論文読み会：*“Countering Popularity Bias by Regularizing Score Differences”*, Recsys 2022

宮本 隆志

ナビプラス株式会社

2023/07/19

紹介する論文

popularity bias

提案手法

既存手法との比較

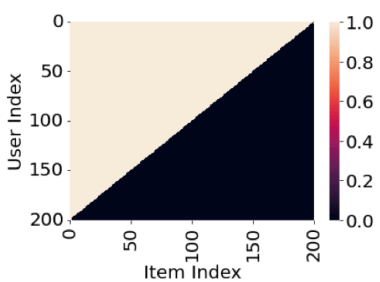
よりリアルなデータで評価

まとめ

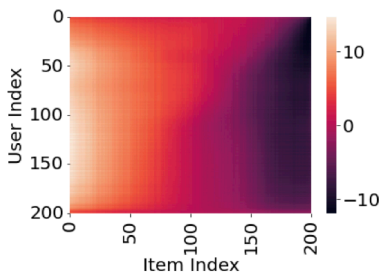
紹介する論文

"Countering Popularity Bias by Regularizing Score Differences" (Recsys'22)

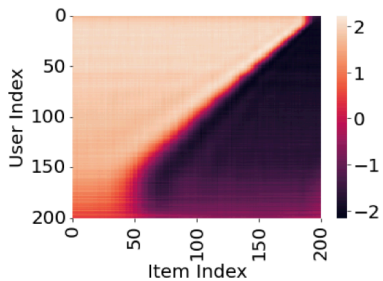
- ソウル大の論文
- popularity bias の補正方法の話
- 人気の高いアイテムの推薦スコアが不当に高くなる問題
- 損失関数に補正のための正則化項を入れて改善した
- コードは `https://github.com/stillpsy/popbias`



(a) Synthetic Data



(a) Score Prediction (Baseline)



(c) Score Prediction (Zerosum)

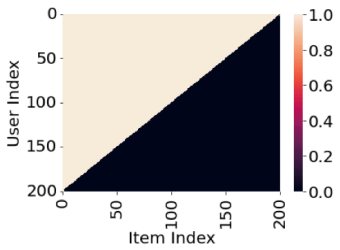
popularity bias

- popularity bias は２種類ある。今回は後者に focus する。
  - data bias 表示されたアイテムしか行動履歴が貯まらないという positive feedback 問題
  - model bias そもそも推薦計算モデル自体が imbalanced data だと bias を生じる問題。人気のあるアイテムに不当に高い推薦スコアを出す。
- baseline = Bayesian Personalized Ranking with Matrix Factorization
- 評価指標

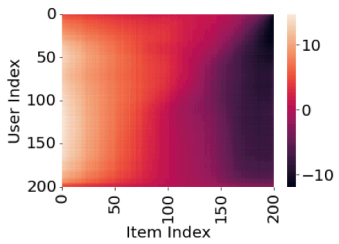
$$\text{PRI} = \text{Spearman 順位相関係数 } (\text{Pop}(I), \text{avg\_rank}(I))$$

$$\text{PopQ@1} = \frac{1}{|U|} \sum_{u \in U} \text{PopQuantile}_u(\text{argmax}_{i \in \text{Pos}_u} \hat{y}_{u,i})$$

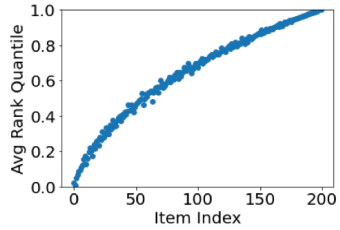
- 合成データ $y_{u,i} = 1(\text{if } u + i \leq 200) \text{ or } = 0$
  - Accuracy $\geq$ 99,99% と学習自体は高精度、
  - $y_{u,i} = 1$ と等しいのに、popular なアイテムのスコアが高くて上位に推薦される。



