Are We Evaluating Rigorously? Benchmarking Recommendation for Reproducible Evaluation and Fair Comparison

1 ADDITIONAL MATERIAL

Table 1. Representative datasets in various domains for recommendation, where 'LBSNs' is location-based social networks.

Domain	Representative Dataset
Movie	MovieLens (100K/1M/10M/20M/25M/Latest), Netflix, Amazon-Movie, FilmTrust, Douban-Moive, Yahoo!Movie, Flixster, CiaoDVD, EachMovie
Music	Last.fm, Yahoo!Music, Douban-Music, KKBox, Kollect.fm, EchoNest
Book	Amazon-Book, Douban-Book, IntentBooks, Book-Crossing, LibraryThing
Image	Pinterest, Flickr, Aesthetic Visual Analysis
Consumable	Amazon Clothing, Amazon Home, Amazon Sports, Amazon Electronic, Taobao, Retailrocket
Social Networks	Epinions, Delicious, Douban, Last.fm, Yelp, Ciao, Xing, FilmTrust
LBSNs	Foursquare, Yelp, Gowalla, BrightKite

Table 2. The equations for the six evaluation metrics.

$Precision@N = \frac{1}{\lvert\mathcal{U}\rvert} \sum_u \frac{1}{N} \sum_{j=1}^{N} rel_j$	$Recall@N = \frac{1}{\lvert\mathcal{U}\rvert} \sum_u \frac{1}{\lvert T(u)\rvert} \sum_{j=1}^{N} rel_j$	$HR@N = \frac{1}{\lvert\mathcal{U}\rvert} \sum_u \delta(R(u) \cap T(u) \neq \emptyset)$
$MRR@N = \frac{1}{\lvert\mathcal{U}\rvert} \sum_u \sum_{j=1}^{N} j^{-1} rel_j$	$MAP@N = \frac{1}{\lvert\mathcal{U}\rvert} \sum_u \frac{1}{N} \sum_{k=1}^{N} Pre@k$	$NDCG@N = \frac{1}{\lvert\mathcal{U}\rvert} \sum_u \frac{DCG@N}{IDCG@N},\ DCG@N = \sum_{j=1}^{N} \frac{2^{rel_j}-1}{log_2(j+1)}$

$R(u), T(u)$ represent sets of recommended items and items in the test set for user u, respectively; $rel_j = 1/0$ indicates whether the item at rank j is in $(R(u) \cap T(u))$; $\delta(x) = 1$ if x is true, otherwise 0; IDCG means the maximum possible DCG through ideal ranking.

Table 3. Objective functions of all baselines.

