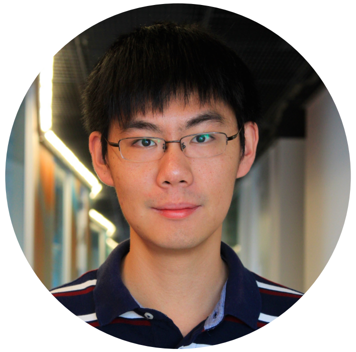
Unbiased Offline Recommender Evaluation for Missing-Not-At-Random Implicit Feedback



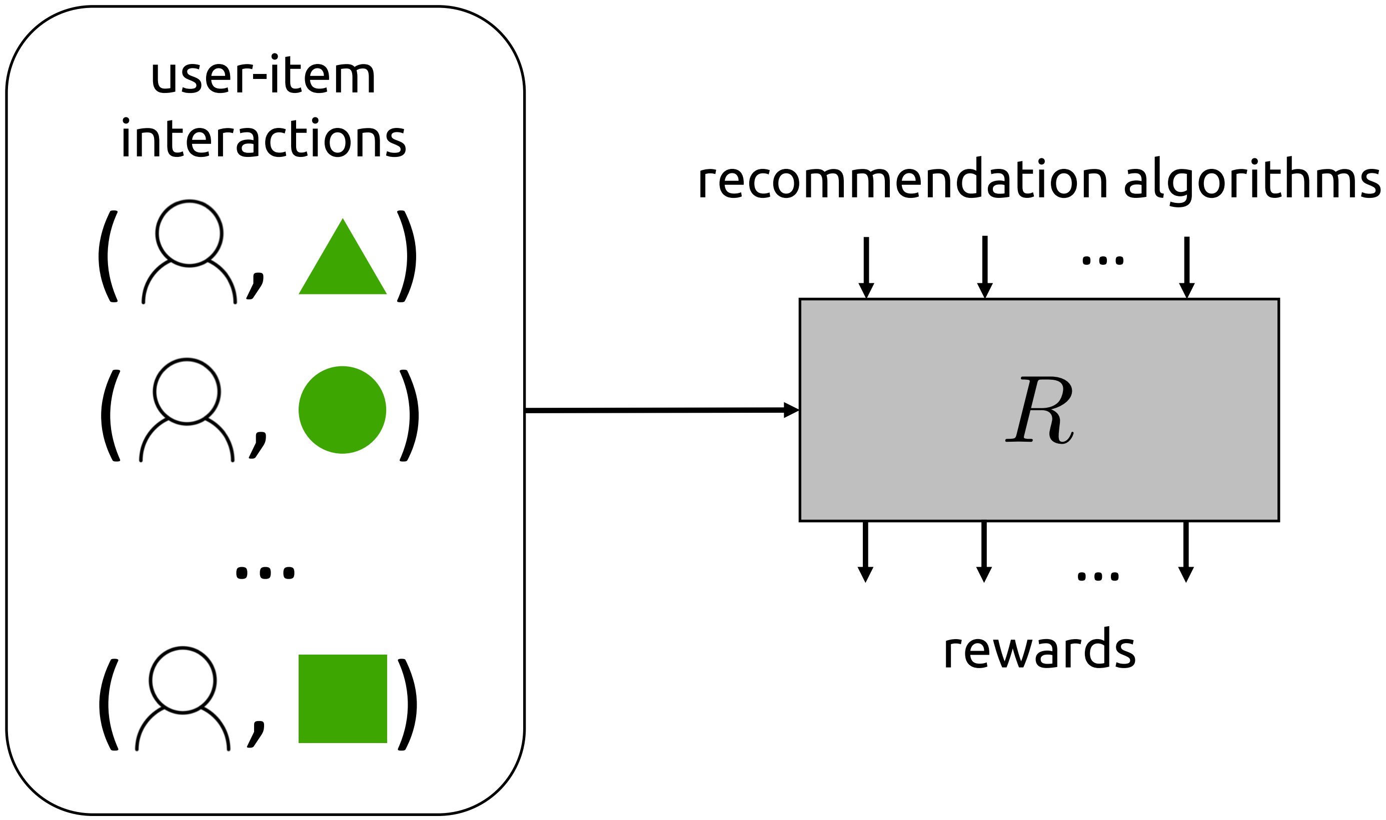
Longqi Yang   Yin Cui   Yuan Xuan   Chenyang Wang   Serge Belongie   Deborah Estrin

