x, a)$$

Use modeled rewards for actions selected by π_{test}

$$\hat{V}_{DM}(\pi_{test}, \mathcal{D}, \hat{r}(x, a)) = \frac{1}{n} \sum_{i=1}^{n} \sum_{a \in \mathcal{A}} \pi_{test}(a|x_i)\ \hat{r}(x_i, a)$$

Simulators typically has low variance, but building response function $\hat{r}(x, a)$ with low bias is hard (active research direction)

source

Importance sampling example (towards lower bias)

		Probability(Red)	Probability(Green)	
π_0 (policy used when data were collected)	X 0.25	80%	20%	X 4
New policy π_{Test}		20%	80%	

		Clicks on Red	Clicks on Green	
π_0	X 0.25	1000	300	X 4
π_{Test}		?	?	

Number of clicks for π_{test}

$$\text{Clicks}_{\text{green}} \cdot \frac{\pi_{\text{test}}(\text{green})}{\pi_0(\text{green})} + \text{Clicks}_{\text{red}} \cdot \frac{\pi_{\text{test}}(\text{red})}{\pi_0(\text{red})} = 300 \cdot \frac{0.8}{0.2} + 1000 \cdot \frac{0.2}{0.8} = 1450$$

IPS

$$\hat{V}_{IPS}(\pi_{test}, \mathcal{D}) = \frac{1}{n} \sum_{i=1}^{n} \frac{\pi_{test}(a_i, x_i)}{\pi_0(a_i, x_i)} \cdot r_i$$

Excersize

Prove that IPS is unbiased

Issues

- problems when π_{test} recommend a which π_0 didn't make
- high varience when π_{test} far from π_0, e.g. $p_0 = 0.001, p_{test} = 0.1$

CIPS

One solution is to ensure that the new recommenders being evaluated don't differ too much from the production recommender

$$\hat{V}_{\mathrm{CIPS}}\left(\pi_{test}, \mathcal{D}_0, \lambda\right) = \frac{1}{n} \sum_{i=1}^{n} \underbrace{\min \left\{ \frac{\pi_{test}\left(a_i \mid x_i\right)}{\pi_0\left(a_i \mid x_i\right)}, \lambda \right\}}_{\text{upper bounded by } \lambda} \cdot r_i$$

Issues

- lower variance than IPS but neigher biased nor consistent
- requires tuning λ

SNIPS

Avoid unstable estimation by rescaling:

$$\hat{V}_{\text{SNIPS}}\left(\pi_{test}, \mathcal{D}_0\right) = \frac{\sum_{i=1}^{n} \frac{\pi_{test}(a_i|x_i)}{\pi_0(a_i|x_i)} \cdot r_i}{\underbrace{\sum_{i=1}^{n} \frac{\pi_{test}\left(a_i \mid x_i\right)}{\pi_0\left(a_i \mid x_i\right)}}_{\text{empirical mean of weights}}}$$

Issue

- consistent but not unbiased

Doubly robust estimator

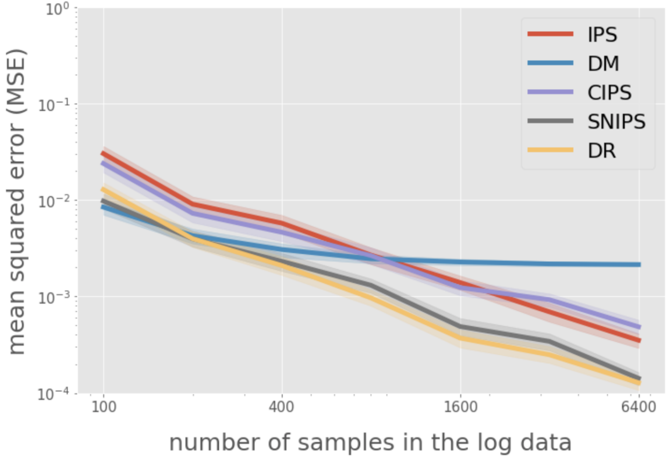
Combine DM and SNIPS:

