机器学习与数据挖掘

Homework 2: Evaluation Metrics

中山大学计算机学院　计算机科学与技术

19335174　施天予

一、Exercise 1: Rank-based Evaluation Metrics, 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.

query	AP@5	AP@10	RR@5	RR@10
1	0.8056	0.7329	1	1
2	0.5	0.375	0.5	0.5
3	0.3333	0.3063	0.3333	0.3333

(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.

MAP@5	MAP@10	MRR@5	MRR@10
0.5463	0.4714	0.6111	0.6111

二、Exercise 2: Rank-based Evaluation Metrics, 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.

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

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

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

P@5	**P@10**
0.8	0.7

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

R@5	**R@10**
0.5714	1

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

$$P@5_{max} = 1$$

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

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

$$P@10_{max} = 0.7$$

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

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

$$R@5_{max} = 0.7143$$

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

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

$$R@10_{max} = 1$$

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

(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.)

R-Precision is the precision after R documents have been retrieved, where R is the number of. relevant documents for the topic. It de-emphasizes the exact ranking of the retrieved relevant. documents, which can be particularly useful in

TREC where there are large numbers of relevant documents.

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

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

$$AP = \frac{1 + 1 + 0.75 + 0.8 + 0.8333 + 0.75 + 0.7}{7} = 0.8333$$

AP 是对一个查询的平均，MAP 是对多个查询的 AP 取平均值

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

$$AP_{max} = 1$$

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

(j) Compute DCG5 (i.e., the discounted cumulative gain at rank 5).

$$DCG_5 = \sum_{i=1}^{5} \frac{rel_i}{\log_2(i+1)} = 4 + 0.6309 + 0 + 1.2920 + 1.5474 = 7.4703$$

(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)?

rank	**docID**	**graded relevance**
1	51	4
5	321	4
4	75	3
10	101	2
2	501	1

(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).)

$$IDCG_5 = \sum_{i=1}^{5} \frac{rel_i}{\log_2(i+1)} = 4 + 2.5237 + 1.5 + 0.8614 + 0.3869 = 9.272$$

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

$$NDCG_5 = \frac{DCG_5}{IDCG_5} = \frac{7.4703}{9.272} = 0.8057$$

(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?

RR@10	**F1-Score@10**	**CG@10**
1	0.8235	16

三、Exercise 3

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.)

使用全部数据，发现无法单调下降

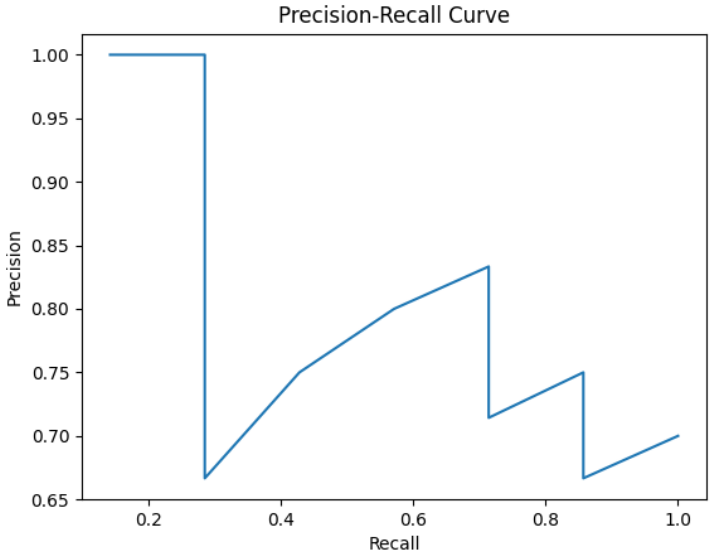

图 2: 使用全部数据的 PR-curve

使用 rank=1，4，7，10 的数据，曲线单调下降

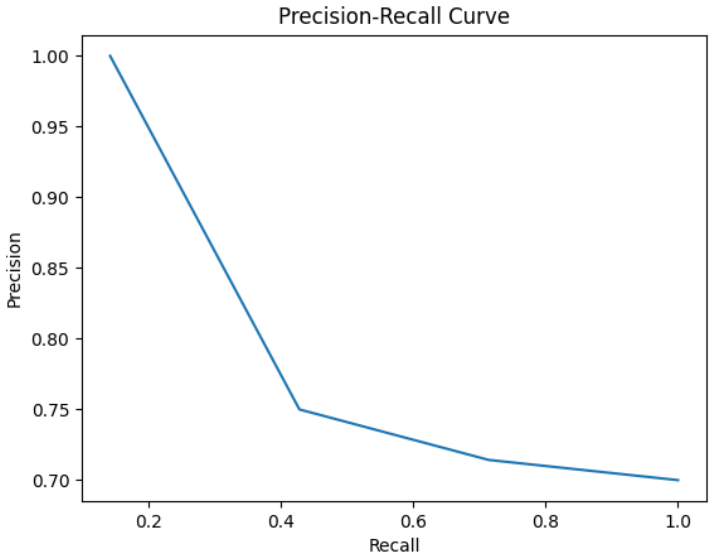

图 3: 使用全部数据的 PR-curve

四、Exercise 4

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

比较两个排序之间，评价存在分歧的数量。

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

其中 $\tau_1(i)$ 和 $\tau_2(i)$ 分别表示元素 i 在两个排序中的位置。如果两个排序完全一样，那么 Kendall tau distance 为 0；如果完全相反，那么为 $n(n-1)/2$；通常该距离都会除以 $n(n-1)/2$ 来归一化。

2. Spearman' s

基本思想类似 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：

$$\frac{\sum_{(i,j)\in\Omega^{test}}(S^*_{ij} - \bar{s}^*)(y^*_{ij} - \bar{y}^*)}{\sqrt{\sum_{(i,j)\in\Omega^{test}}(S^*_{ij} - \bar{s}^*)^2}\ \sqrt{\sum_{(i,j)\in\Omega^{test}}(y^*_{ij} - \bar{y}^*)}}$$

其中 s^*_{ij} 表示模型预测中，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；$\bar{s}^*$ 是 s^*_{ij} 的平均值；$\bar{y}^*$ 是 y^*_{ij} 的平均值。

3. AUC（Area under ROC curve）

AUC 的物理意义为任取一对例和负例，正例得分大于负例得分的几率，AUC 越大，代表方法效果越好。（AUC 的值通常介于 0.5-1）

4. Matthews correlation coefficient

MC 实质上是观察到的和预测的二元分类之间的相关系数; 它返回介于-1 和 +1 之间的值。系数 +1 表示完美预测，0 表示不比随机预测好，-1 表示预测和观察之间的完全不一致。统计数据也称为 phi 系数。当俩个类别具有非常不同的大小时，其他度量（例如正确预测的比例（也称为准确性））无用。例如，将每个对象分配给较大的集合可以实现高比例的正确预测，但通常不是有用的分类。可以使用以下公式直接从混淆矩阵计算 MCC：

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$