CORNELL TECH

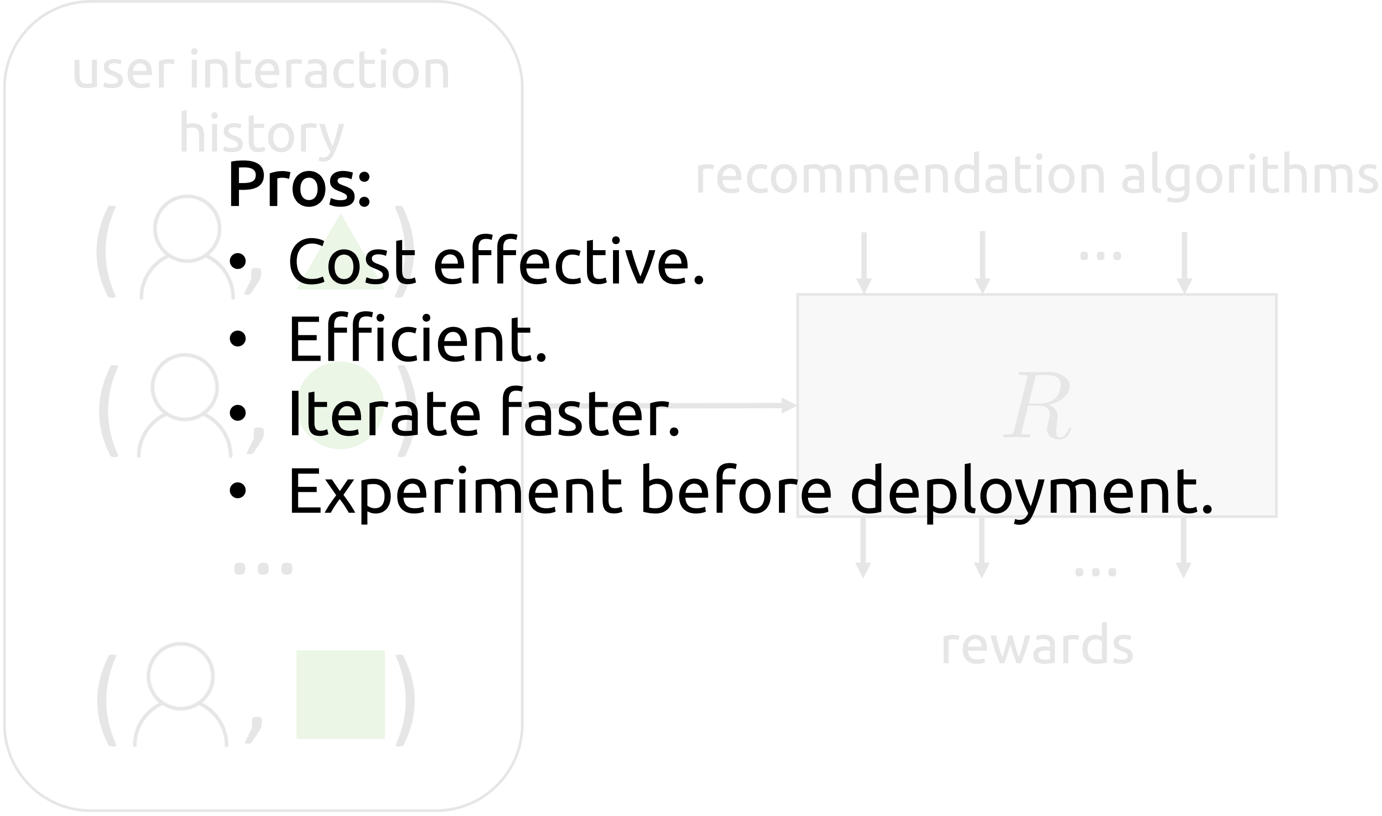
Cornell CIS Computer Science

Funders: NSF Oath: A Verizon company

Offline Evaluation of Recommendation Algorithm



