Debiased Offline Evaluation of Recommender Systems: A Weighted-Sampling Approach (Extended Abstract)

Diego Carraro
diego.carraro@insight-centre.org
Insight Centre for Data Analytics
University College Cork, Ireland

Derek Bridge
derek.bridge@insight-centre.org
Insight Centre for Data Analytics
University College Cork, Ireland

1 INTRODUCTION

Offline evaluations of recommender systems use an *observed dataset*, which records interactions (e.g. clicks, purchases, ratings) that occur during a given period in the lifetime of the recommender system.[1] However, this dataset is *biased*, not only due to the freedom that users have in choosing which items to interact with, but also due to other factors, known as confounders. The user-interface plays an important role, for example: differences in the ways items are exposed to users (e.g. position on the screen) influence the likelihood of a user interacting with those items. The recommender itself sets up a feedback loop: users are typically more likely to interact with the recommender's suggestions than with other items. Because of these and other confounders, interactions that are missing from the observed dataset are Missing Not At Random (MNAR) [4].

Classical offline evaluation uses a heldout MNAR dataset, resulting in biased estimates of performance: for example, such experiments incorrectly reward recommenders that recommend popular items or make recommendations to the more active users. Using a Missing At Random (MAR) test set would, by contrast, give an unbiased estimate of performance.[2] Datasets that are approximately MAR can be collected by randomly selecting a different subset of items for each user and then asking each user to interact with (and rate) these items [4]. This "forced approach" to obtaining ratings results in a dataset that is mostly devoid of confounders and roughly uniformly distributed over users and items.

One way that researchers try to obtain less biased offline evaluation is by designing new supposedly unbiased estimators for use on MNAR test data [6]. We investigate an alternative solution, a *sampling approach*. The general idea is to use a sampling strategy on MNAR data to generate an *intervened* test set with less bias — one that is more MAR-like. An example of this is SKEW, a sampling strategy that aims to adjust for the confounding effect that an item's popularity has on its likelihood of being observed [2].

In this extended abstract, we propose a novel formulation for the sampling approach. We compare our solution to SKEW and to a baseline, REG, which performs a random intervention on MNAR data (and hence is equivalent to no intervention in practice). We empirically validate for the first time the effectiveness of SKEW and we show our approach to be a better estimator of the performance one would obtain on *unbiased* MAR test data.

[1]This abstract emanates from research supported by a grant from Science Foundation Ireland (Grant Number SFI/12/RC/2289), which is co-funded under the European Regional Development Fund.

[2]What we refer to as MAR, in line with papers such as [1, 6], others refer to as Missing Completely At Random, MCAR, e.g. [3].

2 INTERVENED TEST SETS

We define a probabilistic framework. We consider a user-item space, $U \times I$, of size $|U| \cdot |I|$. We denote with $u \in U = \{1, .., |U|\}$ a generic user and with $i \in I = \{1, .., |I|\}$ a generic item. We also denote with the matrix $O \in \mathbb{R}^{U \times I}$ a generic observed dataset. The matrix O records which interactions between users and items have been observed: $O_{u,i} \neq 0$ if an interaction is observed and $O_{u,i} = 0$ otherwise. We also define the binary random variable $\boldsymbol{O} : U \times I \rightarrow \{0, 1\}$ over the set of user-item pairs in O as $\boldsymbol{O} = 1$ if the user-item interaction is observed and $\boldsymbol{O} = 0$ otherwise. (But later we will use abbreviation $P(\boldsymbol{O})$ in place of $P(\boldsymbol{O} = 1)$.) Using this notation, we refer to two kinds of datasets over the same $U \times I$ space, O^{mnar} and O^{mar}, which have MNAR and MAR-like properties respectively.

2.1 The goal of the sampling

The sampling approach consists in performing a debiasing intervention on MNAR data O^{mnar} by means of a given sampling strategy, denoted with S. The result of the intervention is the subset $O^S \subset O^{mnar}$, with the goal that O^S has unbiased-like properties. To model the sampling, we denote with $\boldsymbol{\mathcal{S}} : U \times I \rightarrow \{0, 1\}$ the binary random variable that guides the sampling. $\boldsymbol{\mathcal{S}} = 1$ when a particular user-item pair is sampled from O^{mnar}, 0 otherwise. (Again, we will use abbreviation $P(\boldsymbol{\mathcal{S}})$ in place of $P(\boldsymbol{\mathcal{S}} = 1)$.) In practice, a particular strategy S is characterized by the expression of the probability $P_S(\boldsymbol{\mathcal{S}}|u, i), \forall (u, i) \in O^{mnar}$, which is the probability distribution responsible for guiding the sampling on O^{mnar}.

