Homework 2: Evaluation Metrics

Student ID: 18329015 Student Name: 郝裕玮

Date: 2022.3.19

Lectured by: Shangsong Liang
Information Retrieval Course
Sun Yat-sen University

Exercise 1: Rank-based Evaluation Metrics, MAP@K, MRR@K

练习 1：基于等级的评估指标，MAP@K，MRR@K

Assume you have three queries, and the ranking results that a system in response to these three queries are as follows:

Ranking 1 in response to query #1 is: d1, d2, d3, d4, d5, d6, d7, d8, d9, d10. Here only d1, d3, d4, d6, d7, and d10 are relevant (relevance is binary, i.e., either 1 if relevant or 0 if non-relevant) in response to query #1.

Ranking 2 in response to query #2 is: d3, d8, d7, d1, d2, d4, d5, d9, d10, d6. Here only d8 and d9 are relevant in response to query #2.

Ranking 3 in response to query #3 is: d7, d6, d5, d3, d2, d1, d9, d10, d4, d8. Here only d5, d9, and d8 are relevant in response to query #3.

Answer the questions below.

(a) Compute the scores for these metrics: AP@5 (Average Precision @5), AP@10 for each query; RR@5 (Reciprocal Rank score @5), RR@10 for each query.

(b) Compute the scores for these metrics: MAP@5 (Mean Average Precision @5), MAP@10, MRR@5 (Mean Reciprocal Rank score @5), MRR@10 for this system.

对于问题(a):

（1）对于 Ranking 1：

$$AP@5 = \frac{1+\frac{2}{3}+\frac{3}{4}}{3} = 0.805556$$

$$AP@10 = \frac{1+\frac{2}{3}+\frac{3}{4}+\frac{4}{6}+\frac{5}{7}+\frac{6}{10}}{6} = 0.732937$$

$$RR@5 = 1$$

$$RR@10 = 1$$

（2）对于 Ranking 2：

$$AP@5 = \frac{\frac{1}{2}}{1} = 0.5$$

$$AP@10 = \frac{\frac{1}{2}+\frac{2}{8}}{2} = 0.375$$

$$RR@5 = \frac{1}{2}$$

$$RR@10 = \frac{1}{2}$$

（3）对于 Ranking 3：

$$AP@5 = \frac{\frac{1}{3}}{1} = 0.333333$$

$$AP@10 = \frac{\frac{1}{3} + \frac{2}{7} + \frac{3}{10}}{3} = 0.306349$$

$$RR@5 = \frac{1}{3}$$

$$RR@10 = \frac{1}{3}$$

对于问题(b):

$$MAP@5 = \frac{0.805556 + 0.5 + 0.333333}{3} = 0.546296$$

$$MAP@10 = \frac{0.732937 + 0.375 + 0.306349}{3} = 0.471429$$

$$MRR@5 = \frac{1 + \frac{1}{2} + \frac{1}{3}}{3} = 0.611111$$

$$MRR@10 = \frac{1 + \frac{1}{2} + \frac{1}{3}}{3} = 0.611111$$

Exercise 2: Rank-based Evaluation Metrics, Precision@K, Recall@K, NDCG@K

练习 2：基于等级的评估指标、Precision@K、Recall@K、NDCG@K

Assume the following ranking for a given query (only results 1-10 are shown); see Table 1. The column ‘rank’ gives the rank of the document. The column ‘docID’ gives the document ID associated with the document at that rank. The column ‘graded relevance’ gives the relevance grade associated with the document (4 = perfect, 3 = excellent, 2 = good, 1 = fair, and 0 = bad). The column ‘binary relevance’ provides two values of relevance (1 = relevant and 0 = non-relevant). The assumption is that anything with a relevance grade of ‘fair’ or better is relevant and that anything with a relevance grade of ‘bad’ is non-relevant.

Also, assume that this query has only 7 documents with a relevance grade of fair or better. All happen to be ranked within the top 10 in this given ranking.

Answer the questions below. P@K (Precision@K), R@K (Recall@K), and average precision (AP) assume binary relevance. For those metrics, use the ‘binary relevance’ column. DCG and NDCG assume graded relevance. For those metrics, use the ‘graded relevance’ column.

Table 1 Top-10 ranking result of a system in response to a query.

rank	docID	graded relevance	binary relevance
1	51	4	1
2	501	1	1
3	21	0	0
4	75	3	1
5	321	4	1
6	38	1	1
7	521	0	0
8	412	1	1
9	331	0	0
10	101	2	1

(a) Compute P@5 and P@10.

$$P@5 = \frac{4}{5} = 0.8$$

$$P@10 = \frac{7}{10} = 0.7$$

(b) Compute R@5 and R@10.

$$R@5 = \frac{4}{7}$$

$$R@10 = \frac{7}{7} = 1$$

(c) Provide an example ranking for this query that maximizes P@5.

Example：1 2 4 5 6 3 7 8 9 10

(d) Provide an example ranking for this query that maximizes P@10.

