Homework 2: Evaluation Metrics

Student ID:21307174 Student Name:刘俊杰

Date: 2023.10.16

Lectured by: Shangsong Liang
Machine Learning and Data Mining Course
Sun Yat-sen University

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.

Answer:

Ranking 1:

AP@5=$(\frac{1}{1}+\frac{2}{3}+\frac{3}{4})/3=\frac{29}{36}$

AP@10=$(\frac{1}{1}+\frac{2}{3}+\frac{3}{4}+\frac{4}{6}+\frac{5}{7})/5=\frac{168+112+126+112+120}{168*5}=\frac{638}{840}=\frac{319}{420}$

RR@5=1

RR@10=1

Ranking2:

AP@5=0

AP@10=$(\frac{1}{8}+\frac{2}{9})/2=\frac{25}{144}$

RR@5=0

RR@10=$\frac{1}{8}$

Ranking3:

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

AP@10=($\frac{1}{5}+\frac{2}{8}+\frac{3}{9}$)/3=$\frac{94}{360}$=$\frac{47}{180}$

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

RR@10=$\frac{1}{5}$

(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.

Answer:

MAP@5 (Mean Average Precision @5)=($\frac{29}{36}$ + 0 + $\frac{1}{5}$)/3=$\frac{181}{480}$

MAP@10=($\frac{319}{420}+\frac{25}{144}+\frac{47}{180}$)/3=0.39808201058201058201058201058201

MRR@5 (Mean Reciprocal Rank score @5)=(1+0+$\frac{1}{5}$)/3=$\frac{2}{5}$

MRR@10 =(1+$\frac{1}{8}+\frac{1}{5}$)/3=$\frac{53}{120}$

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.

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.

Answer:

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

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

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

Answer:

R@5 = 1

R@10 = 1

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

Answer:

Rank: 1 2 3 4 5 6 7 8 9 10

docID: 51 501 75 321 38 421 101 21 521 331

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

Answer:

Rank: 1 2 3 4 5 6 7 8 9 10

docID: 51 501 75 321 38 421 101 21 521 331

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

Answer:

Rank: 1 2 3 4 5 6 7 8 9 10

docID: 51 501 75 321 38 421 101 21 521 331

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

Answer:

Rank: 1 2 3 4 5 6 7 8 9 10

docID: 51 501 75 321 38 421 101 21 521 331

(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.)

Answer:

有

在信息检索和文档检测领域，RP Precision 是用于评估模型检测相关文档的性能的关键指标。这些指标用于衡量模型在给定召回率（Recall）水平下的精确度，即模型在检测到相关文档时的准确性。

以下是如何计算和理解与文档检测相关的精确率（RP Precision）：

相关文档（相关的正样本）：这些是模型正确地检测到的实际相关文档的数量。

不相关文档（相关的负样本）：这些是模型错误地将不相关文档视为相关文档的数量。

RP Precision = 相关文档 / (相关文档 + 不相关文档)

较高的 RP Precision 表示模型在检测相关文档时更准确。

此题的 RP Precision=7/10=0.7

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

Answer:

AP@5=$(\frac{1}{1}+\frac{2}{2}+\frac{3}{4}+\frac{4}{5})/4=\frac{71}{80}$

AP@10=$(\frac{1}{1}+\frac{2}{2}+\frac{3}{4}+\frac{4}{5}+\frac{5}{6}+\frac{6}{8}+\frac{7}{10})/7$=0.833333333333333333333333333333333

AP 和 MAP 之间的区别：

AP 是针对单个查询或信息检索任务计算的。

AP 提供了一个指示系统在单个查询中检索相关项目的能力的指标。

MAP：

MAP 是在多个查询或信息检索任务中计算的 AP 值的平均值。

它用于评估信息检索系统在一组查询中的整体性能。

MAP 提供了对系统性能的更全面评估，考虑了系统在多个任务中的一致性。

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

Answer:

Rank: 1 2 3 4 5 6 7 8 9 10
docID: 51 501 75 321 38 421 101 21 521 331

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

Answer:

DCG_5=$4+\frac{1}{log_2 2}+\frac{3}{log_2 4}+\frac{4}{log_2 5}$ =8.222706232293572

(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.

Answer:

$DCG_5 = 4 + \frac{1}{log_2 2} + \frac{3}{log_2 4} + \frac{4}{log_2 5} = 8.222706232293572$

$IDCG_5 = 4 + \frac{4}{log_2 2} + \frac{3}{log_2 3} + \frac{2}{log_2 4} + \frac{1}{log_2 5} = 11.323465818787765$

$NDCG_5 = \frac{DCG_5}{IDCG_5} = 0.7261651480106516$

(k) 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?