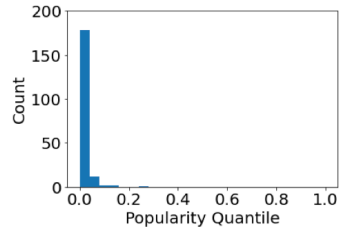
(a) Synthetic Data



(b) Model Score Prediction



(c) Average Rank Quantile of the Items

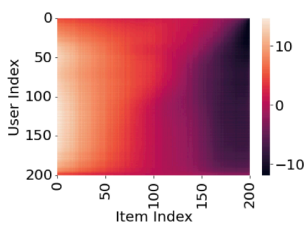


(d) Popularity Quantiles of the Top Positive Items

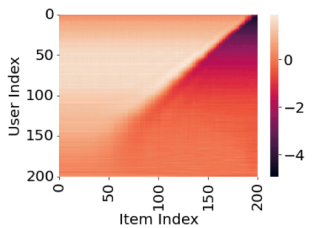
提案手法

Pos2Neg2 Positive item 2 つ、Negative item 2 つを選んで、その score が等しくなる正規化項を追加する。Accuracy Error が悪化。

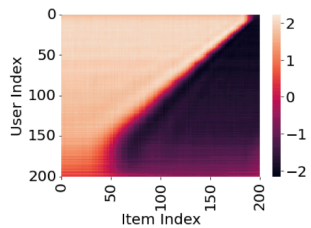
Zerosum Positive item 1 つ、Negative item 1 つを選んで、その score の絶対値が等しくなる正規化項を追加する。Accuracy Error の悪化は無し。



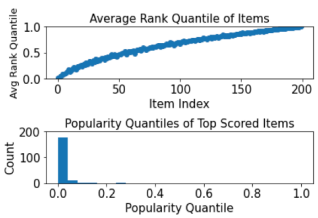
(a) Score Prediction (Baseline)



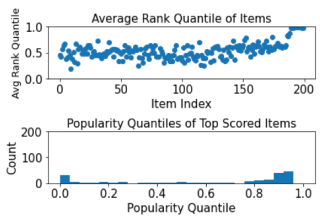
(b) Score Prediction (Pos2Neg2)



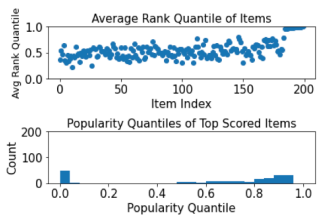
(c) Score Prediction (Zerosum)



(d) Debias Performance (Baseline)



(e) Debias Performance (Pos2Neg2)



(f) Debias Performance (Zerosum)

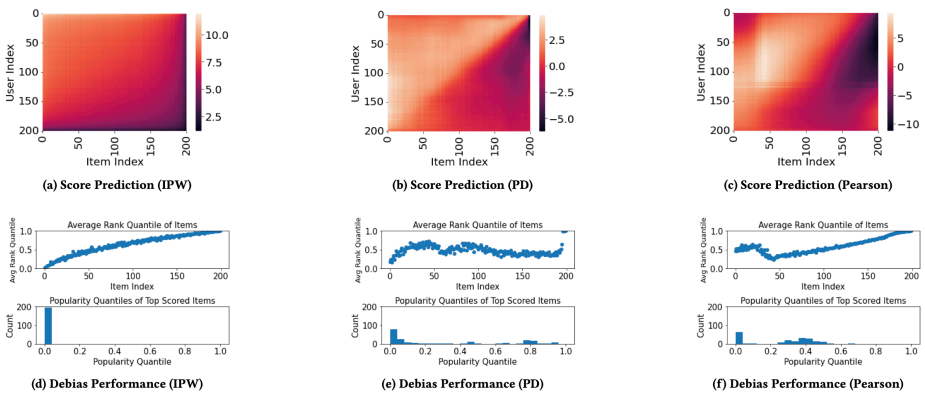
	Baseline	Pos2Neg2	Zerosum
Acc(Error)	0.01%	0.028%	0.007%
PRI	0.99	0.42	0.50
PopQ@1	0.02	0.62	0.61

既存手法との比較

IPW item popularity の逆重みで補正。Pos/Neg での score の差が区別しにくい。

PD Popularity-bias Deconfounding。補正が不均一。NeuCF との相性が悪い。

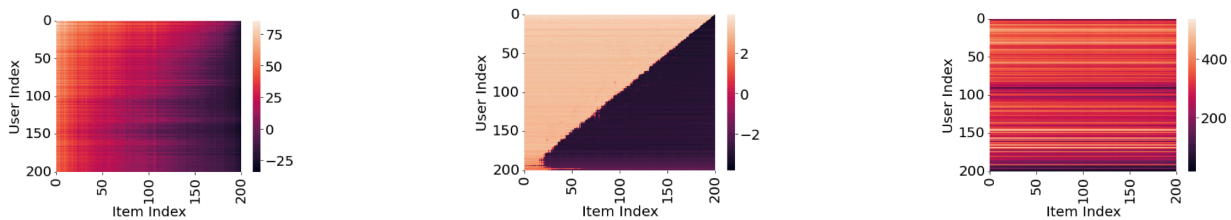
Pearson Pearson 相関係数 (popularity, $\hat{y}$) を 0 にする正規化項。ユーザ間に補正の差がありそう。