Baseline	Objective function
BPRMF	$\mathcal{L}_{poi} + f_{cl} = -\sum_{(u,i)\in\widetilde{O}} r_{ui} log(\hat{r}_{ui}) + (1 - r_{ui}) log(1 - \hat{r}_{ui}) + \lambda_\Theta \lVert\Theta\rVert^2 = -\sum_{(u,i)\in O^+} log(\hat{r}_{ui}) - \sum_{(u,i)\in O^-} log(1 - \hat{r}_{ui}) + \lambda_\Theta \lVert\Theta\rVert^2;\ \ \hat{r}_{ui} = \mathbf{v}_u^\top \mathbf{v}_i$ $\mathcal{L}_{pai} + f_{ll} = \sum_{(u,i,j)\in\widetilde{O}} ln(1 + exp(-r_{uij} \cdot \hat{r}_{uij})) + \lambda_\Theta \lVert\Theta\rVert^2 = -\sum_{(u,i,j)\in\widetilde{O}} ln\sigma(\hat{r}_{uij}) + \lambda_\Theta \lVert\Theta\rVert^2;\ \ \hat{r}_{uij} = \hat{r}_{ui} - \hat{r}_{uj};\ \sigma(x) = 1/(1 + exp(-x))\}$ $\mathcal{L}_{pai} + f_{hl} = \sum_{(u,i,j)\in\widetilde{O}} \max(0, 1 - r_{uij} \cdot \hat{r}_{uij}) + \lambda \lVert\Theta\rVert^2$
BPRFM	$\mathcal{L}_{poi} + f_{cl} = -\sum_{(u,i)\in O^+} log(\hat{r}_{ui}) - \sum_{(u,i)\in O^-} log(1 - \hat{r}_{ui}) + \lambda_\Theta \lVert\Theta\rVert^2;\ \ \hat{r}_{ui} = w_0 + w_u + w_i + \mathbf{v}_u^\top \mathbf{v}_i$ $\mathcal{L}_{pai} + f_{ll} = -\sum_{(u,i,j)\in\widetilde{O}} ln\sigma(\hat{r}_{uij}) + \lambda_\Theta \lVert\Theta\rVert^2;\ \ \hat{r}_{uij} = \hat{r}_{ui} - \hat{r}_{uj}$ $\mathcal{L}_{pai} + f_{hl} = \sum_{(u,i,j)\in\widetilde{O}} \max(0, 1 - r_{uij} \cdot \hat{r}_{uij}) + \lambda \lVert\Theta\rVert^2;\ \ \sigma(x) = 1/(1 + exp(-x))$
SLIM	$\mathcal{L}_{poi} + f_{sl} = \frac{1}{2} \sum_{(u,i)\in\widetilde{O}} (r_{ui} - \hat{r}_{ui})^2 + \lambda_\Theta \lVert\Theta\rVert^2;\ \ \hat{r}_{ui} = \mathbf{r}_u^\top \mathbf{w}_i$
NeuMF	$\mathcal{L}_{poi} + f_{cl} = -\sum_{(u,i)\in O^+} log(\hat{r}_{ui}) - \sum_{(u,i)\in O^-} log(1 - \hat{r}_{ui}) + \lambda_\Theta \lVert\Theta\rVert^2$ $\mathcal{L}_{pai} + f_{ll} = -\sum_{(u,i,j)\in\widetilde{O}} ln\sigma(\hat{r}_{uij}) + \lambda_\Theta \lVert\Theta\rVert^2;\ \ \hat{r}_{uij} = \hat{r}_{ui} - \hat{r}_{uj};\ \sigma(x) = 1/(1 + exp(-x))$ $\mathcal{L}_{pai} + f_{hl} = \sum_{(u,i,j)\in\widetilde{O}} \max(0, 1 - r_{uij} \cdot \hat{r}_{uij}) + \lambda \lVert\Theta\rVert^2$

Table 4. The opitmal hyper-parameter settings found by Bayesian HyperOpt for different baselines on the six datasets.