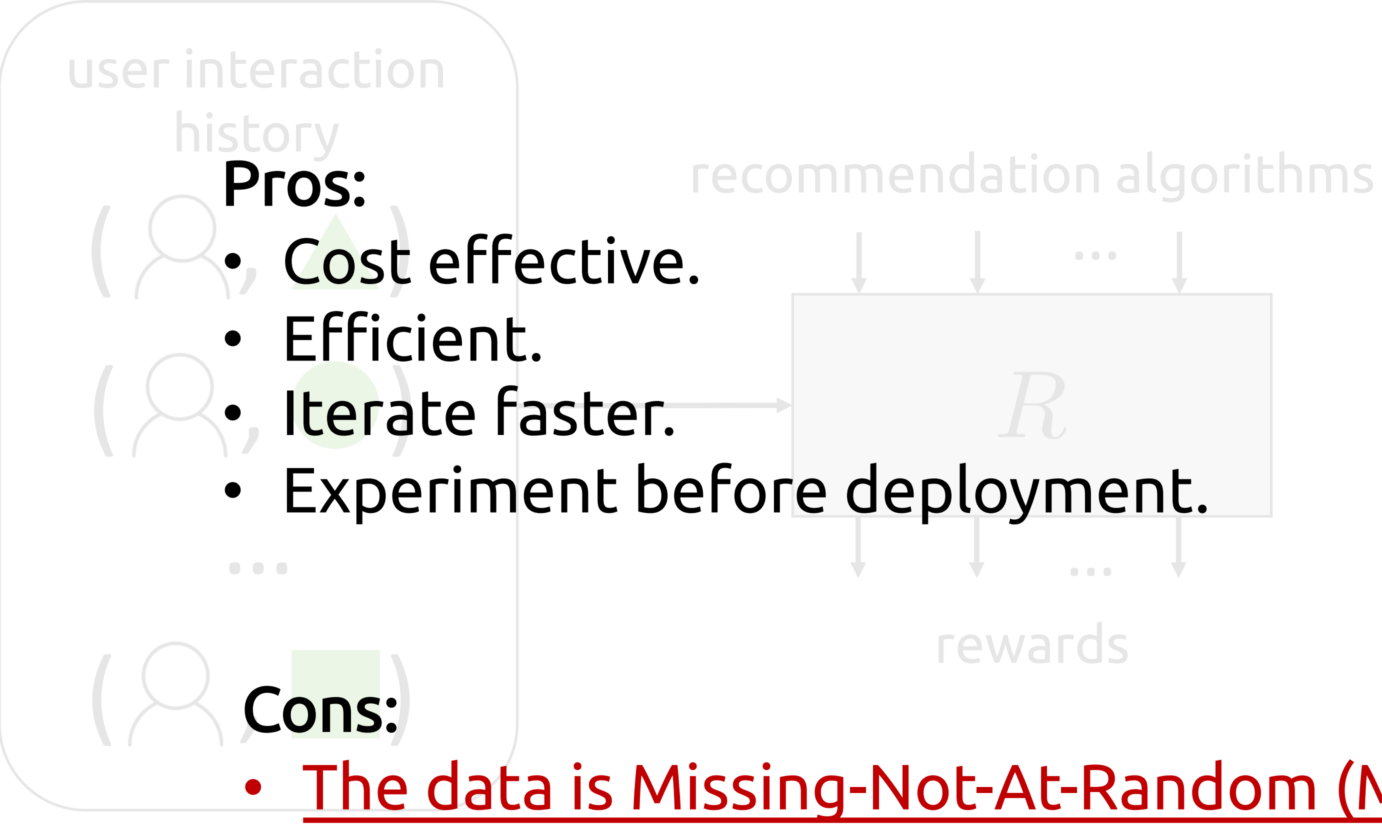
Offline Evaluation of Recommendation Algorithm



Pros:

- Cost effective.
- Efficient.
- Iterate faster.
- Experiment before deployment.

