ADVERSARIAL PERSONALIZED RANKING FOR RECOMMENDATION

Thomas Muñoz

8 de noviembre de 2018

Perturbaciones en Sistemas Recomendadores

Lo robusto de un sistema recomendador es qué tanto aguanta inyecciones de usuarios que apuntan a manipular el modelo mismo.

Perturbaciones en el input

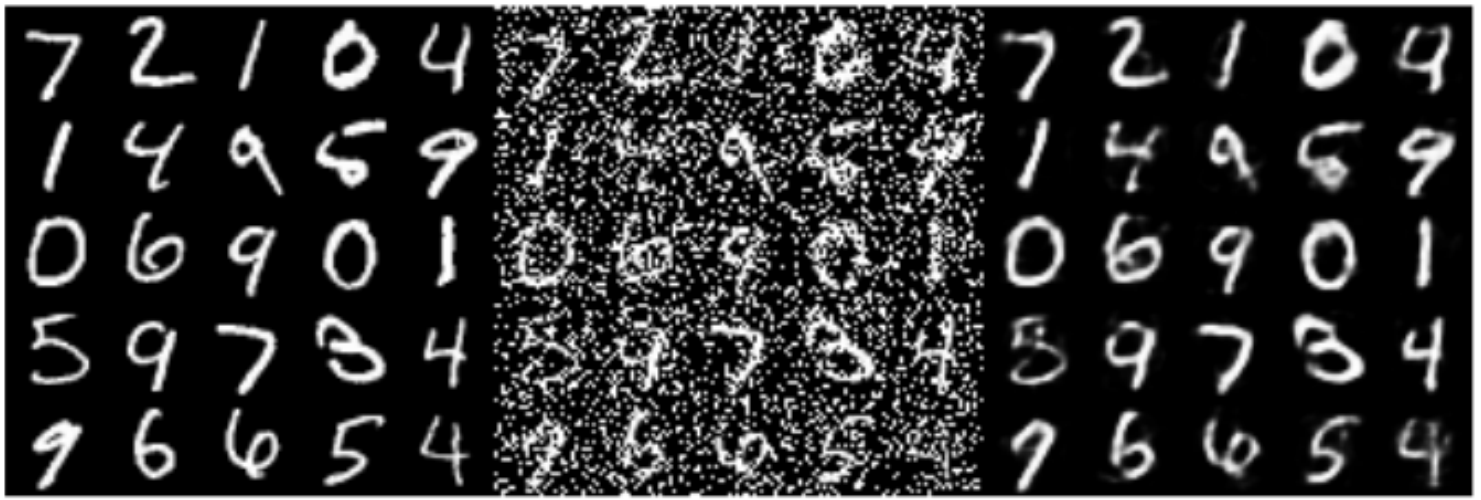

Figure: Perturbaciones aleatorias y adversariales

Las perturbaciones son en los parámetros

El input inicial en general son **entidades** discretas, por lo que perturbaciones en ese dato produce un cambio de significado del dato

$$(u, i, j) \to (u', i, j)$$

Caso de Máquinas de factorización

Los parámetros son los embeddings. Es decir, dados los usuarios $\mathcal{U}$ y los items $\mathcal{I}$, los conjuntos

$$\mathsf{P} = \{p_u\}_{u \in \mathcal{U}}, \mathsf{Q} = \{q_i\}_{i \in \mathcal{I}}$$

