

Countering Popularity Bias by Regularizing Score Differences

Autores: Wondo Rhee, Sung Min Cho, Bongwon Suh

Enzo Morata - Fabrizio Garcia - Jose Caraball

Sistemas Recomendadores - IIC3633

Tabla de contenidos

1. Contexto
2. El problema
3. Soluciones actuales
4. Medición del problema
5. Solución propuesta
6. Resultados: datos sintéticos
7. Resultados: datos reales
8. Conclusión

Contexto

Métodos de recomendación basados en *feedback* implícito

Aprendizaje contrastivo

Contexto

Para cada usuario se definen dos tipos de *items*.

- ***Items positivos:*** Aquellos *items* con los que un usuario interactúa.
- ***Items negativos:*** Aquellos *items* con los que un usuario NO interactúa.

Objetivo: Maximizar diferencia de puntaje entre los *items* “positivos” y “negativos” de cada usuario.

El problema

El Problema

Un sistema de recomendación puede enfrentarse a dos tipos de sesgos:

- **Sesgo de datos:** Datos tienden a centrarse en una cantidad pequeña de *items*. Distribución de cola larga.
- **Sesgo de modelo:** Modelos entrenados sobre datos con sesgos tienden a entregar puntajes más altos a los *items* más populares.

El Problema

Estos sesgos tienen el efecto negativo de afectar la calidad de las recomendaciones.

- Sobre-recomendación de *items* populares.
- Reducción de diversidad de recomendaciones.
- Falta de *serendipity*.

Los autores se concentran en contrarrestar el **sesgo de modelo ó sesgo de popularidad.**

Soluciones actuales

Soluciones actuales

Los autores destacan una serie de soluciones existentes que tratan de solucionar este problema.

- IPW: *Inverse Propensity Weighting*.
- *Causal Intervention*:
 - PD: *Popularity-bias Deconfounding*.
 - MACR: *Model-Agnostic Counterfactual Reasoning*.
 - Pearson.
 - Post-Process.
- *Reranking**

* Los autores consideran que este método no es una comparación pertinente en el contexto de sesgo de modelo.

Soluciones actuales

Limitaciones de estas soluciones:

- **Impacto en la exactitud:** Penalización de *items* positivos e impulso de *items* negativos.
- **Baja validez computacional:** Dependencia en heurísticas.
- **Baja eficiencia computacional:** Alto costo de cómputo.

Medición del problema

Medición del sesgo en la predicción del modelo

El sesgo lo entenderemos como la correlación de la popularidad y el ranking, o posicionamiento, de las recomendaciones de ítems para los distintos usuarios. Es por ello, que se plantea el uso de las siguientes métricas:

- Popularity Rank Correlation For Items (**PRI**)
- Average Popularity Quantile (**PopQ@1**) *Métrica propuesta

Popularity Rank Correlation For Items (PRI)

$$PRI = -SRC(\text{popularity}(I), \text{avg_rank}(I))$$

donde:

- SRC: Spearman rank correlation coefficient
- popularity: Posición de popularidad del ítem
- avg_rank: Posición promedio de recomendación del ítem

$PRI \sim 1 \implies$ Correlación

$PRI \sim 0 \implies$ No correlación

$PRI \sim -1 \implies$ Correlación

Average Popularity Quantile (PopQ@1)

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

donde:

- PopQuantile: Posición de popularidad del ítem
- $\text{argmax}_{i \in \text{Pos}_u}(\hat{y}_{ui})$: El ítem más recomendado al usuario u

$\text{PopQ@1} \sim 0 \implies$ Sesgado

$\text{PopQ@1} \sim 0,5 \implies$ No sesgado

$\text{PopQ@1} \sim 1 \implies$ Sesgado

Solución propuesta

Solución propuesta

Los autores proponen extender la función de pérdida **BPR** (Bayesian Personalized Ranking) mediante un término de regularización.

$$\text{Total Loss} = \text{BPR Loss} + \text{Reg Term}$$

Mientras BPR Loss maximiza la diferencia de puntaje entre ítems positivos y negativos, Reg Term minimiza la diferencia entre ítems positivos, y negativos, respectivamente.

Respecto al término de regularización, proponen de forma incial dos variaciones: **Pos2Neg2** y **Zerosum**

Reg Term - Pos2Neg2

$$\begin{aligned} \text{Reg Term} = & - \sum_{u \in U} \sum_{\substack{p_1, p_2 \in \text{Pos}_u, \\ n_1, n_2 \in \text{Neg}_u}} \log(1 - \tanh(|\hat{y}_{u,p_1} - \hat{y}_{u,p_2}|)) \\ & + \log(1 - \tanh(|\hat{y}_{u,n_1} - \hat{y}_{u,n_2}|)) \end{aligned}$$

Suma de la diferencia ajustada de puntajes entre pares positivos, y negativos, para cada usuario.