Origin	**Parameter**	**ML-1M**	**Lastfm**	**Yelp**	**Epinions**	**Book-X**	**AMZe**	**Searching space**	**Description**
ItemKNN	−makx	8	49	31	14	83	9	$[1, 100]$	the number of neighbors
PureSVD	−factor	39	24	96	99	10	6	$[1, 100]$	the number of singular values
BPRMF	−num_ng	9	10	10	9	10	10	$[1, 10]$	the number of negative items
	−factors	81	33	58	89	81	100	$[1, 100]$	the dimension of latent factors
	−epochs	50	50	50	50	50	50	–	the number of epochs
	−lr	0.0068	0.0094	0.0091	0.0097	0.0038	0.0018	$[10^{-4}, 10^{-2}]$	learning rate
	−lambda	0.0005	0.0097	0.0087	0.0001	0.0005	0.0006	$[10^{-4}, 10^{-2}]$	regularization coefficient
BPRFM	−num_ng	8	10	10	7	10	9	$[1, 10]$	the number of negative items
	−factors	64	91	98	58	91	42	$[1, 100]$	the dimension of latent factors
	−epochs	50	50	50	50	50	50	–	the number of epochs
	−lr	0.0013	0.0100	0.0093	0.0096	0.0094	0.0016	$[10^{-4}, 10^{-2}]$	learning rate
	−lambda	0.0030	0.0023	0.0005	0.0001	0.0014	0.0002	$[10^{-4}, 10^{-2}]$	regularization coefficient
NeuMF	−num_ng	4	2	5	6	4	7	$[1, 10]$	the number of negative items
	−factors	15	49	68	96	16	83	$[1, 100]$	the dimension of latent factors
	−epochs	50	50	50	50	50	50	–	the number of epochs
	−lr	0.0002	0.0016	0.0001	0.0007	0.0021	0.0004	$[10^{-4}, 10^{-2}]$	learning rate
	−lambda	0.0010	0.0016	0.0009	0.0002	0.0056	0.0015	$[10^{-4}, 10^{-2}]$	regularization coefficient
	−num_layers	3	3	2	2	3	2	$[1, 3]$	the number of layers for MLP
	−dropout	0.9531	0.7890	0.7371	0.8351	0.8592	0.7162	$[0, 1]$	dropout ratio
	−batch_size	64	512	1024	128	256	256	$[2^6, 2^7, 2^8, 2^9, 2^{10}]$	batch size
5-filter	**Parameter**	**ML-1M**	**Lastfm**	**Yelp**	**Epinions**	**Book-X**	**AMZe**	**Searching space**	**Description**
ItemKNN	−makx	73	55	75	30	65	41	$[1, 100]$	the number of neighbors
PureSVD	−factor	2	33	100	70	80	2	$[1, 100]$	the number of singular values
BPRMF	−num_ng	1	9	10	10	9	10	$[1, 10]$	the number of negative items
	−factors	10	98	58	76	61	100	$[1, 100]$	the dimension of latent factors
	−epochs	50	50	50	50	50	50	–	the number of epochs
	−lr	0.0019	0.0062	0.0091	0.0081	0.0037	0.0018	$[10^{-4}, 10^{-2}]$	learning rate
	−lambda	0.0003	0.0008	0.0087	0.0004	0.0023	0.0006	$[10^{-4}, 10^{-2}]$	regularization coefficient
BPRFM	−num_ng	9	10	10	9	8	9	$[1, 10]$	the number of negative items
	−factors	2	75	52	96	100	42	$[1, 100]$	the dimension of latent factors
	−epochs	50	50	50	50	50	50	–	the number of epochs
	−lr	0.0020	0.0037	0.0090	0.0070	0.0096	0.0016	$[10^{-4}, 10^{-2}]$	learning rate
