Beyond Parity: New Fairness Objectives for Collaborative Filtering

Sirui Yao
Dept. of Computer Science
Virginia Tech

Bert Huang
Dept. of Computer Science
Virginia Tech

Introduction

We study how discrimination that exists in historical data affects fairness in collaborative-filtering recommender systems. We identify the insufficiency of existing fairness metrics and propose four new metrics that address different forms of unfairness. These fairness metrics can be optimized by adding fairness terms to the learning objective.

Matrix Factorization

The basic matrix-factorization formulation builds on the assumption that each rating can be represented as the product of vectors representing the user and item.

$$r_{ij} \approx \boldsymbol{p}_i^\top \boldsymbol{q}_j + u_i + v_j$$

The learning algorithm seeks to learn these parameters from observed ratings, typically by minimizing a regularized, squared reconstruction error.

$$J(\boldsymbol{P}, \boldsymbol{Q}, \boldsymbol{u}, \boldsymbol{v}) = \frac{\lambda}{2}\left(||\boldsymbol{P}||_{\mathrm{F}}^2 + ||\boldsymbol{Q}||_{\mathrm{F}}^2\right) + \frac{1}{|X|}\sum_{(i,j)\in X}\left(y_{ij} - r_{ij}\right)^2$$

New Fairness Metrics

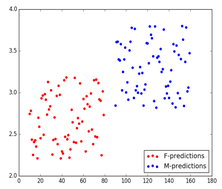


figure 1

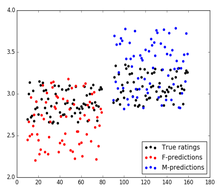


figure 2

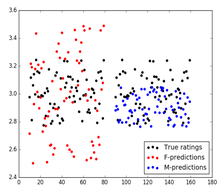


figure 3

"Demographical Parity" evaluates the absolute difference between the overall average ratings of disadvantaged users and those of advantaged users (*figure 1*), **which does not consider the real ratings and disregard the difference in preference between user groups.**

$$U_{\mathrm{par}} = \left|\mathrm{E}_g\left[y\right] - \mathrm{E}_{\neg g}\left[y\right]\right|$$

We instead **compare the predicted ratings against the real ratings and measure the difference in error across groups.** We address two main types of differences: difference in average error (*figure 2*) and difference in prediction quality (*figure 3*).

With binary groups, the formula of our fairness metrics is

$$U = \frac{1}{n}\sum_{j=1}^{n}\left|\mathrm{error}\left(g\right) - \mathrm{error}\left(\neg g\right)\right|$$

We consider various forms of error, each captures a different type of unfairness that may have different consequences.

- Value

$$\mathrm{error}\left(g\right) = \left(\mathrm{E}_g\left[y\right]_j - \mathrm{E}_g\left[r\right]_j\right)$$

$$\mathrm{E}_g\left[y\right]_j := \frac{1}{\left|\{i : ((i,j) \in X) \wedge g_i\}\right|}\sum_{i:((i,j)\in X)\wedge g_i} y_{ij}$$

Value unfairness computes on signed error, it occurs when one class of user is consistently given higher or lower predictions than their true preferences.

- Absolute

$$\mathrm{error}\left(g\right) = \left|\mathrm{E}_g\left[y\right]_j - \mathrm{E}_g\left[r\right]_j\right|$$

Absolute unfairness considers unsigned error, it captures a single statistic representing the quality of prediction for each user type

- Underestimation

$$\mathrm{error}\left(g\right) = \max\{0,\ \mathrm{E}_g\left[y\right]_j - \mathrm{E}_g\left[r\right]_j\}$$

Underestimation unfairness is important in settings where missing recommendations are more critical than extra recommendations.

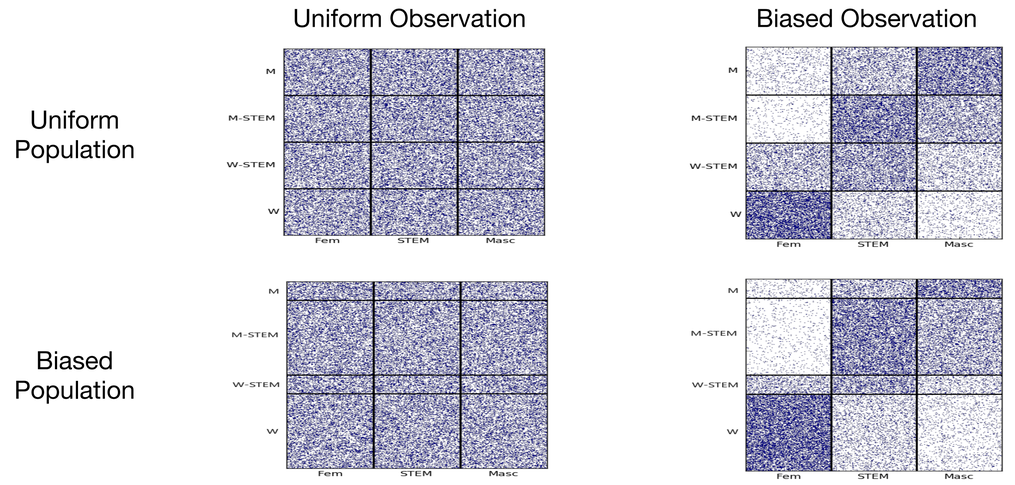
- Overestimation

$$\mathrm{error}\left(g\right) = \max\{0,\ \mathrm{E}_g\left[r\right]_j - \mathrm{E}_g\left[y\right]_j\}$$

Overestimation unfairness may be important in settings where users may be overwhelmed by recommendations, so providing too many recommendations would be especially detrimental.