Offline Evaluation of Recommendation Algorithm



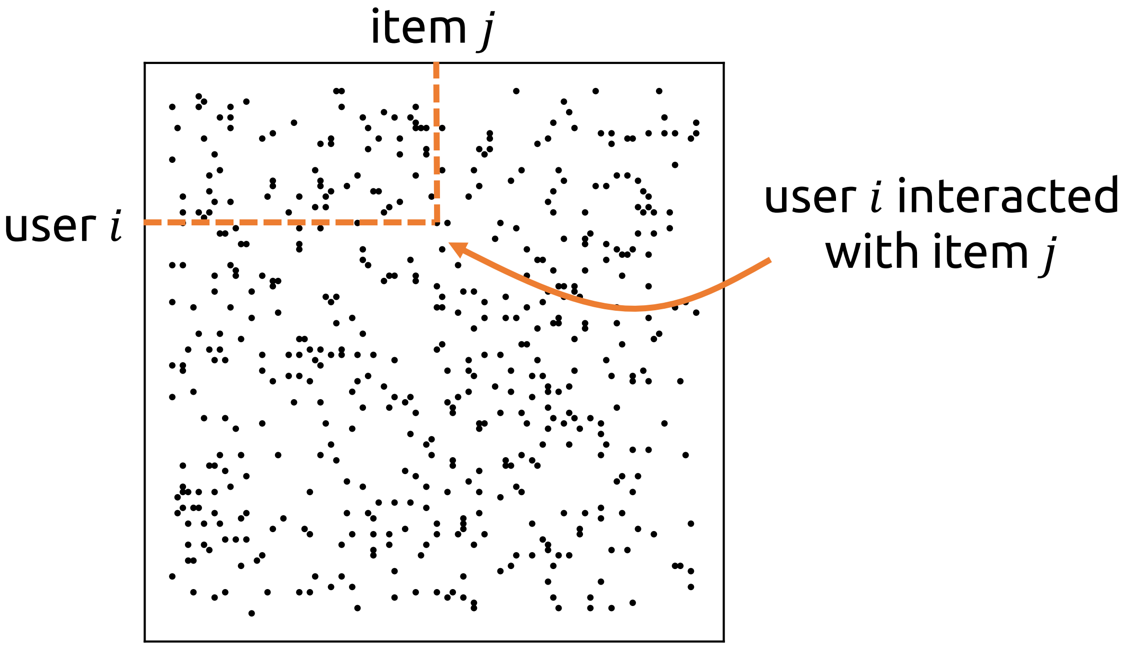
Pros:

- Cost effective.
- Efficient.
- Iterate faster.
- Experiment before deployment.

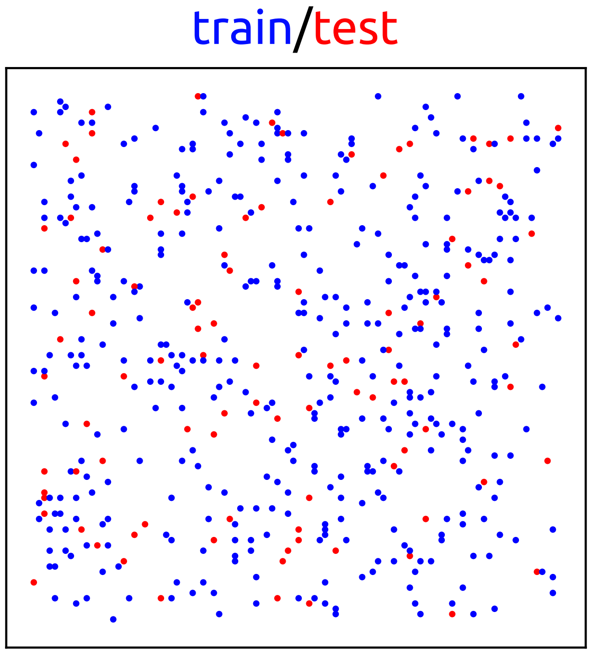
Cons:

