Holistic AI × The Alan Turing Institute

Bias in Recommender Systems Part IV

Content by: Sachin Beepath, Giulio Filippi, Cristian Munoz, Roseline Polle, Nigel Kingsman, Sara Zannone

Speaker: Giulio Filippi

Contents

IV – Mitigating Bias in Recommender Systems

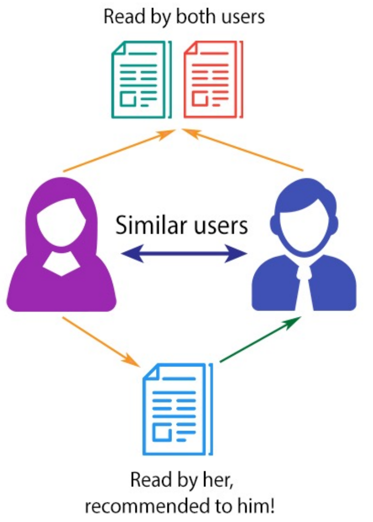
- 1) Get a feel for the taxonomy around bias mitigation.
- 2) Introduce preliminaries: collaborative filtering and matrix factorization.
- 3) Introduce a bias mitigation procedure for user fairness and one for item fairness.

Taxonomy

- Mitigation of bias can be split into three broad categories: Pre-processing, In-processing, Post-processing.
- Pre-processing: the changes are made to the dataset, before the model is trained
- In-processing: the bias mitigation is included in the way we devise and train our model
- Post-processing: the data and model are left unchanged, but we alter the outputs of the model

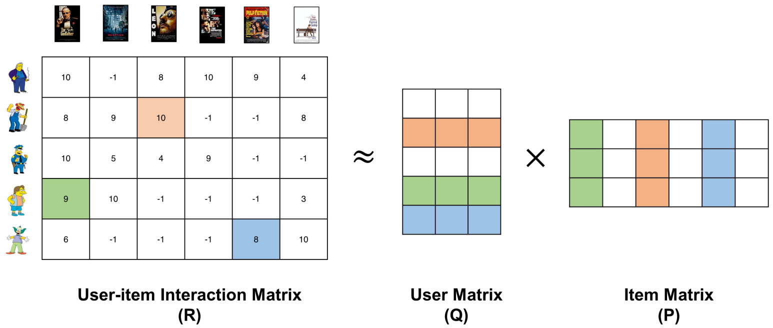
Collaborative filtering

- Collaborative filtering is one approach to recommendation.
- The assumption is that similar users will like similar items.



Traditional Matrix Factorization

– Matrix factorization is a class of collaborative filtering algorithms used in recommender systems. Matrix factorization algorithms work by decomposing the user-item interaction matrix into the product of two lower dimensionality rectangular matrices.



Traditional Matrix Factorization

- $R = QP$
- Otherwise written as
- $R_{ui} = \sum_k Q_{uk} P_{ki}$
- Where Q is of shape (num_users, K) and P is of shape (K, num_items).
- The idea is that the rectangular matrices contain K latent features describing the users and items, and the dot product of the latent features is a good approximation to the rating.

Traditional Matrix Factorization

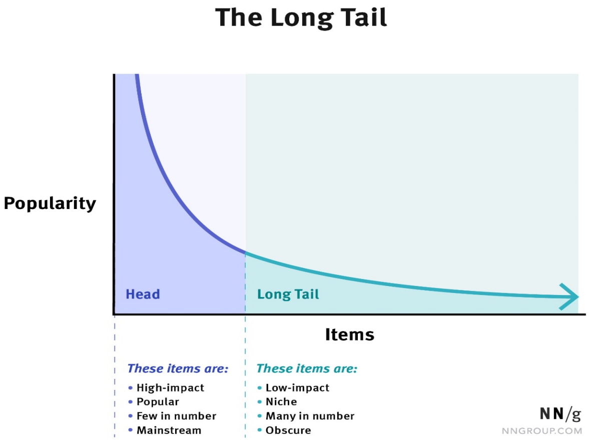
- In practice we find the matrices Q and P by solving an optimization problem.
- $Argmin_{Q,P} ||R - QP||^2 + \lambda(||Q||^2 + ||P||^2)$
- $||R - QP||^2$ is the efficacy term.
- $\lambda(||Q||^2 + ||P||^2)$ is a regularization term.
- There will be some reconstruction error after training. That error is often used as an efficacy metric.

Example 1 – Item Fairness

- Individual Item Bias Mitigation
- In-processing method from (Sun et al, 2019)
- The referenced paper contains 4 different mitigation strategies for interested students

Popularity Bias

- Popularity Bias is a common type of bias found in recommendation.
- Few popular items tend to dominate the recommendations (they are recommended much more than others)
- We usually see this as a fast decrease in the long tail plot.



