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

Year: 2018

Venue: RecSyc

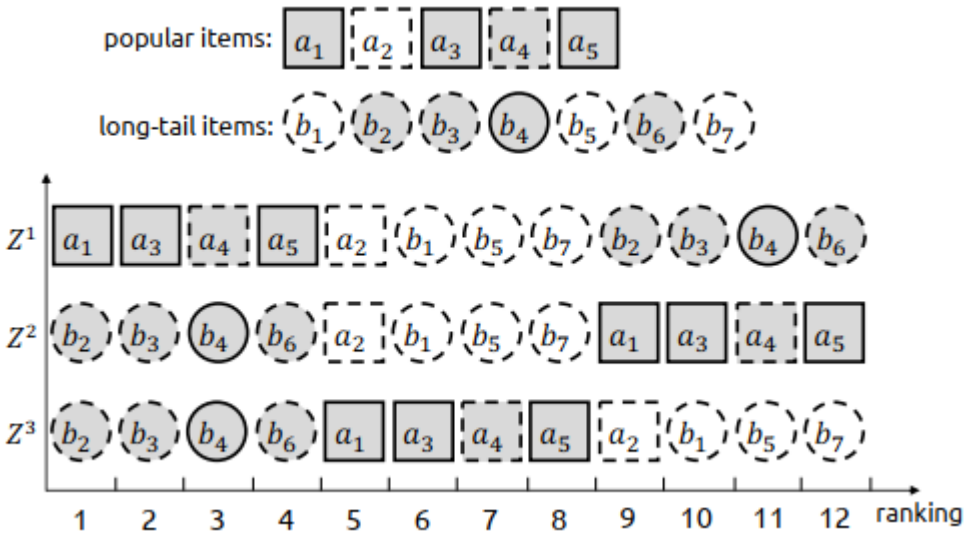
Main Problem:

For the purposes of evaluating the recommender systems, it is common to use the datasets gathered from online platforms. These online platforms are often subject to popularity bias. This basically means the more popular items will be more probable to be recommended by our system. Hence, logged ground truth data are Missing-Not-At-Random (MNAR). With that being the cause, the common method for evaluation of the recommender systems namely Average-Over-All (AOA) is biased towards trendy items. To have a clear view of the true performance of a recommender system, we need an unbiased evaluation method.

Drawbacks of Previous Works:

In contrast to explicit ratings, implicit feedback signals are one-sided and positive only. It means that an ideal recommender would never observe user interactions with irrelevant items, whereas, in explicit-feedback recommenders, complete observations assume that each user has a latent preference score for every item. As a result, for implicit-feedback recommenders, it is unclear whether a missing item in a user's history is not favoured by the user or has simply not yet been observed. Existing work simplifies the evaluation of implicit-feedback recommenders by assuming that positive signals are Missing-At-Random (MAR), that is, each favoured item is equal-likely to be clicked or viewed by a user. This assumption does not hold in real-world settings because online recommenders manifest popularity bias+. Such a bias leads to the phenomenon that relevant and trendy items are more likely to be interacted with by users.

Eventually, the Average-Over-All (AOA) evaluator implicitly places greater weight on the accuracy of serving popular items than on serving long-tail ones.



The example above shows 3 different lists of recommendations for one user. Among the shaded items that were preferred by the user, the ones with a solid border were observed by recommenders. We can observe

that the true performance of Z^1 and Z^2 is the same, but they are different in servings of popular and long-tailed items. Finally, Z^3 has the best true performance. Now if we use DCG and AOA methods for evaluation purposes, the results in the table below indicate that R_{AOA} assumed that Z^1 is the best recommender since it listed more popular items in the higher ranks. This is proof of the presence of popularity bias in the AOA method.

Estimator	Z^1	Z^2	Z^3
$R(\hat{Z})$	0.463	0.463	0.494
$\hat{R}_{\text{AOA}}(\hat{Z})$	0.585	0.340	0.390

Proposed Method:

To conduct an unbiased evaluation of biased observations, we leverage the IPS framework [16, 22] that weights each observation with the inverse of its propensity, where the term propensity refers to the tendency or the likelihood of an event happening. The intuition is to down-weight the commonly observed interactions while up-weighting the rare ones. In the context of this paper, the probability $P_{u,i}$ is treated as the pointwise propensity score. Therefore, the IPS unbiased evaluator is defined as follows:

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

$$= \frac{1}{|\mathcal{U}|}\sum_{u\in\mathcal{U}}\frac{1}{|\mathcal{S}_u|}\sum_{i\in\mathcal{S}_u}\frac{c(\hat{Z}_{u,i})}{P_{u,i}}\cdot O_{u,i}$$

The function c is any top N ranking algorithm. U refers to the whole set of users. S_u is the set of items preferred by the user from the whole available items, I. The function P refers to the propensity score and O is the binary mapping of the observation of an item.

We assume that the propensity score $P_{u,i}$ is user independent, that is, $P_{u,i} = P(O_{u,i} = 1) = P(O_{*,i} = 1) = P_{*,i}$. This simplified assumption is made to address the lack of auxiliary user information in many user-item interaction records. We derive $P_{*,i}$ by constructing a two-step generative process of user–item interactions: (1) Select, where a recommender system selects a set of items to present to a user; and (2) Interact, where the user browses the recommended items and interacts with the ones she likes. Therefore, P∗,i can be calculated as follows: $P_{*,i} = P_{*,i}^{\text{select}} \cdot P_{*,i}^{\text{interact}|\text{select}}$

where $P_{*,i}^{\text{select}}$ is the probability that item i is recommended and $P_{*,i}^{\text{interact}|\text{select}}$ is the conditional probability that the user interacts with item i was given that it is recommended

Datasets:

For each dataset, they randomly and independently hold out 15% of user–item interactions for validation and 15% for testing, and they used the remaining 70% of records for training. During testing, they excluded cold-start users and items that have no record in the training set.

- Citeulike
- Tradesy
- Amazon Book

Gaps:

- User-independent propensity. In the absence of detailed metainformation about users, they assumed that the propensity was user-independent and that the probability of an item being presented was determined by its observed popularity. The propensity may be affected by user-specific traits and preferences.
- Selection-independent interaction. They assumed that the probability that a user interacts with an item is independent of the probability that the item is recommended. This does not capture the potential impact of recommendations and item presentation orders on users' preferences.

Codebase:

https://github.com/ylongqi/unbiased-offline-recommender-evaluation