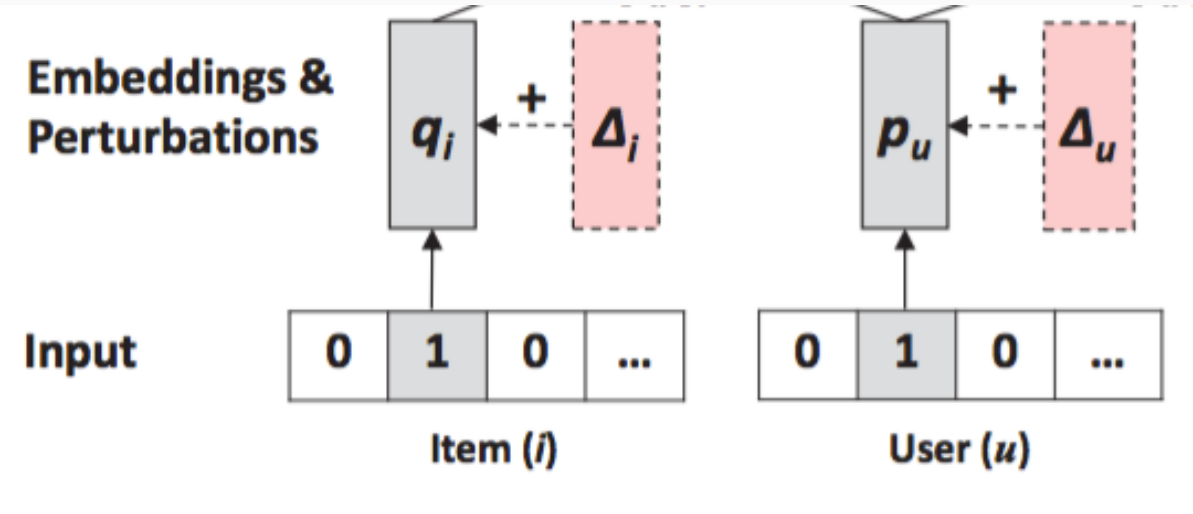

Figure: Perturbaciones en MF

BPR: un método L2R de a pares

Asume que las interacciones observadas tienen que estar mejor rankeadas que las no observadas, y maximiza ese margen.

Es decir, para cada usuario, aprende un orden de items (función de dos variables).

Itera sobre los parámetros de cualquier modelo de recomendación y minimiza la función

$$L_{BPR}(\mathcal{D}|\Theta) = \sum_{(u,i,j)\in\mathcal{D}} -\ln \sigma(\hat{y}_{ui}(\Theta) - \hat{y}_{uj}(\Theta)) + \lambda_\Theta ||\Theta||^2,$$

Perturbaciones adversariales en MF-BPR

Primero se entrena el sistema recomendador y luego se le entrena con perturbaciones.

Se define la perturbación en cada paso de BPR como:

$$\Delta_{adv} = \arg \max_{\Delta, ||\Delta|| \leq \epsilon} L_{BPR}(\mathcal{D}|\hat{\Theta} + \Delta)$$

Mediante linealizaciones se obtiene la perturbación

En cada paso, es decir, dado $\hat{\Theta}$, se tiene que

$$\Delta_{adv} = \epsilon \frac{\Gamma}{||\Gamma||}$$

Donde

$$\Gamma = \frac{\partial L_{BPR}(\mathcal{D}|\hat{\Theta} + \Delta)}{\partial \Delta}$$

