Doubly Robust Estimator for Ranking Metrics with Post-Click Conversions

ACM Conference on Recommender Systems (RecSys'20)

Yuta Saito (https://usaito.github.io/)

Tokyo Institute of Technology

Introduction & Problem Setting

Motivation: Offline Evaluation with Click -> Conversion data

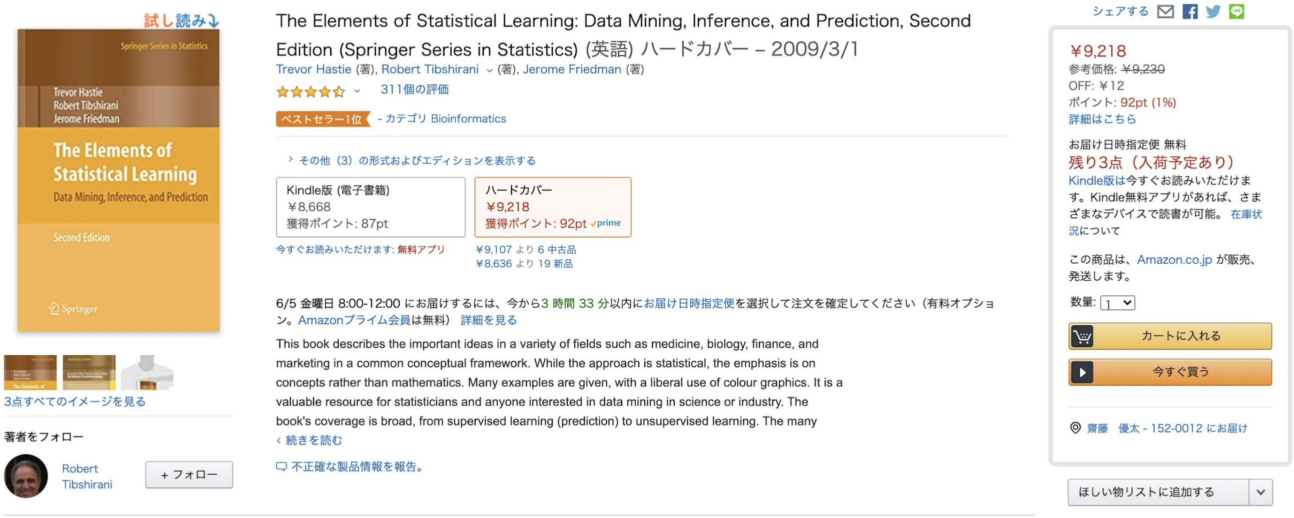
In an Amazon example, a user first **click** the item in a recommendation list

- **query: “statistics”**
- **click “ESL” here**
- **click itself is not our outcome**

Motivation: Offline Evaluation with Click -> Conversion data

We observe the conversion indicator only for an item with a click

User's intended action on the item is revealed as a conversion indicator



Motivation: Offline Evaluation with Click -> Conversion data

Recommend **Items with high conversion rate (CVR)**

example) Top-3 Recommendation in E-commerce

Ranking	Recommender A	Recommender B
1	CV=1	CV=0
2	CV=1	CV=1
3	CV=1	CV=0
———	———	———
9	CV=0	CV=1
10	CV=0	CV=1

Recommender A
is better than
Recommender B

simply because

Recommender A
creates a list of more conversions

Motivation: Offline Evaluation with Click -> Conversion data

Recommend Items with high conversion rate (CVR)

example) Top-3 Recommendation in E-commerce

Ranking	Recommender A	Recommender B
1	missing	missing
2	CV=1	missing
3	missing	CV=0
———	———	———
9	missing	CV=1
10	CV=0	missing

We cannot use conversion indicators for unclicked items in offline evaluation

Ground-truth Ranking Performance

We want to calculate the *ground-truth ranking measure* to evaluate the ranking performance of recommenders offline

conversion rate of u and i

$$\mathcal{R}_{GT}(\hat{Z}) = \frac{1}{|\mathcal{U}|} \sum_{u \in \mathcal{U}} \sum_{i \in \mathcal{I}} p^{cvr}_{u,i} \cdot c(\hat{Z}_{u,i})$$

a set of predicted rankings for user-item paris

user

item

ranking function (weighting function)