	−lambda	0.0015	0.0007	0.0006	0.0004	0.0018	0.0002	$[10^{-4}, 10^{-2}]$	regularization coefficient
NeuMF	−num_ng	5	4	5	6	4	7	$[1, 10]$	the number of negative items
	−factors	47	67	68	96	16	83	$[1, 100]$	the dimension of latent factors
	−epochs	50	50	50	50	50	50	–	the number of epochs
	−lr	0.0007	0.0002	0.0001	0.0007	0.0021	0.0004	$[10^{-4}, 10^{-2}]$	learning rate
	−lambda	0.0039	0.0009	0.0009	0.0002	0.0056	0.0015	$[10^{-4}, 10^{-2}]$	regularization coefficient
	−num_layers	1	3	2	2	3	2	$[1, 3]$	the number of layers for MLP
	−dropout	0.5919	0.9973	0.7371	0.8351	0.8592	0.7162	$[0, 1]$	dropout ratio
	−batch_size	128	512	1024	128	256	256	$[2^6, 2^7, 2^8, 2^9, 2^{10}]$	batch size
10-filter	**Parameter**	**ML-1M**	**Lastfm**	**Yelp**	**Epinions**	**Book-X**	**AMZe**	**Searching space**	**Description**
ItemKNN	−makx	100	28	73	14	71	68	$[1, 100]$	the number of neighbors
PureSVD	−factor	39	12	93	98	98	9	$[1, 100]$	the number of singular values
BPRMF	−num_ng	2	8	9	9	10	4	$[1, 10]$	the number of negative items
	−factors	34	92	75	97	98	89	$[1, 100]$	the dimension of latent factors
	−epochs	50	50	50	50	50	50	–	the number of epochs
	−lr	0.0005	0.0053	0.0095	0.0038	0.0087	0.0057	$[10^{-4}, 10^{-2}]$	learning rate
	−lambda	0.0016	0.0034	0.0002	0.0005	0.0099	0.0004	$[10^{-4}, 10^{-2}]$	regularization coefficient
BPRFM	−num_ng	6	9	9	10	10	7	$[1, 10]$	the number of negative items
	−factors	78	61	81	78	100	73	$[1, 100]$	the dimension of latent factors
	−epochs	50	50	50	50	50	50	–	the number of epochs
	−lr	0.0015	0.0031	0.0094	0.0072	0.0055	0.0001	$[10^{-4}, 10^{-2}]$	learning rate
	−lambda	0.0015	0.0006	0.0029	0.0004	0.0075	0.0008	$[10^{-4}, 10^{-2}]$	regularization coefficient
SLIM	−l1_ratio	0.4462	0.8330	0.0459	0.3951	0.8188	0.3299	$(0, 1]$	the ElasticNet mixing parameter
	−lambda	0.0001	0.0013	0.0040	0.0022	0.0072	0.0020	$[10^{-4}, 10^{-2}]$	constant to multiply penalty terms
NeuMF	−num_ng	4	2	5	5	9	7	$[1, 10]$	the number of negative items
	−factors	15	93	68	41	99	83	$[1, 100]$	the dimension of latent factors
	−epochs	50	50	50	50	50	50	–	the number of epochs
	−lr	0.0002	0.0036	0.0001	0.0012	0.0014	0.0004	$[10^{-4}, 10^{-2}]$	learning rate
	−lambda	0.0010	0.0011	0.0009	0.0057	0.0006	0.0015	$[10^{-4}, 10^{-2}]$	regularization coefficient
	−num_layers	3	2	2	3	3	2	$[1, 3]$	the number of layers for MLP
	−dropout	0.9531	0.5608	0.7371	0.4771	0.4834	0.7162	$[0, 1]$	dropout ratio
	−batch_size	64	512	1024	128	128	256	$[2^6, 2^7, 2^8, 2^9, 2^{10}]$	batch size