Perturbaciones aleatorias vs adversariales

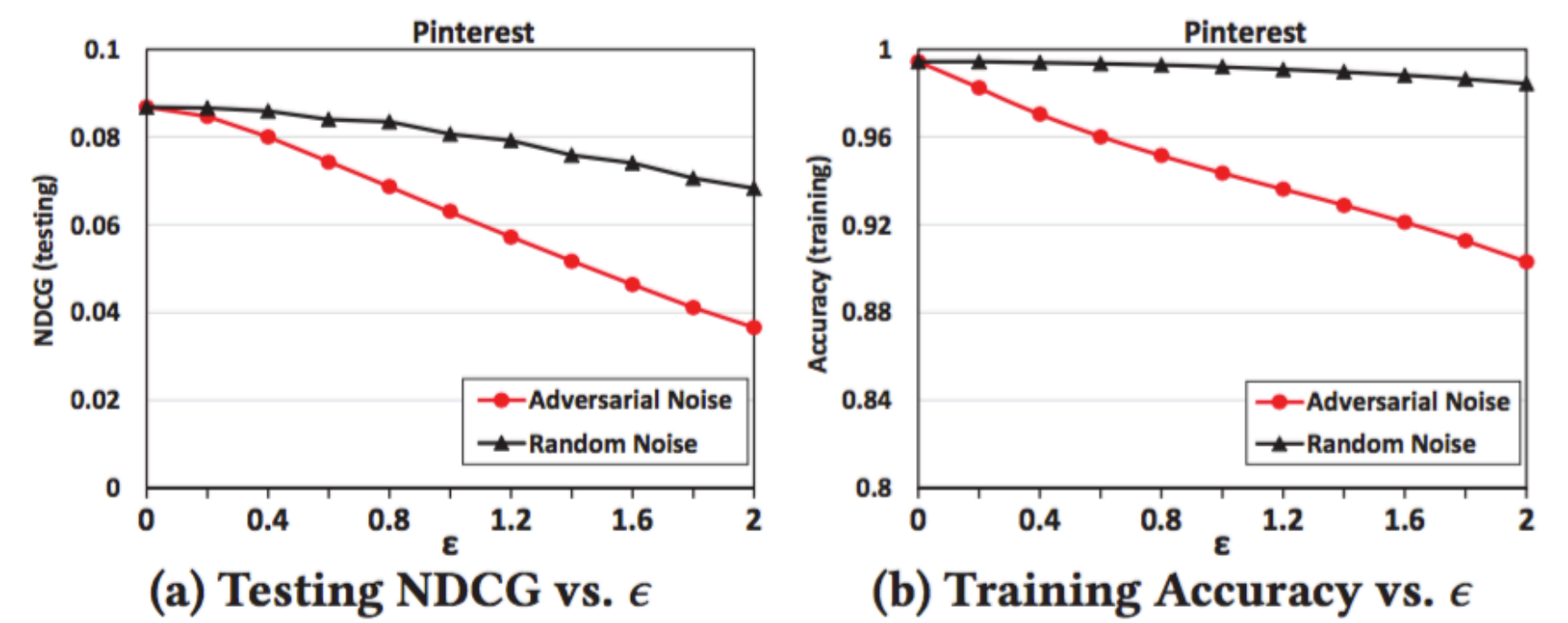

Figure: Comparación en Pinterest

Perturbaciones aleatorias vs adversariales

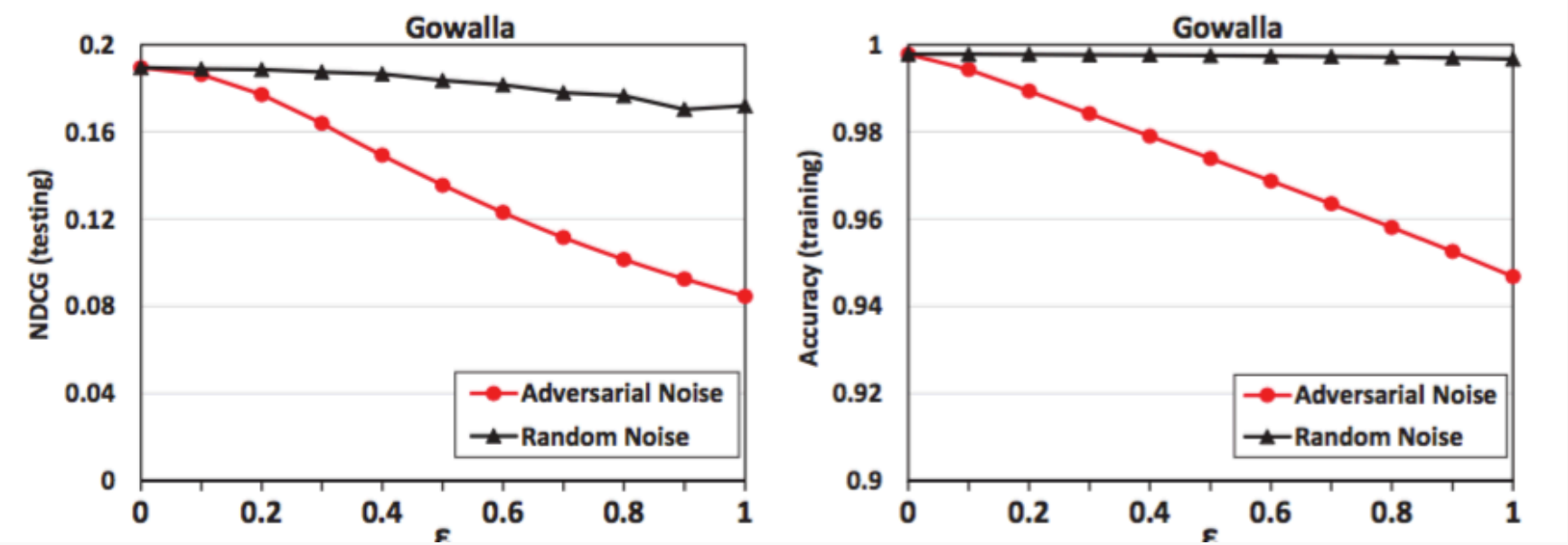

Figure: Comparación en Gowalla

Se minimiza también lo provocado por la perturbación

Luego de entrenar MF con criterio BPR, se sigue entrenando con esta nueva función objetivo a minimizar:

$$L_{APR}(\mathcal{D}|\Theta) = L_{BPR}(\mathcal{D}|\Theta) + \lambda L_{BPR}(\mathcal{D}|\Theta + \Delta_{adv}),$$

$$\text{where} \quad \Delta_{adv} = \arg \max_{\Delta, ||\Delta|| \leq \epsilon} L_{BPR}(\mathcal{D}|\hat{\Theta} + \Delta),$$

Se optimizan los parámetros paso a paso

Se itera por turnos minimizando en Θ y maximizando en Δ.