Ground-truth Ranking Performance

The function c(.) characterizes ranking metrics

Average Relevance Position: $c(\hat{Z}_{u,i}) = \hat{Z}_{u,i}$

Discounted Cumulative Gain: $c(\hat{Z}_{u,i}) = \log_2(1 + \hat{Z}_{u,i})^{-1}$

where Z is the predicted ranking for a user-item pair

$$\hat{Z}_{u,i} = \mathrm{rank}(\hat{S}_{u,i} \mid \{\hat{S}_{u,j}\}_{j \in \mathcal{I}})$$

Offline Evaluation of Recommenders in E-commerce settings

It is desirable to use the ground-truth ranking metric
to identify a recommender that can obtain the maximum CVs

Offline Evaluation of Recommenders in E-commerce settings

It is **desirable to use the ground-truth ranking metric** to identify a recommender that can obtain the maximum CVs

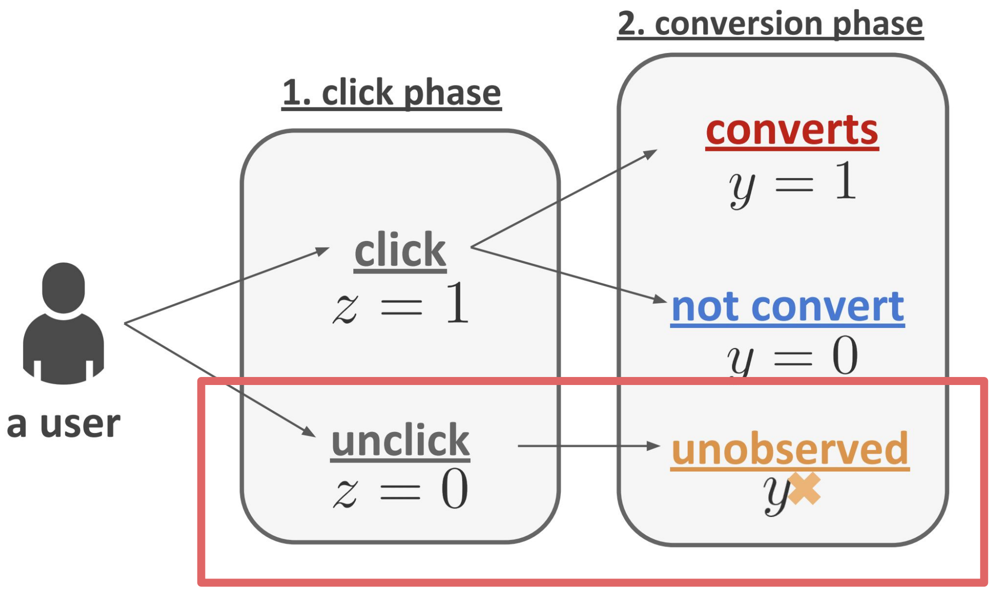
However, there are several difficulties in evaluating recommenders in an offline environment, including...