- <u>The data is Missing-Not-At-Random (MNAR)</u>

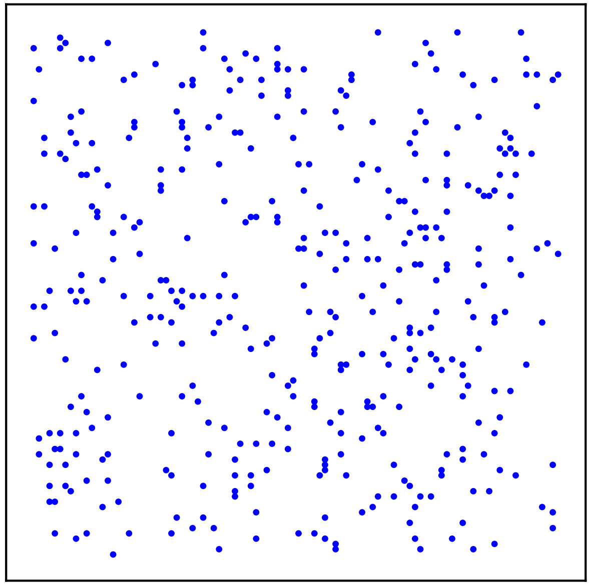
Offline Evaluation procedure



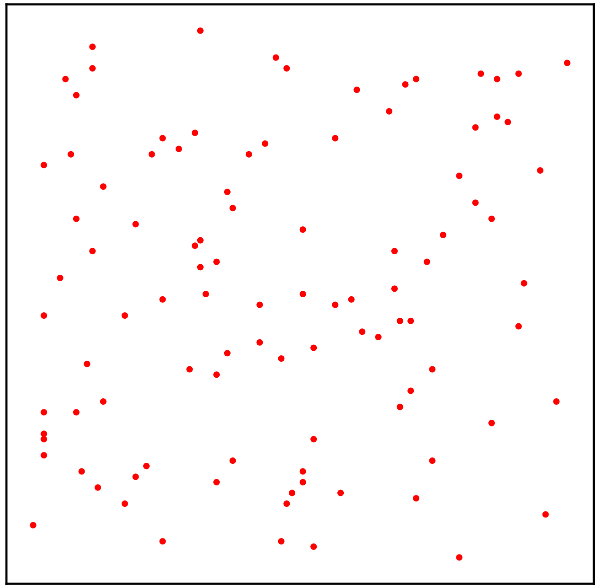
Offline Evaluation procedure



Offline Evaluation procedure



1. Train and validate a recommendation model

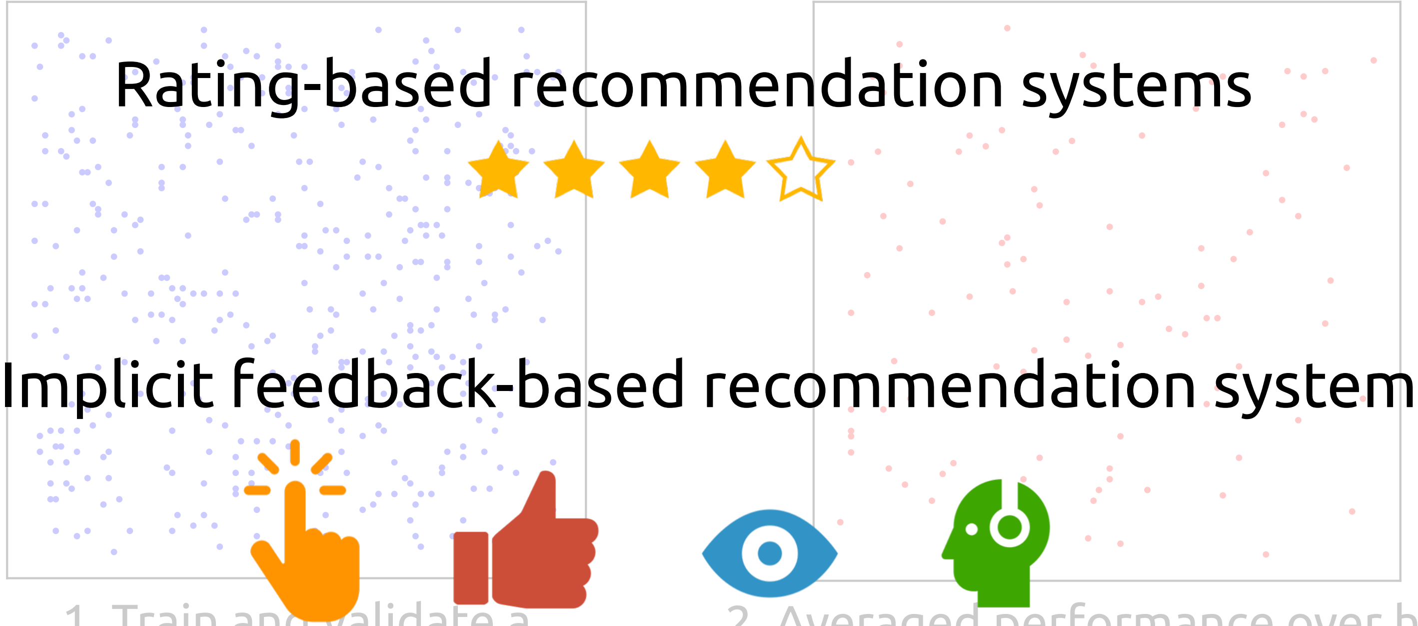


2. Averaged performance over held-out (user, item) interaction pairs (Average-Over-All)

Offline Evaluation procedure

Rating-based recommendation systems

Implicit feedback-based recommendation systems



1. Train and validate a recommendation model (policy) π

2. Averaged performance over held-out (user, item) interaction pairs (Average-Over-All)

Previous work: **Average-Over-All is biased for rating-based recommendation systems, because ratings are MNAR**