Para Δ:

$$l_{adv}((u,i,j)|\Delta) = -\lambda \ln \sigma(\hat{y}_{ui}(\hat{\Theta}+\Delta) - \hat{y}_{uj}(\hat{\Theta}+\Delta)).$$

Y se obtiene

$$\Delta_{adv} = \epsilon \frac{\Gamma}{||\Gamma||} \quad \text{where} \quad \Gamma = \frac{\partial l_{adv}((u,i,j)|\Delta)}{\partial \Delta}.$$

Iteración para Ω

$$l_{APR}((u,i,j)|\Theta) = -\ln \sigma(\hat{y}_{ui}(\Theta) - \hat{y}_{uj}(\Theta)) + \lambda_\Theta ||\Theta||^2 - \lambda \ln \sigma(\hat{y}_{ui}(\Theta + \Delta_{adv}) - \hat{y}_{uj}(\Theta + \Delta_{adv})).$$

Y se obtiene

$$\frac{\partial l_{APR}((u,i,j)|\Theta)}{\partial \Theta} = -(1 - \sigma(\hat{y}_{uij}(\Theta))) \frac{\partial \hat{y}_{uij}(\Theta)}{\partial \Theta} + 2\lambda_\Theta \Theta - \lambda(1 - \sigma(\hat{y}_{uij}(\Theta + \Delta_{adv}))) \frac{\partial \hat{y}_{uij}(\Theta + \Delta_{adv})}{\partial \Theta}.$$

Métricas

Los sistemas recomendadores tienen como output una lista de recomendaciones $top - k$ (K=100)

- **HR:** *hit ratio*, una métrica que se basa en el *recall*, si están o no los elementos relevantes.
- **NDCG:** *Normalized Discounted Cumulative Gain*, es sensible a la posición.

