Ranking-Oriented Evaluation Metrics

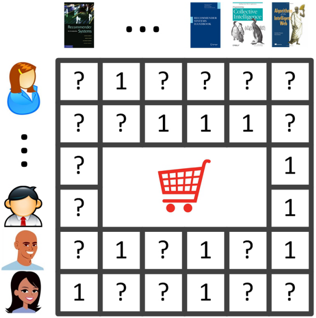
Weike Pan

College of Computer Science and Software Engineering
Shenzhen University

Outline

Recommendation with Implicit Feedback

- We may represent users' implicit feedback in a ***matrix*** form:



- If we can estimate the missing values (denoted as "?") in the matrix or **rank the items directly**, we can make recommendations for each user.

Notations (1/2)

Table: Some notations.

n	user number
m	item number
$u \in \{1, 2, \ldots, n\}$	user ID
$i, j \in \{1, 2, \ldots, m\}$	item ID
$\mathcal{R} = \{(u, i)\}$	(user, item) pairs in training data
$y_{ui} \in \{1, 0\}$	indicator variable, $y_{ui} = 1$ if $(u, i) \in \mathcal{R}$
$\mathcal{I}_u$	preferred items by user u in training data
$\mathcal{I}$	the whole item set
$\mathcal{U}$	the whole user set

Notations (2/2)

Table: Some notations.

$\mathcal{R}^{te} = \{(u, i)\}$	(user, item) pairs in test data
$\mathcal{I}_u^{te}$	preferred items by user u in test data
$\mathcal{U}^{te}$	user set in test data
$\hat{r}_{ui}$	predicted rating of user u on item i
$\mathcal{I}_u^{re}$	recommended items for user u

Top-k Recommended Items

Assume we have a ranked recommendation list of items for user u as generated by some recommendation method,

$$\mathcal{I}_u^{re} = \{i(1), \ldots, i(\ell), \ldots, i(k)\} \in \mathcal{I} \backslash \mathcal{I}_u$$

where $i(\ell)$ represents the item located at position ℓ.

How Shall We Evaluate the Recommendation Performance?

- Compare the ranked recommendation list of items for user u, i.e., $\mathcal{I}_u^{re}$, with the preferred items by user u in test data, i.e., $\mathcal{I}_u^{te}$

Pre@k

The precision of user u is defined as,

$$Pre_u@k = \frac{1}{k}\sum_{\ell=1}^{k}\delta(i(\ell) \in \mathcal{I}_u^{te}),$$

where $\delta(x) = 1$ if x is true and $\delta(x) = 0$ otherwise. $\sum_{\ell=1}^{k}\delta(i(\ell) \in \mathcal{I}_u^{te}) = |\mathcal{I}_u^{re} \cap \mathcal{I}_u^{te}|$ thus denotes the number of items in $\mathcal{I}_u^{re} \cap \mathcal{I}_u^{te}$.

Then, we have

$$Pre@k = \sum_{u \in \mathcal{U}^{te}} Pre_u@k / |\mathcal{U}^{te}| \qquad (1)$$

Rec@k

The recall of user u is defined as,

$$Rec_u@k = \frac{1}{|\mathcal{I}_u^{te}|} \sum_{\ell=1}^{k} \delta(i(\ell) \in \mathcal{I}_u^{te}),$$

which means how many preferred items in $\mathcal{I}_u^{te}$ are also in $\mathcal{I}_u^{re}$.

Then, we have

$$Rec@k = \sum_{u \in \mathcal{U}^{te}} Rec_u@k / |\mathcal{U}^{te}| \tag{2}$$

F1@k

The F1 score of user u is defined as,

$$F1_u@k = 2 \times \frac{Pre_u@k \times Rec_u@k}{Pre_u@k + Rec_u@k}.$$

Then, we have

$$F1@k = \sum_{u \in \mathcal{U}^{te}} F1_u@k / |\mathcal{U}^{te}| \quad (3)$$

NDCG@k

The NDCG of user u is defined as,

$$NDCG_u@k = \frac{1}{Z_u} DCG_u@k,$$

where $DCG_u@k = \sum_{\ell=1}^{k} \frac{2^{\delta(i(\ell) \in \mathcal{I}_u^{te})} - 1}{\log(\ell+1)}$ and Z_u is the best $DCG_u@k$ score with preferred items in $\mathcal{I}_u^{te}$ in the beginning of $\mathcal{I}_u^{re}$.