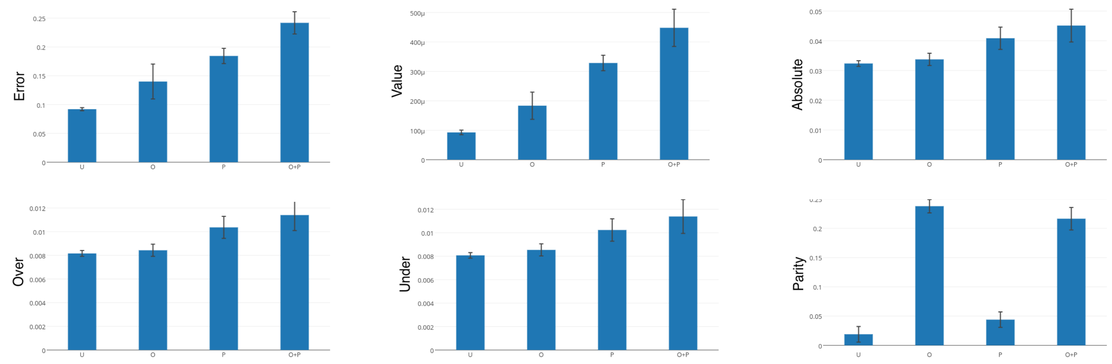
Observation & Population Bias

In our synthetic experiments, we generate simulated course-recommendation data from a block model. We consider four user groups: W (Women), W-STEM, M-STEM, and M (Men), three item groups: Fem, STEM and Masc, we compare four sampling conditions, each with uniform or unbalanced observation probability and user group distribution.



*Block size indicates population size, color thickness represents observation density

For each sampling condition, we generate 400 users and 300 items and sample preferences and observations. Training on these ratings, we then measure the various unfairness metrics as well as error by comparing the predicted rating against the true expected rating.



Learning

To reduce unfairness, we augment the learning objective by adding a smoothed variation of a fairness metric based on the Huber loss

$$\min_{\boldsymbol{P},\boldsymbol{Q},\boldsymbol{u},\boldsymbol{v}} \ J(\boldsymbol{P}, \boldsymbol{Q}, \boldsymbol{u}, \boldsymbol{v}) + U$$

We ran experiments on the Movielens Million Dataset. We manually selected five genres (action, crime, musical, romance, and sci-fi) that each have different forms of gender imbalance and only consider movies that list these genres.

Unfairness	Error	Value	Absolute	Underestimation	Overestimation	Non-Parity
None	0.887 ± 1.9e-03	0.234 ± 6.3e-03	0.126 ± 1.7e-03	0.107 ± 1.6e-03	0.153 ± 3.9e-03	0.036 ± 1.3e-03
Value	0.886 ± 2.2e-03	**0.223 ± 6.9e-03**	0.128 ± 2.2e-03	**0.102 ± 1.9e-03**	**0.148 ± 4.9e-03**	0.041 ± 1.6e-03
Absolute	0.887 ± 2.0e-03	0.235 ± 6.2e-03	**0.124 ± 1.7e-03**	0.110 ± 1.8e-03	0.151 ± 4.2e-03	0.023 ± 2.7e-03
Under	0.888 ± 2.2e-03	0.233 ± 6.8e-03	0.128 ± 1.8e-03	**0.102 ± 1.7e-03**	0.156 ± 4.2e-03	0.058 ± 9.3e-04
Over	**0.885 ± 1.9e-03**	0.234 ± 5.8e-03	**0.125 ± 1.6e-03**	0.112 ± 1.9e-03	**0.148 ± 4.1e-03**	0.015 ± 2.0e-03
Non-Parity	0.887 ± 1.9e-03	0.236 ± 6.0e-03	0.126 ± 1.6e-03	0.110 ± 1.7e-03	0.152 ± 3.9e-03	**0.010 ± 1.5e-03**

References

1. Hardt, Moritz, Eric Price, and Nati Srebro. "Equality of opportunity in supervised learning." *Advances in Neural Information Processing Systems*. 2016.
2. Kamishima, Toshihiro, et al. "Model-based approaches for independence-enhanced recommendation." ICDMW. 2016.
3. Koren, Yehuda, Robert Bell, and Chris Volinsky. "Matrix factorization techniques for recommender systems." *Computer* 42.8 (2009).