Resultados

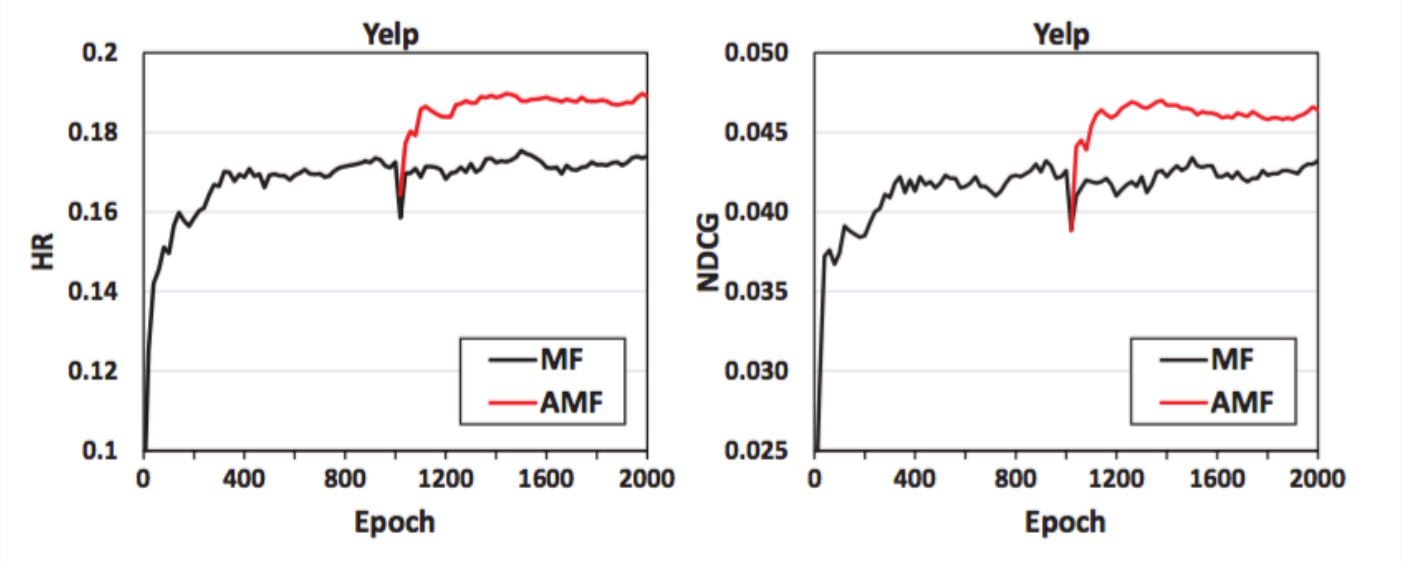

Figure: Entrenamiento en Yelp

Resultados

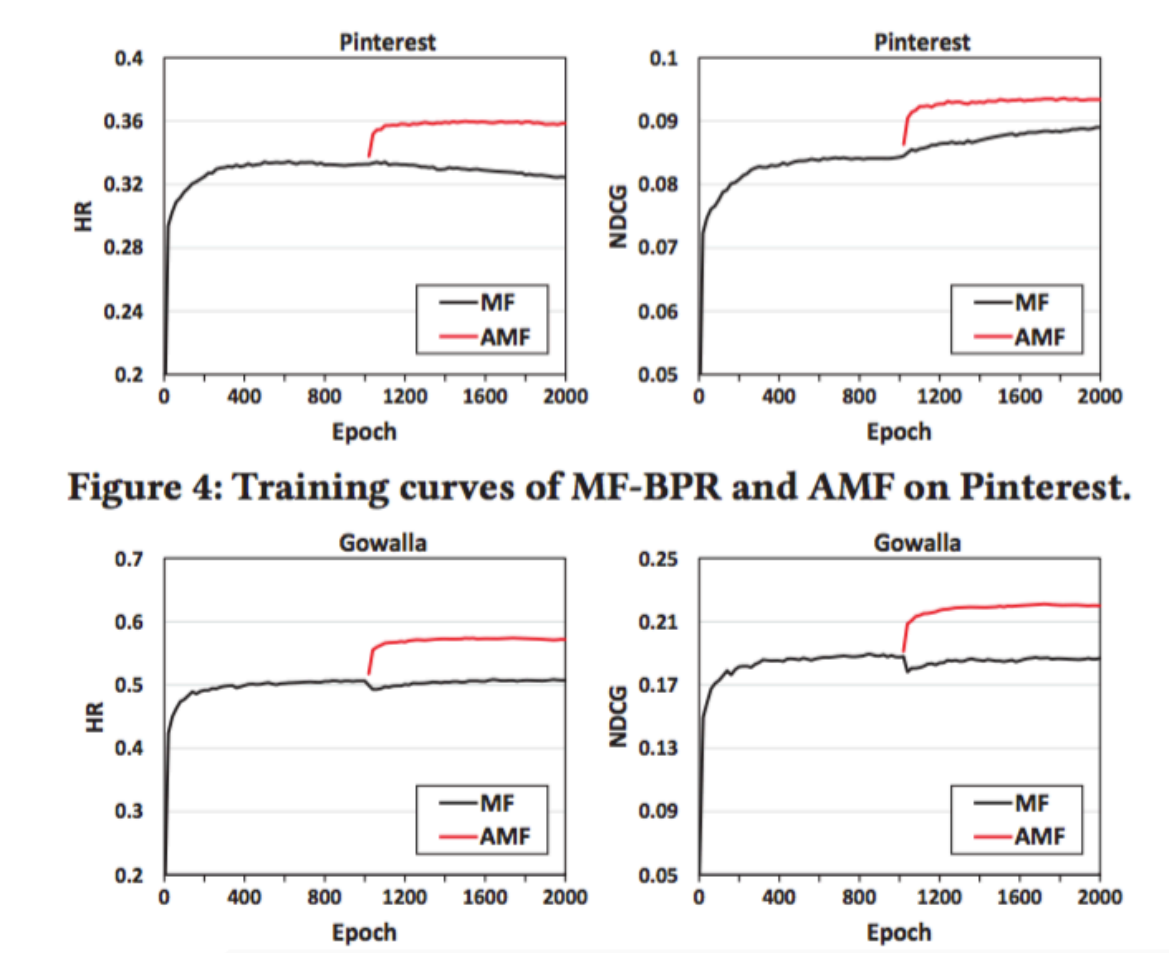

Figure 4: Training curves of MF-BPR and AMF on Pinterest.