$$\hat{V}_{\mathrm{DR}}\left(\pi_{test}, \mathcal{D}_0, \hat{r}\right) = \hat{V}_{\mathrm{DM}}\left(\pi_{test}, \mathcal{D}_0, \hat{r}\right) + \frac{1}{n} \sum_{i=1}^{n} \frac{\pi_{test}\left(a_i \mid x_i\right)}{\pi_0\left(a_i \mid x_i\right)}\left(r_i - \hat{r}\left(x_i, a_i\right)\right)$$

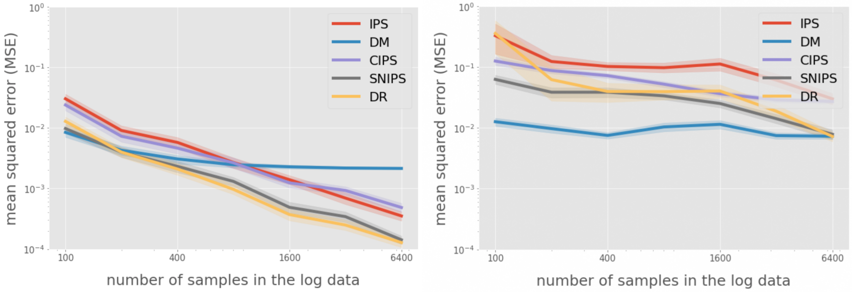
Review

- unbiased and consistent
- potential to reach optimal varience (informal)
- useful when using estimated propensities $\hat{p}_i \approx \pi_0(a_i|x_i)$
- default in Vowpal Wabbit

Comparing OPE



Similar vs far policies



- left: max importance weight: 18.60
- right: max importance weight: 451.13

Preparing for off-policy evaluation

Log everything:

$$\langle x_i, a_i, r_i, p_i \rangle ,$$

where $p_i = \pi_0(a_i, x_i)$ — propensities

If impossible to get p_i, log enough to estimate $\hat{p}_i$:

- candidate set of actions
- features for each candidate
- exploration parameters
- etc.

Lecture summary

Approach 0: «A/B test»

Estimation via model deployment

- Pro: unbiased $\mathbb{E}_{\mathcal{D}}\left[\hat{V}_{A/B}(\pi, \mathcal{D})\right] = V(\pi)$
- Con: costly to obtain

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

- Pro: unbiased for known propensities
- Con: suffer when the recommendation policy to be evaluated is far from the logging policy

Literature

1. *P. Rosenbaum et al.* (1983) The Central Role of Propensity in Observational Studies for Causal Effects (IPS)
2. *Adith Swaminathan, Thorsten Joachims* (2015) Batch Learning from Logged Bandit Feedback through Counterfactual Risk Minimization (CIPS)
3. *Adith Swaminathan, Thorsten Joachims* (NIPS 2015) The Self-Normalized Estimator for Counterfactual Learning (SNIPS)
4. RecSys 2021 tutorial
5. SIGIR 2016 tutorial
6. *Y. Saito et al.* (NeurIPS'21) Open Bandit Dataset and Pipeline: Towards Realistic and Reproducible Off-Policy Evaluation

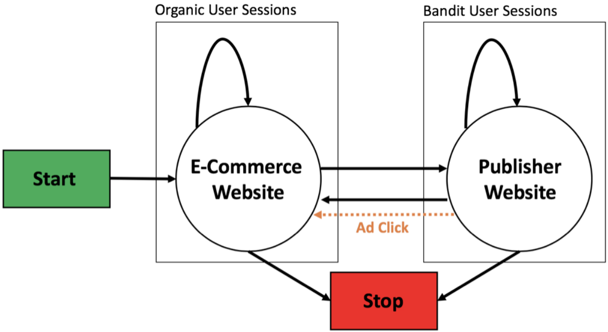


Figure 1: Markov Chain of the organic and bandit user sessions

source