1. The detailed explanation for the parameters of SLIM is available at https://lijiancheng0614.github.io/scikit-learn/modules/generated/sklearn.linear_model.ElasticNet.html

2. For NeuMF, the searching space of batch size on ML-1M, Lastfm, Book-X, and Epinions is $[2^6, 2^7, 2^8, 2^9]$; while on Yelp and AMZe is $[2^8, 2^9, 2^{10}]$ to speed up the training.

Table 5. Paper Collection.

KDD Paper List (9)
[1] Collaborative variational autoencoder for recommender systems, 2017; [2] Learning tree-based deep model for recommender systems, 2018;
[3] Local latent space models for top-n recommendation, 2018;
[4] Leveraging meta-path based context for top-n recommendation with a neural co-attention model, 2018;
[5] KGAT: Knowledge graph attention network for recommendation, 2019;
[6] AKUPM: Attention-enhanced knowledge-aware user preference model for recommendation, 2019;
[7] Knowledge-aware graph neural networks with label smoothness regularization for recommender systems, 2019;
[8] LambdaOpt: Learn to regularize recommender models in finer levels, 2019;
[9] IntentGC: A scalable graph convolution framework fusing heterogeneous information for recommendation, 2019
SIGIR Paper List (10)
[1] Attentive collaborative filtering: Multimedia recommendation with item-and component-level attention, 2017;
[2] Adversarial personalized ranking for recommendation, 2018; [3] Collaborative memory network for recommendation systems, 2018;
[4] Graphcar: Content-aware multimedia recommendation with graph autoencoder, 2018;
[5] Should i follow the crowd?: A probabilistic analysis of the effectiveness of popularity in recommender systems, 2018;
[6] Streaming ranking based recommender systems, 2018;
[7] Bayesian personalized feature interaction selection for factorization machines, 2019; [8] Neural graph collaborative filtering, 2019;
[9] Noise contrastive estimation for one-class collaborative filtering, 2019;
[10] Relational collaborative filtering: modeling multiple item relations for recommendation, 2019
WWW Paper List (9)
[1] Collaborative metric learning, 2017; [2] Neural collaborative filtering, 2017; [3] Aesthetic-based clothing recommendation, 2018;
[4] Multi-task feature learning for knowledge graph enhanced recommendation, 2019;
[5] Jointly learning explainable rules for recommendation with knowledge graph, 2019;
[6] Unifying knowledge graph learning and recommendation: Towards a better understanding of user preferences, 2019;
[7] Signed distance-based deep memory recommender, 2019;
[8] Knowledge graph convolutional networks for recommender systems, 2019; [9] Collaborative similarity embedding for recommender systems, 2019
IJCAI Paper List (16)
[1] Learning discriminative recommendation systems with side information, 2017;
[2] MRLR: Multi-level representation learning for personalized ranking in recommendation, 2017;
[3] Deep matrix factorization models for recommender systems, 2017;
[4] Dynamic Bayesian logistic matrix factorization for recommendation with implicit feedback, 2018;
[5] Matrix completion with preference ranking for top-n recommendation, 2018; [6] Discrete factorization machines for fast feature-based recommendation, 2018;
[7] PLASTIC: Prioritize long and short-term information in top-n recommendation using adversarial training, 2018;
[8] Improving implicit recommender systems with view data, 2018; [9] DELF: A dual-embedding based deep latent factor model for recommendation, 2018;
[10] CFM: convolutional factorization machines for context-aware recommendation, 2019;
[11] Convolutional Gaussian embeddings for personalized recommendation with uncertainty, 2019;
[12] Discrete trust-aware matrix factorization for fast recommendation, 2019;
[13] Learning shared vertex representation in heterogeneous graphs with convolutional networks for recommendation, 2019;
[14] Deep adversarial social recommendation, 2019; [15] Quaternion collaborative filtering for recommendation, 2019;
[16] Unified embedding model over heterogeneous information network for personalized recommendation, 2019
AAAI Paper List (13)
[1] Exploiting both vertical and horizontal dimensions of feature hierarchy for effective recommendation, 2017;
[2] ERMMA: Expected risk minimization for matrix approximation-based recommender systems, 2017;
[3] Walkranker: A unified pairwise ranking model with multiple relations for item recommendation, 2018;
[4] Collaborative filtering with social exposure: A modular approach to social recommendation, 2018;
[5] Coupled poisson factorization integrated with user/item metadata for modeling popular and sparse ratings in scalable recommendation, 2018;
[6] Hierarchical reinforcement learning for course recommendation in MOOCs, 2019;
[7] CAMO: A collaborative ranking method for content based recommendation, 2019;
[8] Non-compensatory psychological models for recommender systems, 2019; [9] From zero-shot learning to cold-start recommendation, 2019;
[10] Discrete social recommendation, 2019; [11] Explainable reasoning over knowledge graphs for recommendation, 2019;
[12] Hers: Modeling influential contexts with heterogeneous relations for sparse and cold-start recommendation, 2019;
[13] Deepcf: A unified framework of representation learning and matching function learning in recommender system, 2019
RecSys Paper List (11)
[1] Mpr: Multi-objective pairwise ranking, 2017; [2] Learning to rank with trust and distrust in recommender systems, 2017;
[3] On the robustness and discriminative power of information retrieval metrics for top-n recommendation, 2018; [4] Spectral collaborative filtering, 2018;
[5] Recurrent knowledge graph embedding for effective recommendatio, 2018;
[6] Asymmetric Bayesian personalized ranking for one-class collaborative filtering, 2019;
[7] Deep generative ranking for personalized recommendation, 2019; [8] Personalized diffusions for top-n recommendation, 2019;
[9] Collective embedding for neural context-aware recommender systems, 2019; [10] HybridSVD: When collaborative information is not enough, 2019;
[11] Variational low rank multinomials for collaborative filtering with side-information, 2019
WSDM Paper List (8)
[1] Multi-product utility maximization for economic recommendation, 2017; [2] Discrete deep learning for fast content-aware recommendation, 2018;
[3] Recommendation in heterogeneous information networks based on generalized random walk model and Bayesian personalized ranking, 2018;
[4] Neural personalized ranking for image recommendation, 2018; [5] Gated attentive-autoencoder for content-aware recommendation, 2019;
[6] Recwalk: Nearly uncoupled random walks for top-n recommendation, 2019; [7] Spiral of silence in recommender Systems, 2019;
[8] Social attentional memory network: Modeling aspect-and friend-level differences in recommendation, 2019
CIKM Paper List (9)
[1] Joint representation learning for top-n recommendation with heterogeneous information sources, 2017;
[2] Indexable Bayesian personalized ranking for efficient top-k recommendation, 2017; [3] Interacting attention-gated recurrent networks for recommendation, 2017;
[4] An attentive interaction network for context-aware recommendations, 2018;
[5] Ripplenet: Propagating user preferences on the knowledge graph for recommender systems, 2018;
[6] Regularizing matrix factorization with user and item embeddings for recommendation, 2018;
[7] DBRec: Dual-bridging recommendation via discovering latent groups, 2019;
[8] Candidate generation with binary codes for large-scale top-n recommendation, 2019;
[9] Relation-aware graph convolutional networks for agent-initiated social e-commerce recommendation, 2019