Figure: Entrenamiento en Pinterest y Gowalla

Resultados

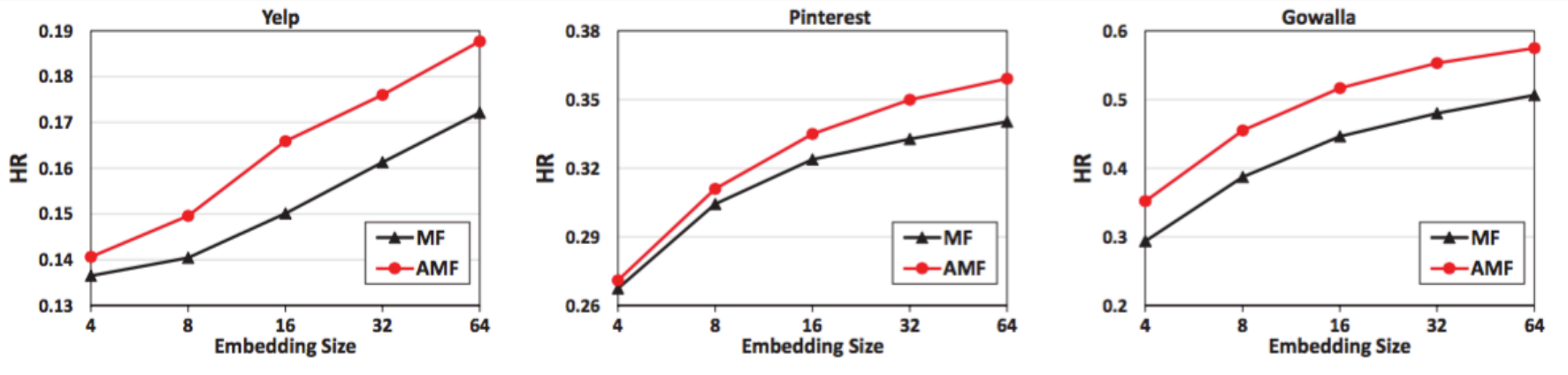

Figure: Efecto del tamaño del embedding

Resultados

Ahora, después de entrenar, se ve el efecto de perturbaciones adversariales en los modelos.

	$\epsilon = 0.5$		$\epsilon = 1.0$		$\epsilon = 2.0$	
Dataset	**BPR**	**APR**	**BPR**	**APR**	**BPR**	**APR**
Yelp	-22.1%	-4.7%	-42.7%	-12.5%	-63.8%	-31.0%
Pinterest	-9.5%	-2.6%	-25.1%	-7.2%	-55.7%	-23.4%
Gowalla	-26.3%	-2.9%	-53.0%	-13.2%	-78.0%	-29.2%

Figure: NDCG: MF-APR vs MF-BPR

Métodos a comparar

- **ItemPop:** más populares.
- **MF-BPR.** máquina de factorización con criterio BRP.
- **CDAE:** modelo que sirve para generalizar factores latentes.
- **NeuMF:** combina MF con perceptrones multicapa para aprender la función de interacción.
- **IRGAN:** discrimina si un dato es generado (perturbación) o es del *dataset*.

Resultados

	Yelp, HR		Yelp, NDCG		Pinterest, HR		Pinterest, NDCG		Gowalla, HR		Gowalla, NDCG	
	K=50	**K=100**	**K=50**	**K=100**	**K=50**	**K=100**	**K=50**	**K=100**	**K=50**	**K=100**	**K=50**	**K=100**
ItemPop	0.0405	0.0742	0.0114	0.0169	0.0294	0.0485	0.0085	0.0116	0.1183	0.1560	0.0367	0.0428
MF-BPR	0.1053	0.1721	0.0312	0.0420	0.2226	0.3403	0.0696	0.0886	0.4061	0.5072	0.1714	0.1878
CDAE [35]	0.1041	0.1733	0.0293	0.0405	0.2254	0.3495	0.0672	0.0873	0.4435	0.5483	0.1837	0.2007
IRGAN [31]	0.1119	0.1765	**0.0361***	**0.0465***	0.2254	0.3363	0.0724	0.0904	0.4157	0.518	0.1853	0.2019
NeuMF [17]	0.1135	0.1817	0.0335	0.0445	0.2342	0.3526	0.0734	0.0925	0.4558	0.5642	0.1962	0.2138
AMF	**0.1176***	**0.1885***	0.0350	**0.0465***	**0.2375***	**0.3595***	**0.0741***	**0.0938***	**0.4693***	**0.5763***	**0.2039***	**0.2212***

