

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

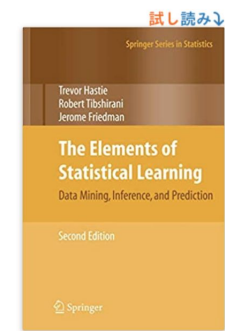
- query: “statistics”
- click “ESL” here
- click itself is not our outcome



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



著者をフォロー



Robert Tibshirani

+ フォロー

The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition (Springer Series in Statistics) (英語) ハードカバー – 2009/3/1

Trevor Hastie (著), Robert Tibshirani (著), Jerome Friedman (著)

★★★★☆ 311個の評価

ベストセラー1位 ← カテゴリ Bioinformatics

> その他 (3) の形式およびエディションを表示する

Kindle版 (電子書籍)
¥8,668
獲得ポイント: 87pt

今すぐお読みいただけます: 無料アプリ

ハードカバー
¥9,218
獲得ポイント: 92pt ✓prime

¥9,107 より 6 中古品
¥8,636 より 19 新品

6/5 金曜日 8:00-12:00 にお届けするには、今から3 時間 33 分以内にお届け日時指定便を選択して注文を確定してください (有料オプション。Amazonプライム会員は無料) 詳細を見る

This book describes the important ideas in a variety of fields such as medicine, biology, finance, and marketing in a common conceptual framework. While the approach is statistical, the emphasis is on concepts rather than mathematics. Many examples are given, with a liberal use of colour graphics. It is a valuable resource for statisticians and anyone interested in data mining in science or industry. The book's coverage is broad, from supervised learning (prediction) to unsupervised learning. The many < 続きを読む

🚩 不正確な製品情報を報告。

シェアする

¥9,218
参考価格: ¥9,239
OFF: ¥12
ポイント: 92pt (1%)
詳細はこちら

お届け日時指定便 無料
残り3点 (入荷予定あり)
Kindle版は今すぐお読みいただけます。Kindle無料アプリがあれば、さまざまなデバイスで読書が可能。在庫状況について

この商品は、Amazon.co.jp が販売、発送します。

数量: 1

カートに入れる

今すぐ買う

📍 斎藤 優太 - 152-0012 にお届け

ほしい物リストに追加する

Motivation: Offline Evaluation with Click -> Conversion data

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

example) Top-3 Recommendation in E-commerce

Ranking	<u>Recommender A</u>	<u>Recommender B</u>
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	<u>Recommender A</u>	<u>Recommender B</u>
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

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

a set of predicted rankings for user-item paris (points to \hat{Z})

user (points to $u \in \mathcal{U}$)

item (points to $i \in \mathcal{I}$)

conversion rate of u and i (points to $p_{u,i}^{cvr}$)

ranking function (weighting function) (points to $c(\hat{Z}_{u,i})$)

Ground-truth Ranking Performance

The function $c(\cdot)$ 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

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

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

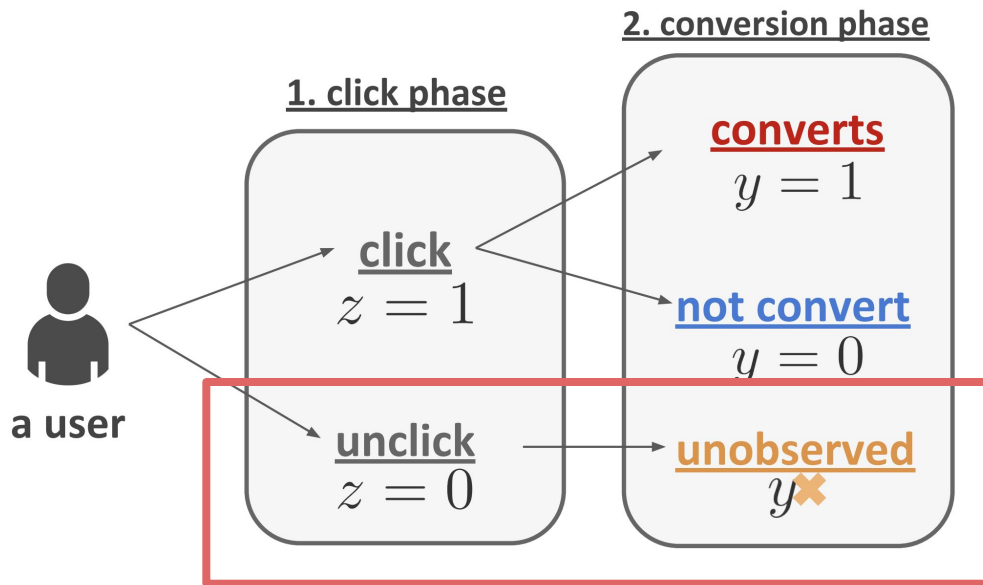
However, there are **several difficulties in evaluating recommenders in an offline environment**, including...

- 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

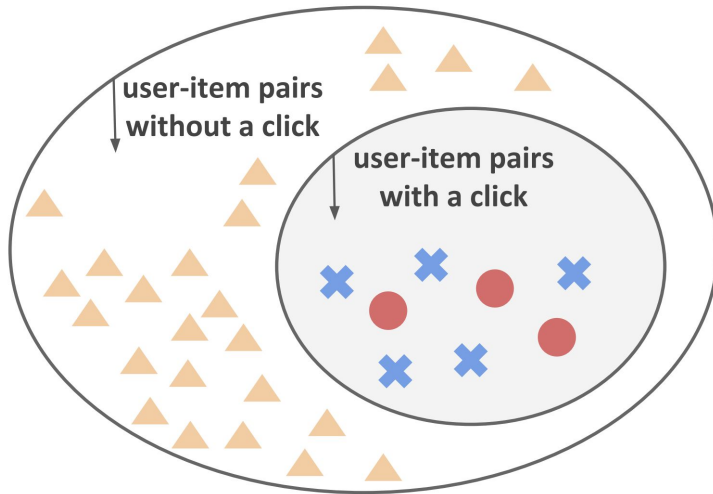


(b) user behavior pattern

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:	$\mathcal{R}_{GT}(\hat{Z}) = \frac{1}{ \mathcal{U} } \sum_{u \in \mathcal{U}} \sum_{i \in \mathcal{I}} \underline{p_{u,i}^{cvr}} \cdot c(\hat{Z}_{u,i})$
	<div style="text-align: center;">\downarrow</div>
An Estimator:	$\hat{\mathcal{R}}(\hat{Z}) = \frac{1}{ \mathcal{U} } \sum_{u \in \mathcal{U}} \sum_{i \in \mathcal{I}} \underline{???} c(\hat{Z}_{u,i})$

In summary,

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

Using offline (observable) data:

$$\{ (u, i, \underline{y_{u,i}}) \mid \underline{z_{u,i}} = 1 \}$$

conversion indicator

with a click

A Previous Solution: IPS Estimator

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

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

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

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

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

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

Our Approach: Doubly Robust Estimator

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

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

Diagram illustrating the components of the doubly robust estimator formula:

- click indicator**: Points to $z_{u,i}$.
- inverse of CTR**: Points to $p_{u,i}^{ctr}$.
- estimated CVRs**: Points to $\hat{p}_{u,i}^{cvr}$ (appearing twice).

Variance Reduction by the DR estimator

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

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

in most cases, the DR estimator has a lower variance

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

Solutions & Experiments

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

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

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

Datasets	Estimators	DCG@K			Recall@K		
		$K = 5$	$K = 10$	$K = 50$	$K = 5$	$K = 10$	$K = 50$
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