- missing, sparse conversions
- selection bias issue

Challenge 1: Missing, Sparse Conversions

Users first **click** the item then they decide whether they should **convert**

When a click does not happen, then the **conversion is unobserved**

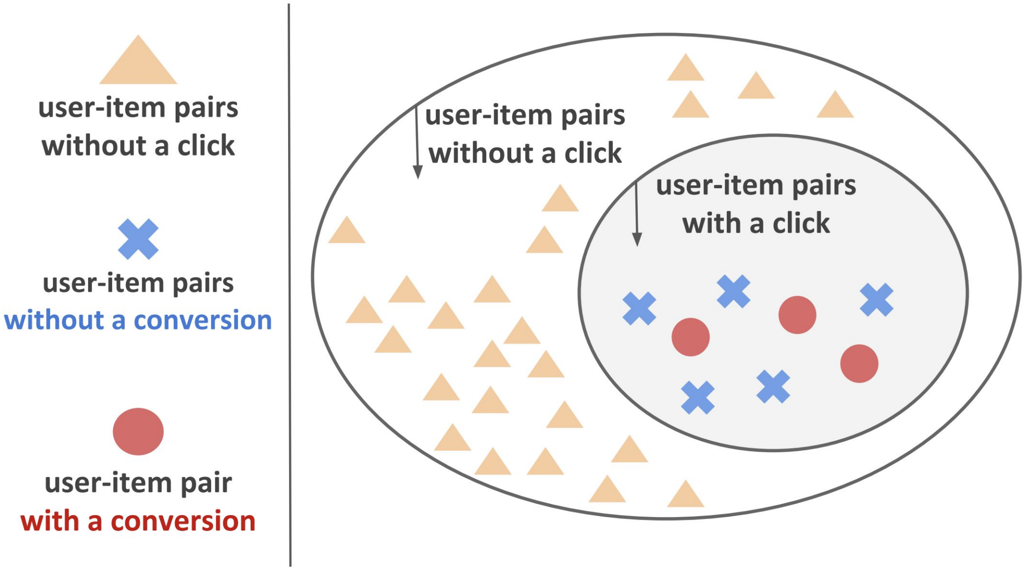


(b) user behavior pattern

Challenge 2: Selection Bias

We can use only conversions with a click in offline eval

Observed data is biased and **not representative of the whole data**



(a) selection bias problem

In summary,

It is essential to estimate the ground-truth using only observed CVs

Ground-truth:

$$\mathcal{R}_{GT}(\hat{Z}) = \frac{1}{|\mathcal{U}|} \sum_{u \in \mathcal{U}} \sum_{i \in \mathcal{I}} \underline{p_{u,i}^{cvr}} \cdot c(\hat{Z}_{u,i})$$

An Estimator:

$$\hat{\mathcal{R}}(\hat{Z}) = \frac{1}{|\mathcal{U}|} \sum_{u \in \mathcal{U}} \sum_{i \in \mathcal{I}} \underline{???} \; c\left(\hat{Z}_{u,i}\right)$$

In summary,

It is essential to estimate the ground-truth using only observed CVs

Using offline (observable) data:

$$\{(u, i, \underbrace{y_{u,i}}_{\text{conversion indicator}}) \mid \underbrace{z_{u,i} = 1}_{\text{with a click}}\}$$

Solutions & Experiments

A Previous Solution: IPS Estimator

(Yang et al. 2018) proposed *the IPS estimator* to estimate the ground-truth ranking metrics

$$\hat{\mathcal{R}}_{IPS}(\hat{Z}) = \frac{1}{|\mathcal{U}|} \sum_{u \in \mathcal{U}} \sum_{i \in I : z_{u,i} = 1} \frac{y_{u,i}}{p^{ctr}_{u,i}} c\left(\hat{Z}_{u,i}\right)$$

using only clicked data

weight conversions by the inverse of the CTRs

Pros and Cons of the IPS Estimator

The IPS estimator is *unbiased* for the ground-truth ranking metrics

$$\mathbb{E}\left[\widehat{\mathcal{R}}_{IPS}(\widehat{Z})\right] = \mathcal{R}_{GT}(\widehat{Z})$$

but, the variance is huge, when conversions are highly sparse

THEOREM 3.3. *(Variance of the IPS estimator) When the set of true CTRs and scoring set $\hat{Z}$ are given, the variance of the IPS estimator is*

$$\mathbb{V}\left(\widehat{\mathcal{R}}_{IPS}(\widehat{Z})\right) = \frac{1}{|\mathcal{U}|^2} \sum_{u \in \mathcal{U}} \sum_{i \in \mathcal{I}} \left(\frac{1}{p^{ctr}_{u,i}} - p^{cvr}_{u,i}\right) p^{cvr}_{u,i}\, c(\hat{Z}_{u,i})^2$$