[Marlin et al. 09], [Schnabel et al. 16], [Steck 10], [Steck 11], and [Steck 13]

Previous work: **Average-Over-All is biased for rating-based recommendation systems, because ratings are MNAR**

[Marlin et al. 09], [Schnabel et al. 16], [Steck 10], [Steck 11], and [Steck 13]

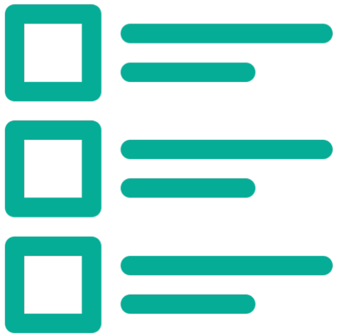
Previous work: **Average-Over-All is unbiased for implicit feedback-based recommendation systems, because implicit feedback is missing uniformly at random.**

[Lim 15]

This work: Average-Over-All is biased for implicit feedback-based recommendation systems, because implicit feedback is NOT missing uniformly at random.

This work: **Average-Over-All is biased for implicit feedback-based recommendation systems, because implicit feedback is NOT missing uniformly at random.**

trending



recommendation

Popularity bias (Users are more likely to be exposed to popular items)

A Hypothetical Example

	Popular Items		Long-tail Items
# of liked items (over all items)	1	:	10
# of liked items (over observations)	10	:	1
Algorithm 1 Performance	0.8		0
Algorithm 2 Performance	0.75		0.75

A Hypothetical Example

	Popular Items		Long-tail Items
# of liked items (over all items)	1	:	10
# of liked items (over observations)	10	:	1
Algorithm 1 Performance	0.8		0
Algorithm 2 Performance	0.75		0.75

A Hypothetical Example

	Popular Items		Long-tail Items
# of liked items (over all items)	1	:	10
# of liked items (over observations)	10	:	1
Algorithm 1 Performance	0.8		0
Algorithm 2 Performance	0.75		0.75

A Hypothetical Example

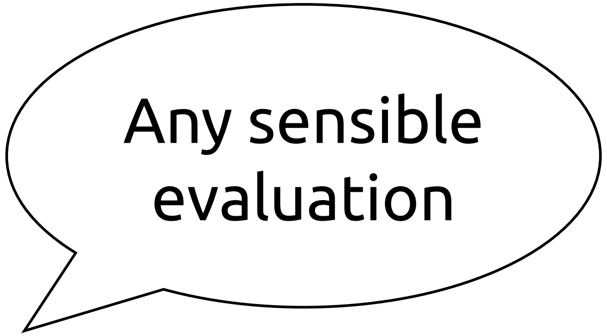
	Popular Items		Long-tail Items
# of liked items (over all items)	1	:	10
# of liked items (over observations)	10	:	1
Algorithm 1 Performance	0.8		0
Algorithm 2 Performance	0.75		0.75

A Hypothetical Example

	Popular Items		Long-tail Items
# of liked items (over all items)	1	:	10
# of liked items (over observations)	10	:	1
Algorithm 1 Performance	0.8		0
Algorithm 2 Performance	0.75		0.75

A Hypothetical Example

	Popular Items		Long-tail Items
# of liked items (over all items)	1	:	10
# of liked items (over observations)	10	:	1
Algorithm 1 Performance			0
Algorithm 2 Performance	0.75		0.75



A Hypothetical Example

	Popular Items		Long-tail Items
# of liked items (over all items)	1	:	10
# of liked item (over observati	0	:	1
Algorithm 1 Performance	0.8		0
Algorithm 2 Performance	0.75		0.75

Average-
Over-All

Formalize Reward R

Item rankings predicted by an algorithm

Ideal evaluation: $$R(\hat{Z}) = \frac{1}{|\mathcal{U}|} \sum_{u \in \mathcal{U}} \frac{1}{|\mathcal{S}_u|} \sum_{i \in \mathcal{S}_u} c(\hat{Z}_{u,i})$$