Figure: Comparación con distintos métodos

Resultados

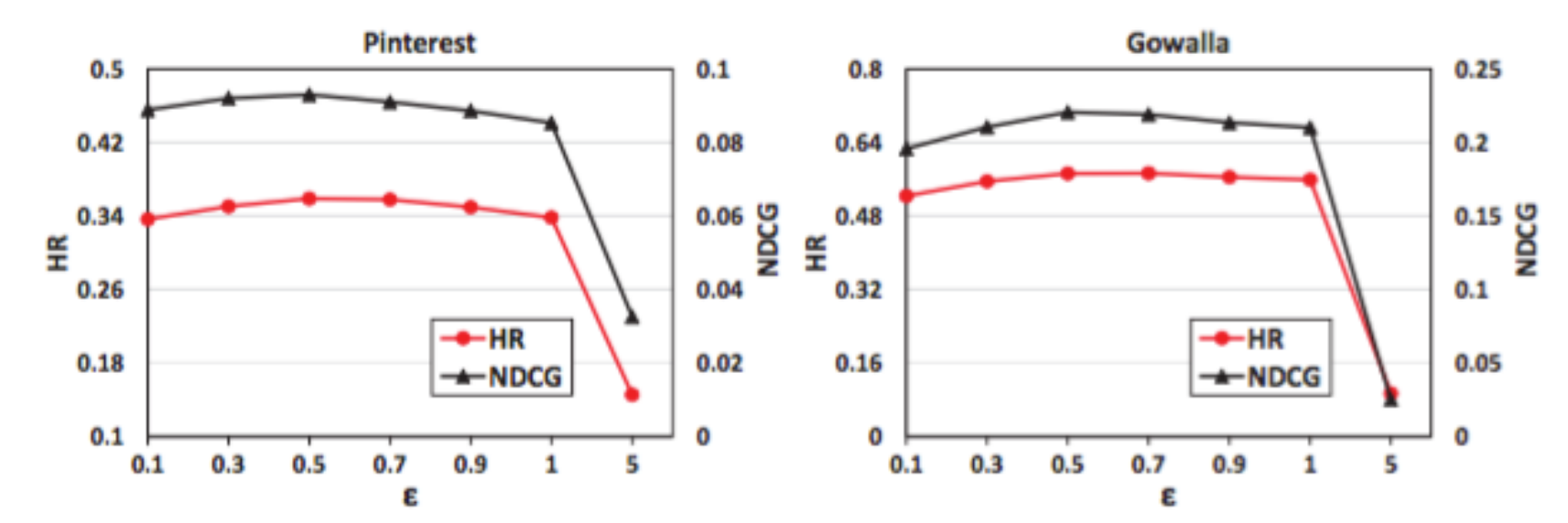

Figure: Efecto de ε

Resultados

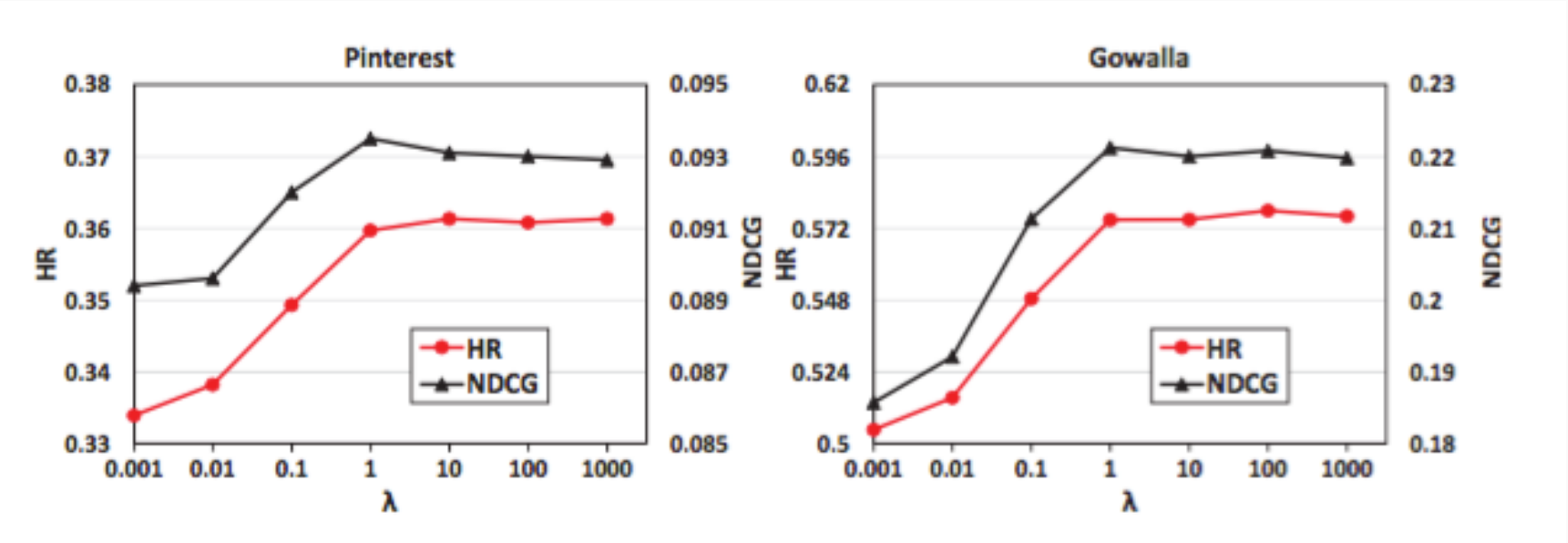

Figure: Efecto de λ