In REG, because every (u, i) entry has the same probability of being sampled, we have $P_S(\boldsymbol{\mathcal{S}}|u, i) = 1/|O^{mnar}|$. In SKEW, we have $P_S(\boldsymbol{\mathcal{S}}|u, i) = 1/\text{pop}(i)$, where $\text{pop}(i)$ counts the number of interactions available for the item i in O^{mnar}.

2.2 Our approach: weights for the sampling

Our approach assumes the availability of some unbiased MAR-like data O^{mar} in addition to MNAR data O^{mnar}.

Our idea is to make the posterior probability of each user-item pair in the sampled O^S, i.e. $P_S(u, i|\boldsymbol{\mathcal{S}})$, approximately the same as the posterior distribution observed for the corresponding user-item pair in O^{mar}, i.e. $P_{mar}(u, i|\boldsymbol{O})$. Writing this as a formula, we want:

$$P_S(u, i|\boldsymbol{\mathcal{S}}) \approx P_{mar}(u, i|\boldsymbol{O}) \quad \forall (u, i) \in O^S \tag{1}$$

To obtain this approximation, we adjust the posterior distributions of the sampling space O^{mnar}, i.e. $P_{mnar}(u, i|\boldsymbol{O})$, using user-item weights $w = (w_{ui})_{u \in U, i \in I}$ (similarly to [5]). We denote the modified weighted MNAR posteriors by $P_{mnar}(u, i|\boldsymbol{O}, w)$ and we use w to obtain the following equality:

Table 1: Percentage difference of performances on the intervened test sets with respect to O^{gt}.

Recommender	REG	SKEW	WTD
PosPop	+274.5	**-13.7**	-68.4
AvgRating	-75.6	**-47.8**	-50.5
UB_KNN	+272.5	+42.7	**-4.4**
IB_KNN	+312.1	+46.5	**-3.78**
RankMF	+259.9	+86.0	**+54.8**

$$P_{mnar}(u,i|O,w) = P_{mar}(u,i|O) \quad \forall (u,i) \in O^{mnar} \tag{2}$$

We know that a typical MAR dataset is uniformly distributed over users and items, and so we have that $P_{mar}(u|O) \approx 1/|U|$ and $P_{mar}(i|O) \approx 1/|I|$. Therefore, $P_{mar}(u,i|O) \approx 1/|U||I|$, and we can infer $P_{mar}(u,i|O) \approx P_{mar}(u|O)P_{mar}(i|O)$. We use the independence of user and item posteriors to split the left side of formula 2 into the two following equalities:

$$P_{mnar}(u|O,w) = P_{mar}(u|O) \quad \forall u \in U \tag{3}$$

$$P_{mnar}(i|O,w) = P_{mar}(i|O) \quad \forall i \in I \tag{4}$$

As a consequence of the independence between user and item posteriors which led to formulas 3 and 4, we can define and calculate user-specific weights $w = (w_u)_{u \in U}$ and item-specific weights $w = (w_i)_{i \in I}$ instead of weights that are user-item specific.

We propose the most straightforward solution to model the weighted MNAR posteriors, i.e. $P_{mnar}(.|O,w) = w_. P_{mnar}(.|O)$. We plug this into formulas 3 and 4 and we obtain $w_u P_{mnar}(u|O) = P_{mar}(u|O)$, $w_i P_{mnar}(i|O) = P_{mar}(i|O)$ for each user and item weighted distribution respectively. Simply reversing the two formulas, we have the expressions for calculating the weights:

$$w_u = \frac{P_{mar}(u|O)}{P_{mnar}(u|O)} \quad \forall u \in U \tag{5}$$

$$w_i = \frac{P_{mar}(i|O)}{P_{mnar}(i|O)} \quad \forall i \in I \tag{6}$$

We can think of the calculated weights as quantities that measure the divergence between the MNAR distributions of the sampling space and the target MAR distribution. Because a specific weight adjusts the corresponding MNAR distribution, we directly use weights to model the sampling distribution, i.e. $P_S(\mathcal{S}|u,i) = w_u w_i$. During the sampling, the effect of the weights is to increase or decrease the probability that a particular user-item pair is sampled depending on how divergent are the user and item posterior probabilities in the MNAR sampling space with respect to MAR distributions.