Then, we have

$$NDCG@k = \sum_{u \in \mathcal{U}^{te}} NDCG_u@k / |\mathcal{U}^{te}| \tag{4}$$

- NDCG emphasizes the items ranked in the beginning (i.e., location dependent)

1-call@k

The 1-call of user u is defined as,

$$1\text{-}call_u@k = \delta(\sum_{\ell=1}^{k} \delta(i(\ell) \in \mathcal{I}_u^{te}) \geq 1),$$

which means whether there is at least one preferred item in $\mathcal{I}_u^{re}$.

Then, we have

$$1\text{-}call@k = \sum_{u \in \mathcal{U}^{te}} 1\text{-}call_u@k / |\mathcal{U}^{te}| \tag{5}$$

Mean Reciprocal Rank (MRR)

The reciprocal rank of user u is defined as,

$$RR_u = \frac{1}{\min_{i \in \mathcal{I}_u^{te}}(p_{ui})}$$

where $\min_{i \in \mathcal{I}_u^{te}}(p_{ui})$ is the position of the first preferred item in $\mathcal{I}_u^{re}$.

Then, we have

$$MRR = \sum_{u \in \mathcal{U}^{te}} RR_u / |\mathcal{U}^{te}| \tag{6}$$

Mean Average Precision (MAP)

The average precision of user u is defined as,

$$AP_u = \frac{1}{|\mathcal{I}_u^{te}|} \sum_{i \in \mathcal{I}_u^{te}} \left[\frac{1}{p_{ui}} \left(\sum_{j \in \mathcal{I}_u^{te}} \delta(p_{uj} \prec p_{ui}) + 1 \right) \right]$$

where p_{ui} is the ranking position of item i. $p_{uj} \prec p_{ui}$ means that item j is ranked before item i for user u.

Then, we have

$$MAP = \sum_{u \in \mathcal{U}^{te}} AP_u / |\mathcal{U}^{te}| \tag{7}$$

Average Relative Position (ARP)

The relative position of user u is defined as,

$$RP_u = \frac{1}{|\mathcal{I}_u^{te}|} \sum_{i \in \mathcal{I}_u^{te}} \frac{p_{ui}}{|\mathcal{I}| - |\mathcal{I}_u|}$$

where $\frac{p_{ui}}{|\mathcal{I}|-|\mathcal{I}_u|}$ is the relative position of item i.

Then, we have

$$ARP = \sum_{u \in \mathcal{U}^{te}} RP_u / |\mathcal{U}^{te}| \tag{8}$$

Area Under the Curve (AUC)

The AUC of user u is defined as,

$$AUC_u = \frac{1}{|\mathcal{R}^{te}(u)|} \sum_{(i,j) \in \mathcal{R}^{te}(u)} \delta(\hat{r}_{ui} > \hat{r}_{uj})$$

where $\mathcal{R}^{te}(u) = \{(i,j) | (u,i) \in \mathcal{R}^{te}, (u,j) \notin \mathcal{R} \cup \mathcal{R}^{te}\}$.

Then, we have

$$AUC = \sum_{u \in \mathcal{U}^{te}} AUC_u / |\mathcal{U}^{te}| \tag{9}$$

PopRank

We may use the bias of each item as the predicted rating,

$$\hat{r}_{ui} = b_i = \sum_{u=1}^{n} y_{ui}/n - \mu \tag{10}$$

where $\mu = \sum_{u=1}^{n} \sum_{i=1}^{m} y_{ui}/n/m$.

Data Set

- We use the files u1.base and u1.test of MovieLens100K[1] as our training data and test data, respectively.
- user number: $n = 943$; item number: $m = 1682$.
- u1.base (training data): 80000 rating records, and the density (or sparsity) is $80000/943/1682 = 5.04\%$.
- u1.test (test data): 20000 rating records.
- Pre-processing (for simulation): we only keep the (user, item) pairs with ratings 4 or 5 in u1.base and u1.test as preferred (user,item) pairs, and remove all other records. Finally, we obtain u1.base.OCCF and u1.test.OCCF.

[1] http://grouplens.org/datasets/

Results

Table: Prediction performance of PopRank on MovieLens100K (u1.base.OCCF, u1.test.OCCF).

	PopRank
Pre@5	0.2338
Rec@5	0.0571
*F*1@5	0.0775
NDCG@5	0.2568
1–*call*@5	0.5877
MRR	0.4657
MAP	0.1516
ARP	0.1551
AUC	0.8516

Conclusion

- Different ranking-oriented evaluation metrics with different emphasis

Homework

- Implement the ranking-oriented evaluation metrics and conduct empirical studies of PopRank on u2.base.OCCF, u2.test.OCCF of MovieLens100K with similar pre-processing
- Design some new evaluation metrics
- Design a new recommendation method to beat PopRank