Example：1 2 4 5 6 8 10 3 7 9

(e) Provide an example ranking for this query that maximizes R@5.

Example：1 2 4 5 6 3 7 8 9 10

(f) Provide an example ranking for this query that maximizes R@10.

Example：1 2 4 5 6 8 10 3 7 9

(g) You have reason to believe that the users of this system will want to examine every relevant document for a given query. In other words, you have recent to believe that users want perfect recall. You want to evaluate based on P@K. Is there a query-specific method for setting the value of K that would be particularly appropriate in this scenario? What is it? (**Hint**: there is an evaluation metric called R-Precision, which we did not talk about in the lectures. Your answer should be related to R-Precision. Wikipedia/Google might help.)

答：R-Precision：若在前 R 个检索文档中有 r 个相关（relevant）文档，则

$$\text{R-Precision} = \frac{r}{R}$$

所以可设置 K = 7（本题共 7 个相关文档），使得 R-Precision 越大越好。

(h) Compute average precision (AP). What are the difference between AP and MAP (Mean Average precision)?

$$AP = \frac{1+1+\frac{3}{4}+\frac{4}{5}+\frac{5}{6}+\frac{6}{8}+\frac{7}{10}}{7} = 0.833333$$

区别：AP 是求单个 Query 的平均 Precision（Average of P@K）， MAP 则是求多个 Query 的 AP 的平均值。

(i) Provide an example ranking for this query that maximizes average precision (AP).

Example：1 2 4 5 6 8 10 3 7 9

(j) Compute DCG_5 (i.e., the discounted cumulative gain at rank 5).

使用公式为：

$$\mathrm{DCG_p} = \sum_{i=1}^{p} \frac{rel_i}{\log_2(i+1)} = rel_1 + \sum_{i=2}^{p} \frac{rel_i}{\log_2(i+1)}$$

所以计算代码如下：

```
import numpy as np
rel = [4.0, 1.0, 0.0, 3.0, 4.0, 1.0, 0.0, 1.0, 0.0, 2.0]
DCG_5 = rel[0]
for i in range (1,5):
    DCG_5 = DCG_5 + rel[i]/np.log2(i+2)
print("%.6f" %DCG_5)
```

```
In [2]: import numpy as np
        rel = [4.0, 1.0, 0.0, 3.0, 4.0, 1.0, 0.0, 1.0, 0.0, 2.0]
        DCG_5 = rel[0]
        for i in range (1,5):
            DCG_5 = DCG_5 + rel[i]/np.log2(i+2)
        print("%.6f" %DCG_5)

        7.470371
```

$$DCG_5 = 7.470371$$

(k) $NDCG_5$ is given by

$$NDCG_5 = \frac{DCG_5}{IDCG_5}$$

where $IDCG_5$ is the DCG_5 associated with the *ideal* top-5 ranking associated with this query. Computing $NDCG_5$ requires three steps.

(i) What is the *ideal* top-5 ranking associated with this query (notice that the query has 2 *perfect* documents, 1 *excellent* document, 1 *good* document, 3 *fair* documents, and the rest of the documents are *bad*)?

(ii) $IDCG_5$ is the DCG_5 associated with the *ideal* ranking. Compute $IDCG_5$. (**Hint:** compute DCG_5 for your ranking proposed in part (i).)

(iii) Compute $NDCG_5$ using the formula above.

计算代码如下：

```
import numpy as np
rel = [4.0, 1.0, 0.0, 3.0, 4.0, 1.0, 0.0, 1.0, 0.0, 2.0]
DCG_5 = rel[0]
for i in range (1,5):
    DCG_5 = DCG_5 + rel[i]/np.log2(i+2)

rel.sort(reverse=True)

IDCG_5 = rel[0]
for i in range (1,5):
    IDCG_5 = IDCG_5 + rel[i]/np.log2(i+2)
```

```
NDCG_5 = DCG_5/IDCG_5
print("%.6f" %NDCG_5)
```

In [3]:
```
import numpy as np
rel = [4.0, 1.0, 0.0, 3.0, 4.0, 1.0, 0.0, 1.0, 0.0, 2.0]
DCG_5 = rel[0]
for i in range (1,5):
    DCG_5 = DCG_5 + rel[i]/np.log2(i+2)

rel.sort(reverse=True)

IDCG_5 = rel[0]
for i in range (1,5):
    IDCG_5 = IDCG_5 + rel[i]/np.log2(i+2)

NDCG_5 = DCG_5/IDCG_5
print("%.6f" %NDCG_5)
```

```
0.805698
```

$$NDCG_5 = 0.805698$$

(l) Are there other evaluation metrics to be used to evaluate the performance of the rankings in the table? What are the evaluation scores obtained by these metrics?