Formalize Reward R

Item rankings predicted by an algorithm

Predicted ranking of item i for user u

Ideal evaluation: $$R(\hat{Z}) = \frac{1}{|\mathcal{U}|} \sum_{u \in \mathcal{U}} \frac{1}{|\mathcal{S}_u|} \sum_{i \in \mathcal{S}_u} \boxed{c(\hat{Z}_{u,i})}$$

Items liked by user u among <u>the entire item set</u>

scoring metric

Reward for $(u,\ i)$ pair

Formalize Reward R

Item rankings predicted by an algorithm

Predicted ranking of item i for user u

Ideal evaluation: $$R(\hat{Z}) = \frac{1}{|\mathcal{U}|} \sum_{u \in \mathcal{U}} \boxed{\frac{1}{|\mathcal{S}_u|} \sum_{i \in \mathcal{S}_u} c(\hat{Z}_{u,i})}$$

Items liked by user u among the entire item set

scoring metric

Reward for user u

Formalize Reward R

Item rankings predicted by an algorithm

Predicted ranking of item i for user u

Ideal evaluation: $$R(\hat{Z}) = \boxed{\frac{1}{|\mathcal{U}|} \sum_{u \in \mathcal{U}} \frac{1}{|\mathcal{S}_u|} \sum_{i \in \mathcal{S}_u} c(\hat{Z}_{u,i})}$$

Items liked by user u among <u>the entire item set</u>

scoring metric

Reward for the algorithm

Formalize Reward R

Average-Over-All: $\hat{R}_{\mathrm{AOA}}(\hat{Z}) = \frac{1}{|\mathcal{U}|} \sum_{u \in \mathcal{U}} \frac{1}{|\mathcal{S}_u^*|} \sum_{i \in \mathcal{S}_u^*} c(\hat{Z}_{u,i})$

Items liked by user u (observed)

Formalize Bias

$$\mathbb{E}_O\left[\hat{R}_{\mathrm{AOA}}(\hat{Z})\right] \neq R(\hat{Z})$$

$$O_{u,i} = 1 \text{ if } (u,i) \text{ is observed, and } O_{u,i} = 0 \text{ otherwise}$$

$$O_{u,i} \sim \mathcal{B}(1, P_{u,i})$$

Inverse-Propensity-Scoring (IPS)

$$\hat{R}_{\mathrm{AOA}}(\hat{Z}) = \frac{1}{|\mathcal{U}|} \sum_{u \in \mathcal{U}} \boxed{\frac{1}{|\mathcal{S}_u^*|} \sum_{i \in \mathcal{S}_u^*} c(\hat{Z}_{u,i})}$$

$$\hat{R}_{\mathrm{IPS}}(\hat{Z}|P) = \frac{1}{|\mathcal{U}|} \sum_{u \in \mathcal{U}} \boxed{\frac{1}{|\mathcal{S}_u|} \sum_{i \in \mathcal{S}_u^*} \frac{c(\hat{Z}_{u,i})}{P_{u,i}}}$$

Inverse-Propensity-Scoring (IPS)

$$\hat{R}_{\mathrm{AOA}}(\hat{Z}) = \frac{1}{|\mathcal{U}|} \sum_{u \in \mathcal{U}} \frac{1}{|\mathcal{S}_u^*|} \sum_{i \in \mathcal{S}_u^*} c(\hat{Z}_{u,i})$$

$$\hat{R}_{\mathrm{IPS}}(\hat{Z}|P) = \frac{1}{|\mathcal{U}|} \sum_{u \in \mathcal{U}} \frac{1}{|\mathcal{S}_u|} \sum_{i \in \mathcal{S}_u^*} \frac{c(\hat{Z}_{u,i})}{P_{u,i}}$$

$$\boxed{\mathbb{E}_O \left[\hat{R}_{\mathrm{IPS}}(\hat{Z}|P) \right] = R(\hat{Z})}$$

Self-Normalized Inverse-Propensity-Scoring (SNIPS)

[Swaminathan et al.15]

$$\hat{R}_{\mathrm{IPS}}(\hat{Z}|P) = \frac{1}{|\mathcal{U}|} \sum_{u \in \mathcal{U}} \boxed{\frac{1}{|\mathcal{S}_u|}} \sum_{i \in \mathcal{S}_u^*} \frac{c(\hat{Z}_{u,i})}{P_{u,i}}$$

$$\hat{R}_{\mathrm{SNIPS}}(\hat{Z}|P) = \frac{1}{|\mathcal{U}|} \sum_{u \in \mathcal{U}} \boxed{\frac{1}{\sum_{i \in \mathcal{S}_u^*} \frac{1}{P_{u,i}}}} \sum_{i \in \mathcal{S}_u^*} \frac{c(\hat{Z}_{u,i})}{P_{u,i}}$$