(a) Score Prediction (IPW) (b) Score Prediction (PD) (c) Score Prediction (Pearson)

(d) Debias Performance (IPW) (e) Debias Performance (PD) (f) Debias Performance (Pearson)

	Baseline	Zerosum	IPW	PD	Pearson
Acc(Error)	0.01%	0.007%	0.05%	0.01%	0.01%
PRI	0.99	0.50	0.99	-0.52	0.80
PopQ@1	0.02	0.61	0.0	0.35	0.31



(a) Score Prediction (Baseline) (b) Score Prediction (Zerosum) (c) Score Prediction (PD)

よりリアルなデータで評価

- 4 つのデータセットで評価
- 推薦アルゴリズム = BPR, NeuCF, NGCF, LightGCN
- 補正方法 = baseline, IPW, PD, MACR, Pearson, Post-process, Zerosum
- Zerosum は Ciao データセットで性能悪い。baseline の $\hat{y}$ の推測が悪いと上手くいかない。

	#users	#items	#interactions	sparsity
Movielens	6,040	3,260	998,539	0.0507
Gowalla	65,253	57,445	1,339,108	0.0003
Goodreads	14,512	12,385	3,053,619	0.0169
Ciao	4,920	4,394	100,000	0.0046

	Dataset - Movielens											
	MF			NeuCF			NGCF			LightGCN		
	Hit	NDCG	PopQ	Hit	NDCG	PopQ	Hit	NDCG	PopQ	Hit	NDCG	PopQ
Baseline	0.728	0.475	0.181	0.682	0.435	0.172	0.709	0.455	0.163	0.705	0.451	0.137
IPW	0.405	0.224	0.044	0.429	0.233	0.097	0.397	0.218	0.050	0.419	0.235	0.035
PD	**0.715**	**0.457**	**0.266**	0.404	0.198	0.642	0.698	0.441	0.193	0.684	0.431	0.119
MACR	0.475	0.270	0.017	0.326	0.184	0.071	0.478	0.272	0.017	0.476	0.271	0.017
Pearson	**0.729**	**0.457**	**0.414**	0.682	0.430	0.181	0.619	0.347	0.404	0.588	0.322	0.295
Post-Process	0.682	0.371	0.517	0.692	0.433	0.227	0.620	0.319	0.576	0.670	0.378	0.428
Zerosum	**0.718**	**0.449**	**0.383**	**0.662**	**0.344**	**0.291**	**0.710**	**0.444**	**0.318**	**0.703**	**0.437**	**0.314**
	Dataset - Ciao											
Baseline	0.486	0.307	0.201	0.428	0.257	0.194	0.509	0.319	0.187	0.480	0.308	0.118
IPW	0.393	0.244	0.059	0.325	0.209	0.087	0.307	0.141	0.148	0.378	0.235	0.051
PD	0.469	0.289	0.246	0.278	0.134	0.509	0.491	0.301	0.215	0.476	0.301	0.151
MACR	0.419	0.269	0.105	0.310	0.200	0.139	0.359	0.224	0.075	0.404	0.260	0.111
Pearson	0.286	0.165	0.421	0.426	0.259	0.215	0.324	0.150	0.470	0.128	0.064	0.466
Post-Process	0.437	0.239	0.364	0.450	0.285	0.118	0.417	0.197	0.456	0.434	0.247	0.266
Zerosum	0.444	0.286	0.195	0.409	0.250	0.228	0.504	0.306	0.215	0.468	0.287	0.162

まとめ

- そもそも推薦計算モデル自体が imbalanced data だと bias を生じる問題。人気のあるアイテムに不当に高い推薦スコアを出す。
- 提案手法の Zerosum は推薦モデルに依らずに、予測精度低下を抑えつつ、バイアスを提言する補正方法であることが、複数のデータセットを用いた評価で確認出来た。
- 感想
  - 正規化項足すだけで上手くいくのはお手軽で良いと思った。
  - BPRMF 強い。NeuCF や GNN に勝ってる。
  - BPR じゃない普通の MF (iALS とか) には使えないのかな・・・