Our Approach: Doubly Robust Estimator

To alleviate the variance issue of IPS,
we propose the following *doubly robust* estimator

$$\widehat{\mathcal{R}}_{DR}(\widehat{Z}) = \frac{1}{|\mathcal{U}|} \sum_{u \in \mathcal{U}} \sum_{i \in \mathcal{I}} \left(\frac{z_{u,i}}{p^{ctr}_{u,i}} \left(y_{u,i} - \hat{p}^{cvr}_{u,i} \right) + \hat{p}^{cvr}_{u,i} \right) c(\hat{Z}_{u,i})$$

click indicator

inverse of CTR

estimated CVRs

Variance Reduction by the DR estimator

The DR estimator is also *unbiased* for the ground-truth ranking metrics

$$\mathbb{E}\left[\widehat{\mathcal{R}}_{DR}(\widehat{Z})\right] = \mathcal{R}_{GT}(\widehat{Z})$$

in most cases, the DR estimator has a lower variance

$$\mathbb{V}\left(\widehat{\mathcal{R}}_{DR}(\widehat{Z})\right) \leq \mathbb{V}\left(\widehat{\mathcal{R}}_{IPS}(\widehat{Z})\right)$$

Real-World Experiment (with Yahoo! R3 and Coat)

We compared the estimation performances of estimators

Yahoo! R3 and Coat datasets

- contain ***ground-truth relevance label*** (5 star-rating)
- contain train-test data with ***different item distributions***

These datasets are especially convenient for **the evaluation of offline evaluation** with the presence of selection bias

Performance measures for offline estimators

We used the following *relative-RMSE* to evaluate the performance of estimators

$$\textit{relative-RMSE}\ (\widehat{\mathcal{R}}) = \sqrt{\frac{1}{|\mathcal{M}|} \sum_{\hat{Z} \in \mathcal{M}} \left(\frac{\mathcal{R}_{GT}(\hat{Z}) - \widehat{\mathcal{R}}(\hat{Z})}{\mathcal{R}_{GT}(\hat{Z})} \right)^2}$$

an estimator to be evaluated

a set of 32 recommenders

Brief Experimental Results on Yahoo! and Coat

DR outperforms the others (lower values mean accurate evaluation!)

Table 4: Comparison of *relative-RMSE* (model evaluation performances) of alternative estimators

		DCG@K			Recall@K		
Datasets	Estimators	$K = 5$	$K = 10$	$K = 50$	$K = 5$	$K = 10$	$K = 50$
Yahoo! R3	Naive	0.613 (± 0.070)	0.470 (± 0.057)	0.245 (± 0.027)	0.615 (± 0.067)	0.442 (± 0.047)	0.207 (± 0.017)
	IPS	0.767 (± 0.022)	0.780 (± 0.024)	0.850 (± 0.015)	0.473 (± 0.040)	0.308 (± 0.032)	0.158 (± 0.013)
	DR (ours)	**0.461** (± 0.053)	**0.316** (± 0.040)	**0.181** (± 0.022)	**0.397** (± 0.042)	**0.261** (± 0.029)	**0.101** (± 0.011)
Coat	Naive	0.666 (± 0.037)	0.430 (± 0.013)	0.208 (± 0.005)	0.617 (± 0.027)	0.387 (± 0.011)	0.184 (± 0.004)
	IPS	0.785 (± 0.020)	0.805 (± 0.010)	0.915 (± 0.004)	0.605 (± 0.028)	0.374 (± 0.011)	0.181 (± 0.004)
	DR (ours)	0.661 (± 0.066)	**0.359** (± 0.020)	**0.137** (± 0.004)	0.599 (± 0.050)	**0.318** (± 0.014)	**0.118** (± 0.003)

* relative-RMSE measures the accuracy of offline evaluation, (not that of predictions)

Conclusions

- We study *offline evaluation with biased click -> conversion data*
- Previous unbiased estimator has a large variance
- We proposed *the doubly robust estimator* to estimate the ground-truth ranking performance efficiently
- Proposed estimator evaluates the performance of recommenders accurately in a real-world experiment

Thank you for listening!

theoretical analysis, semi-synthetic experiment, related work
are all in the full paper!

email: saito.y.bj@m.titech.ac.jp