Reg Term - Zerosum

$$\text{Reg Term} = - \sum_{u \in U} \sum_{\substack{p \in \text{Pos}_u, \\ n \in \text{Neg}_u}} \log(1 - \tanh(|\hat{y}_{u,p} - \hat{y}_{u,n}|))$$

Suma de la diferencia ajustada de puntajes entre valores positivos y negativos, para cada usuario.

Esta expresión busca propagar una recomendación simétrica de puntajes para los valores positivos, y negativos.

Resultados: datos sintéticos

Resultados: datos sintéticos

Datos utilizados:

- Datos sintéticos con sesgo explícito reflejados en una matriz de interacción usuario-elemento de 200 x 200.

$$\begin{cases} R[u, i] &= 1 \quad \text{si } i + j \leq 200 \\ R[u, i] &= 0 \quad \text{o.w.} \end{cases}$$

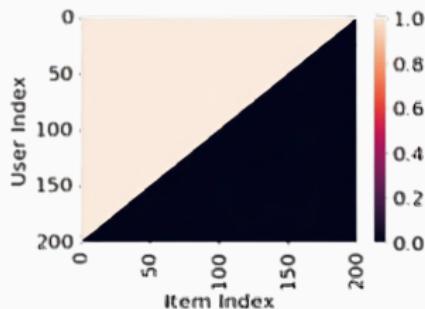


Figura 1: Data Sintética.

Modelo línea base: Factorización Matricial (MF) - Bayesian Personalized Ranking (BPR)

Resultados: datos sintéticos

Comparación entre términos de regulación (Pos2Neg2 y Zerosum) y Línea base para minimizar la diferencia de scores entre los items positivos y negativos

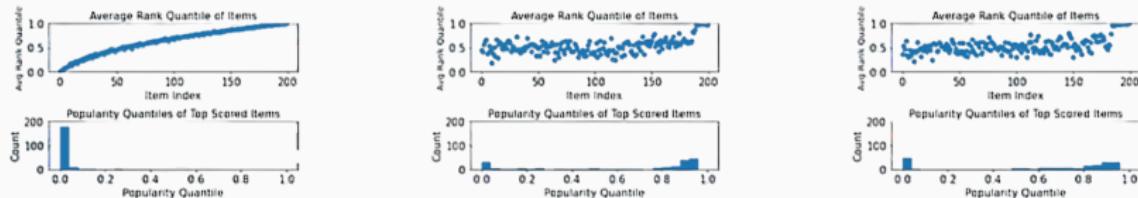


Figura 2: Rendimiento de sesgo para Modelo Línea base, Pos2Neg2 y Zerosum respectivamente

	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

Tabla 1: Precisión y rendimiento del sesgo del modelo linea base, Pos2Neg2 y Zerosum.

Resultados: datos sintéticos

Zerosum > Pos2Neg2

Resultados: datos sintéticos

Comparación con modelos anteriormente utilizados para la corrección de sesgo: IPW, PD y Pearson, junto con el modelo Zerosum y Línea base.

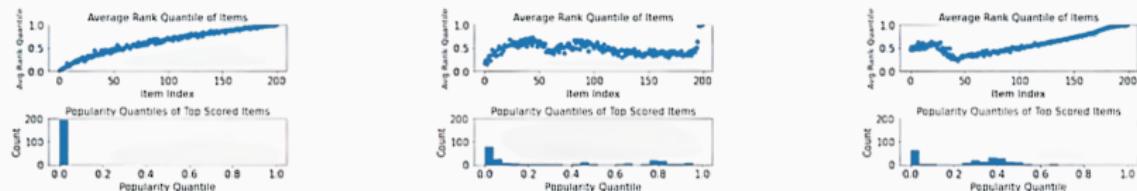


Figura 3: Rendimiento de sesgo para IPW, PD y Pearson respectivamente

	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

Tabla 2: Precisión y rendimiento del sesgo del modelo linea base, Zerosum, IPW, PD, Pearson.

Resultados: datos reales

Resultados: datos reales

Se utilizaron los siguientes 4 datasets preprocesados.

- MovieLens: Sistema recomendador y plataforma de rating de películas.
- Gowalla: Red social basada en ubicaciones.
- Goodreads: Plataforma social de catalogación de libros.
- Ciao: Dataset derivado de la plataforma *Opinions*.