F-Measure：

$$F = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} = \frac{2 \cdot 0.7 \cdot 1}{0.7 + 1} = 0.823529$$

Exercise 3: Precision-Recall Curves

练习 3：精确召回曲线

A Precision-Recall (PR) curve expresses precision as a function of recall. Usually, a PR-curve is computed for each query in the evaluation set and then averaged. For simplicity, the goal in this question is to draw a PR-curve for a *single* query. Draw the PR-curve associated with the ranking in Exercise 2 (same query, same results). (**Hint:** Your PR curve should always go down with increasing levels of recall.)

Precision-Recall (PR) 曲线将精度表示为召回率的函数。 通常，为评估集中的每个查询计算 PR 曲线，然后取平均值。 为简单起见，此问题的目标是为单个查询绘制 PR 曲线。 绘制与练习 2 中的排名相关的 PR 曲线（相同的查询，相同的结果）。 （提示：你的 PR 曲线应该随着召回水平的提高而下降。）

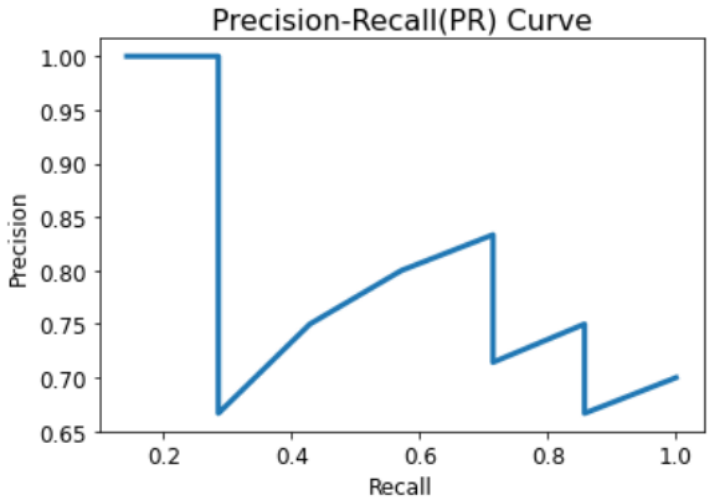
画图代码如下：

```
import matplotlib.pyplot as plt

#横纵坐标数据（Recall & Precision ）
precision = [1, 1, 2/3, 3/4, 4/5, 5/6, 5/7, 6/8, 6/9, 7/10]
```

```
recall = [1/7, 2/7, 2/7, 3/7, 4/7, 5/7, 5/7, 6/7, 6/7, 1]

#画图参数设置
plt.title("Precision-Recall(PR) Curve", fontsize=16)
plt.xlabel("Recall", fontsize=12)
plt.ylabel("Precision", fontsize=12)
plt.tick_params(axis='both', labelsize=12)
plt.plot(recall, precision, linewidth=3)
plt.show()
```



Exercise 4: Other Evaluation Metrics

练习 4：其他评估指标

Except the metrics we have in our lecture slides, are there other evaluation metrics that can be used to evaluate the performance of specific tasks in data mining? What are the tasks and how do to compute such evaluation metrics? (**Hint:** Use the internet to find your answers.)

（1）Kendall tau Distance

可用作衡量搜索结果之间相似性的指标。 例如，可以比较Google和Bing（针对同一查询）产生的前10名结果的接近程度。即两个排序间，评价存在分歧的对的数量。具体定义如下：

$$K(\tau_1,\tau_2)=|\{(i,j):i<j,(\tau_1(i)<\tau_1(j)\wedge\tau_2(i)>\tau_2(j))\vee(\tau_1(i)>\tau_1(j)\wedge\tau_2(i)<\tau_2(j))\}|$$

其中$\tau_1(i)$和$\tau_2(i)$分别为元素i在两个排序中的序位，如果两个排序完全一样。则Kendall tau distance为0。否则，若两个排序完全相反，则为$\frac{n(n-1)}{2}$。通常 Kendall tau distance都会通过除以$\frac{n(n-1)}{2}$来归一化。

（2）Spearman’s ρ

其基本思想类似于Kendall tau distance：比较两个排序（通常一个是理想排序）的（排序值的）皮尔逊相关系数。

比如在推荐中，一个推荐排序列表采用物品实际的评分值（用户实际的偏好程度）排序。一个是你的模型对物品的实际排序。s_{ij}^* 表示你模型预测中，物品 j 在用户 i 的推荐列表上的排序位置；y_{ij}^* 表示按实际用户 i 对物品的评分来排序时物品 j 在 i 的推荐列表上的排序位置。$\bar{s}^*$ 表示 s_{ij}^* 的平均值；$\bar{y}^*$ 表示 y_{ij}^* 的平均值.。则：

$$\frac{\sum_{(i,j)\in\Omega^{\text{test}}}\left(s_{ij}^*-\bar{s}^*\right)\left(y_{ij}^*-\bar{y}^*\right)}{\sqrt{\sum_{(i,j)\in\Omega^{\text{test}}}\left(s_{ij}^*-\bar{s}^*\right)^2}\sqrt{\sum_{(i,j)\in\Omega^{\text{test}}}\left(y_{ij}^*-\bar{y}^*\right)^2}}$$