Blind Spot Aware Matrix Factorization

- There is a family of mitigation techniques (Sun et al, 2019) that change the objective of matrix factorization to account for popularity bias.
- $Argmin_{Q,P}[\sum_{O_{u,i}\neq 0}\left\|R_{u,i}-(QP)_{u,i}\right\|^2+\lambda(\|Q_u\|^2+\|P_i\|^2)+\beta(\|Q_u-P_i\|^2)]$
- Where $O_{u,i}=1$ where we have a rating from user u to item i.
- $\left\|R_{u,i}-(QP)_{u,i}\right\|^2$ is the accuracy term.
- $\lambda(\|Q_u\|^2+\|P_i\|^2)$ is a regularizer.
- The $\beta(\|Q_u-P_i\|^2)$ term is a fairness regularizer, explained in next slide.

Blind Spot Aware Matrix Factorization

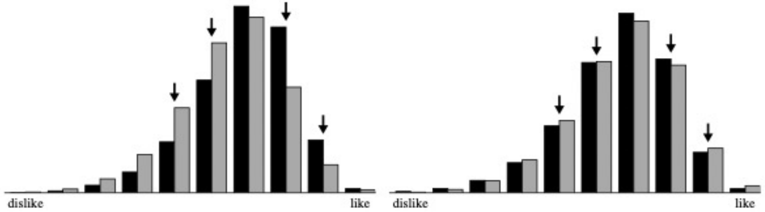
- The $\beta(||Q_u - P_i||^2)$ term is a fairness regularizer
- making all the latent vectors more homogeneous so giving more items a chance at being picked.
- If beta is very large, then we are forcing all the latent vectors to be the same, so the scores will be the same for all items.

Example 2 – User Fairness

- User Group Bias Mitigation
- In-processing method from (Kamishima et al, 2018).
- A variety of other methods can be found in (Boratto et al, 2022).

Exposure Bias

– We wish to have independence of outputs of model and chosen sensitive attribute.



(a) standard (b) independence-enhanced

Figure 1: Distributions of the predicted ratings for each sensitive value

Mitigation Strategy

– We have seen matrix factorization in previous part, this paper also uses a variant on matrix factorization.

– The loss function is composed of three parts: the first is an accuracy term, the second is a regularizer, the third is an independence term.

– In training our model (minimizing the loss), adding an independence term ensures that the outputs of the model are approximately independent from the sensitive attribute.

– $Loss = \sum_{O_{u,i} \neq 0} \left\| R_{u,i} - \widehat{R_{u,i}} \right\|^2 + \lambda RegTerm + \eta IndTerm$

Independence Term

- The paper proposes several independence terms, we will introduce one of them here.
- This independence term uses the fact that under the assumption of independence between ratings and the sensitive attribute, the expected scores should be equal for both groups.
- In equations $IndTerm = (\mathbb{E}[R|S=0] - \mathbb{E}[R|S=1])^2$
- Where $\mathbb{E}$ denotes the expectation and S is the sensitive attribute.
- In practice we compute an empirical expectation.

Mitigating bias with holisticai library

– The first step is installing the library

– The documentation for the mitigation strategies can be found here.

Mitigating bias with holisticai library

– The training and bias mitigation both happen in fitting the object.

```
# import model
from holisticai.bias.mitigation import BlindSpotAwareMF

# instantiate and train model
mf = BlindSpotAwareMF(K=40, beta=0.02, steps=10, alpha=0.002, lamda=0.008, verbose=1)
mf.fit(data_matrix)

# predictions
print(mf.pred)
```

– The matrix of predictions can be accessed with mf.pred.

Exercise Notebooks

– We have created exercise Notebooks for mitigating bias.

References

[1] Dash et al, 2021, When the Umpire is also a Player

[2] Deldjoo et al, 2022, A Survey of Research on Fair Recommender Systems

[3] Sun et al, 2019, Debiasing the Human-Recommender System Feedback Loop in Collaborative Filtering

[4] V. Tsintzou, E. Pitoura, P. Tsaparas, 2019, Bias disparity in recommendation systems

[5] M. Mansoury, B. Mobasher, R. Burke, M. Pechenizkiy, 2019, Bias disparity in collaborative recommendation

[6] *Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, Jun Sakuma, 2018,* Recommendation Independence

Sources

https://fairness-tutorial.github.io/files/Tutorial_on_Fairness_in_Recommendation_Slides.pdf

https://ir.library.louisville.edu/cgi/viewcontent.cgi?article=1440&context=faculty

https://arxiv.org/pdf/2205.11127.pdf

https://arxiv.org/pdf/2005.01148.pdf

https://arxiv.org/pdf/2105.05779.pdf