In fact, based on preliminary experiments, we use $P_S(\mathcal{S}|u,i) = w_u (w_i)^2$ instead. This variant, denoted by WTD, raises the importance of the item weight relative to the user weight. Specifically, $(w_i)^2$ will be bigger than w_i if w_i is greater than one, and $(w_i)^2$ will be smaller than w_i if w_i is less than one.

3 EXPERIMENTS

In this extended abstract, we illustrate our proposed approach using just one dataset, the Webscope R3 dataset, a 5-stars rating dataset on the music domain [4].[3] This dataset is composed of two parts: one part was collected by the "forced approach" we mentioned earlier and thus has MAR-like properties (O^{mar}); the other part is a traditional observed dataset, having MNAR properties (O^{mnar}).

We randomly split O^{mnar} into a training set O^{tr} and a heldout set O^{he}. Since the split is random, MNAR distributions are preserved in both sets. We use the former to train a given recommender. We use the latter as the sampling space to obtain the three intervened test sets, one per strategy (REG, SKEW, WTD).

We split O^{mar} into two, i.e. O^w and O^{gt}. O^w is used to calculate the weights; see the next paragraph. We use O^{gt} as an unbiased test set (or, at least, as a test set that is much less biased than an MNAR test set would be). In other words, we want the performance of a recommender on an intervened test set to be close to its performance on this unbiased test set. The best intervention strategy is the one that produces test sets where performance most closely resembles performance on O^{gt}.

To calculate w_u for a specific u (Eq. 5), we need the posteriors $P_.(u|O)$. These are calculated as the proportion of interactions associated with user u in O^w over the proportions in O^{tr}. The item weights (Eq. 6) are calculated analogously.

We train five recommender systems on O^{tr}. AvgRating and PosPop rank items in descending order of their mean rating and number of positive ratings, respectively. UB_KNN and IB_KNN are user-based and item-based nearest-neighbour algorithms. RankMF is Matrix Factorization with a ranking loss function.

Table 1 shows how Recall@10 on the three intervened sets differs from Recall@10 on O^{gt}. We see that WTD is superior for UB_KNN, IB_KNNN and RankMF. It is close to SKEW in the case of AvgRating. SKEW is expected to be the best for PosPop as it is designed to account for the item popularity bias in a test set.

We conclude that WTD is the most robust across various recommenders. We are investigating using it to debias *training sets* to complement this work on debiasing *test sets*. We also have a way of calculating the sampling weights without requiring the unbiased dataset, O^w, using instead a hypothesized, 'ideal' MAR distribution.

REFERENCES

[1] Rocío Cañamares and Pablo Castells. 2018. Should I Follow the Crowd?: A Probabilistic Analysis of the Effectiveness of Popularity in Recommender Systems. In *41st International ACM SIGIR Conference on Research & Development in Information Retrieval*. 415–424.

[2] Dawen Liang, Laurent Charlin, and David M. Blei. 2016. Causal Inference for Recommendation. In *UAI Workshop on Causation*.

[3] Benjamin M. Marlin and Richard S. Zemel. 2009. Collaborative Prediction and Ranking with Non-random Missing Data. In *Third ACM Conference on Recommender Systems*. 5–12.

[4] Benjamin M. Marlin, Richard S. Zemel, Sam Roweis, and Malcolm Slaney. 2007. Collaborative Filtering and the Missing at Random Assumption. In *Twenty-Third Conference on Uncertainty in Artificial Intelligence*. 267–275.

[5] Arnaud De Myttenaere, Bénédicte Le Grand, Boris Golden, and Fabrice Rossi. 2014. Reducing Offline Evaluation Bias in Recommendation Systems. *CoRR* abs/1407.0822 (2014).

[6] Harald Steck. 2010. Training and Testing of Recommender Systems on Data Missing Not at Random. In *16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 713–722.

[3] Available by request at https://webscope.sandbox.yahoo.com