Resultados

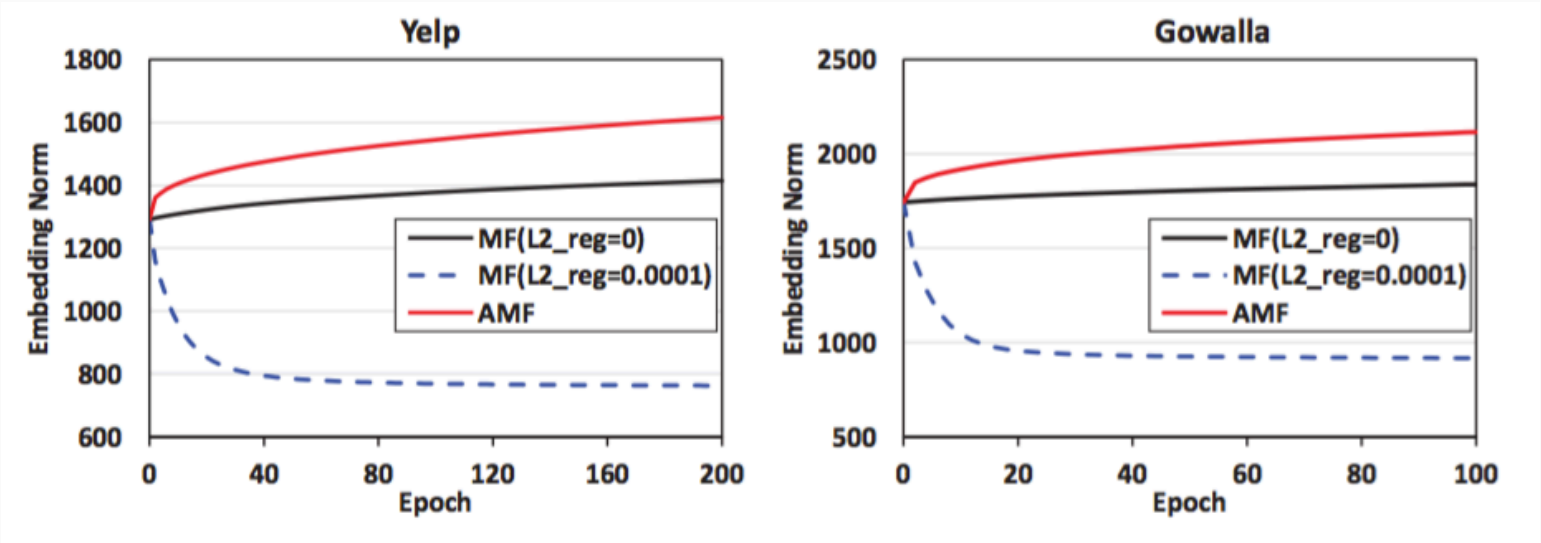

Figure: Efecto en la norma del embedding

Referencias

 He, X., He, Z., Du, X. & Chua, T. *Adversarial Personalized Ranking for Recommendation*, 2018.