Estimating Propensity Scores

Factor: Popularity bias (Users are more likely to be exposed to popular items)

Assumptions:

- User-independence assumption $P_{u,i} = P(O_{u,i} = 1) = P(O_{*,i} = 1) = P_{*,i}$
- Two-steps assumption $P_{*,i} = P_{*,i}^{\text{select}} \cdot P_{*,i}^{\text{interact}|\text{select}}$
- User preference is not affected by item presentation

$$P_{*,i}^{\text{interact}|\text{select}} = P_{*,i}^{\text{interact}}$$

Estimating Propensity Scores

Popularity bias model [Steck 11]:

$$\hat{P}^{\text{select}}_{*,i} \propto (n^*_i)^\gamma$$

Observed item popularity

Estimating Propensity Scores

Popularity bias model [Steck 11]:

$$\hat{P}^{\mathrm{select}}_{*,i} \propto (n^*_i)^{\gamma}$$

Estimated from known online content serving policy

$$\boxed{\hat{P}_{*,i} \propto (n^*_i)^{\left(\frac{\gamma+1}{2}\right)}}$$

Measuring bias in recommender evaluation (Yahoo! music rating dataset)

Mean Absolute Error (MAE), Recall

Model	Average-Over-All	R_{SNIPS} $(\gamma = 1.5)$	R_{SNIPS} $(\gamma = 2.0)$	R_{SNIPS} $(\gamma = 2.5)$	R_{SNIPS} $(\gamma = 3.0)$
U-CML	0.401	0.270	0.260	0.253	0.248
A-CML	0.399	0.274	0.264	0.258	0.253
BPR	0.380	0.275	0.26[illegible]	[illegible]	0.258
PMF	0.386	0.267	[illegible]	[illegible]	[illegible]8

R_{SNIPS} produces significantly lower MAE

Measuring bias in recommender evaluation (Yahoo! music rating dataset)

Mean Absolute Error (MAE), Recall

Model	Average-[illegible]	R_{SNIPS} [illegible]	R_{SNIPS} [illegible]	R_{SNIPS} [illegible]	R_{SNIPS} [illegible] = 3.0)
U-CML	[illegible]	[illegible]	[illegible]	[illegible]	0.248
A-CML	0.399	0.274	[illegible]	0.258	0.253
BPR	0.380	0.275	[illegible]	[illegible]	0.258
PMF	0.386	0.267	[illegible]	[illegible]	[illegible]

R_{SNIPS} produces significantly lower MAE

The accuracy of recommending popular items is a significant **overestimation** of the true recommendation performance

Please come to our poster or refer to our paper for:

- Proofs
- Experimental details.
- More experiments.
- Deeper analysis of the unbiased evaluator.

Conclusions and Future Work

$$\mathbb{E}_O\left[\hat{R}_{\mathrm{IPS}}(\hat{Z}|P)\right] = R(\hat{Z})$$

$$\hat{R}_{\mathrm{SNIPS}}(\hat{Z}|P) = \frac{1}{|\mathcal{U}|}\sum_{u\in\mathcal{U}} \frac{1}{\sum_{i\in\mathcal{S}_u^*}\frac{1}{P_{u,i}}} \sum_{i\in\mathcal{S}_u^*} \frac{c(\hat{Z}_{u,i})}{P_{u,i}}$$

- Understanding variance of evaluators.
- Propensity estimation (e.g., incorporate auxiliary user and item information).
- Debias training of recommendation systems (e.g., [Liang et al. 16]).

OpenRec

http://www.openrec.ai

Github link, documents, and tutorials

Longqi Yang

Ph.D. candidate

Computer Science, Cornell Tech, Cornell University

Email: ylongqi@cs.cornell.edu

Web: bit.ly/longqi

Twitter: @ylongqi

Connected Experiences Lab

http://cx.jacobs.cornell.edu/

Small Data Lab

http://smalldata.io/

CORNELL TECH

Cornell CIS Computer Science

Funders: NSF, Oath: A Verizon company