	#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

Tabla 3: Descripción cuantitativa de los 4 datasets a utilizar.

Resultados: datos reales

Para este experimento se utilizó lo siguiente:

- **Modelos:** BPR-MF, NeuCF, NGCF y LightGCN.
- **Métodos:** Línea Base (BPR), IPW, PD, MACR, Pearson, Post-Process y Zerosum.
- **Entrenamiento:**
 - El dataset se subdividió en dataset de entrenamiento (60 %), validación (20 %) y testeo (20 %).
 - Se utilizó la combinación de hiperparámetros que dieran la mejor métrica de precisión para cada dataset, modelo y método.
- **Evaluación.**
 - **Precisión:** Hit@10 y NDCG@10, un mayor valor es mejor.
 - **Sesgo:** PopQ@1, un valor cercano a 0.5 es mejor.

Resultados: datos reales

	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 - Gowalla												
Baseline	0.923	0.706	0.140	0.845	0.609	0.208	0.896	0.658	0.126	0.876	0.622	0.100
IPW	0.863	0.618	0.099	0.194	0.119	0.217	0.301	0.161	0.186	0.773	0.515	0.129
PD	0.922	0.694	0.218	0.771	0.454	0.619	0.903	0.656	0.127	0.870	0.617	0.094
MACR	0.740	0.500	0.112	0.130	0.075	0.358	0.495	0.316	0.099	0.682	0.449	0.116
Pearson	0.924	0.706	0.210	0.807	0.548	0.316	0.900	0.645	0.186	0.860	0.601	0.149
Post-Process	0.899	0.584	0.622	0.793	0.567	0.173	0.862	0.514	0.626	0.801	0.447	0.633
Zerosum	0.918	0.687	0.233	0.780	0.515	0.366	0.899	0.602	0.240	0.865	0.577	0.198
Dataset - Goodreads												
Baseline	0.843	0.614	0.256	0.809	0.570	0.262	0.776	0.514	0.145	0.758	0.495	0.109
IPW	0.468	0.276	0.064	0.306	0.159	0.137	0.319	0.182	0.051	0.492	0.296	0.047
PD	0.835	0.597	0.388	0.680	0.409	0.709	0.768	0.501	0.238	0.732	0.460	0.326
MACR	0.541	0.326	0.013	0.382	0.222	0.038	0.471	0.281	0.007	0.408	0.240	0.004
Pearson	0.842	0.610	0.295	0.802	0.560	0.202	0.765	0.498	0.226	0.743	0.480	0.165
Post-Process	0.793	0.460	0.790	0.808	0.561	0.384	0.654	0.331	0.769	0.646	0.330	0.636
Zerosum	0.852	0.597	0.384	0.798	0.515	0.308	0.776	0.500	0.271	0.751	0.476	0.236
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

Tabla 4: Precisión y rendimiento de sesgo de cada modelo y método propuesto

Resultados: datos reales

Los autores razonan sobre el mal rendimiento e impacto en la exactitud del método propuesto en el dataset *Ciao*: Zerosum empeora modelos con mal rendimiento base.

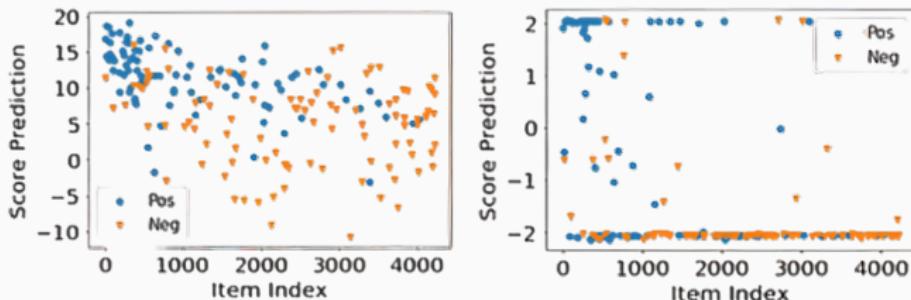


Figura 4: Exacerbación de predicciones erróneas con Zerosum

Conclusión

Conclusión

- Sesgo de los modelos respecto a la popularidad.
- Enfoques previos más costosos, con alto impacto en exactitud, y con menor exactitud y validez computacional.
- Zerosum: Extensión "*drop in*" para modelos que utilizan BPR.
- Disminución significativa del sesgo de popularidad con bajo impacto en exactitud y bajo costo computacional.
- Baja eficacia en modelos con un bajo rendimiento *baseline*.