Answer:

3. F measure

$$F_\beta = (1 + \beta^2) \cdot \frac{precision \cdot recall}{(\beta^2 \cdot precision) + recall}$$

We normally use a balanced F1 measure with β=1.

$$F_1 = 2 \cdot \frac{precision \cdot recall}{precision + recall}$$

可以使用F度量方法

在这里我们使用F1度量方法来计算本题

此题中precision=0.7 recall=1

即F1=1.4/1.7=0.8235294117647058

Exercise 3: Precision-Recall Curves

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.)

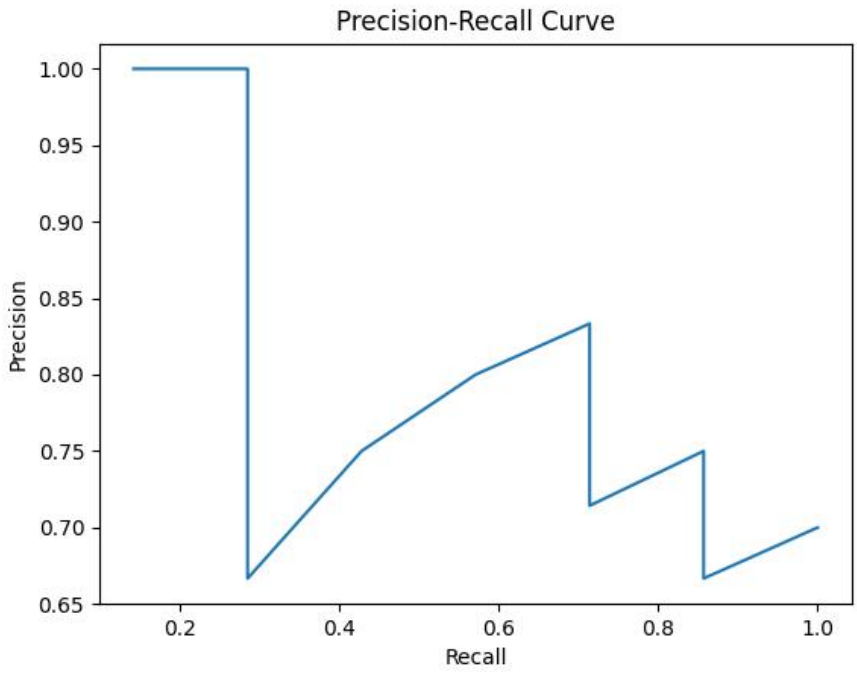
Answer:

Precision = TP/(TP+FP)，即被检索出的文档中真正相关的文档所占的比例。

Recall = TP/(TP+FN)，即真正相关的文档中被成功召回（检索）的文档所占的比例。

```python
import matplotlib.pyplot as plt
def calculate_P(K,relevant):
    relevant_count = 0
    for i in range(K):
        if relevant[i] == 1:
            relevant_count += 1
    return relevant_count/K
def calculate_R(K,relevant):
    relevant_count_k = 0
    relevant_count = 0
    for i in range(len(relevant)):
        if relevant[i]==1:
            relevant_count += 1
            if i<K:
                relevant_count_k += 1
    return relevant_count_k/relevant_count

relevant = [1,1,0,1,1,1,0,1,0,1]
P=[]
R=[]
for i in range(len(relevant)):
    P.append(calculate_P(i+1,relevant))
    R.append(calculate_R(i+1,relevant))
plt.xlabel('Recall')
plt.ylabel('Precision')
plt.title('Precision-Recall Curve')
plt.plot(R,P)
plt.show()
```



Exercise 4: Other Evaluation Metrics

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.)

Answer:

1.回归任务：

平均绝对百分比误差（MAPE）： MAPE = (1/n) * Σ(|(真实值 - 预测值) / 真实值|) * 100 这里，n 是样本数，|x| 表示 x 的绝对值。

Huber损失： Huber损失通常是一个带有阈值的损失函数，适用于多种问题。具体计算方式取决于阈值和数据。

Theil's U 统计量： Theil's U 统计量是一个统计度量，衡量了预测值和实际值之间的相对误差。具体计算方法可能因数据的性质而异。

2.多类分类任务：

多类混淆矩阵：多类混淆矩阵是一个方阵，其行和列对应于真实类别和预测类别。您可以使用混淆矩阵计算精确率、召回率、F1分数等。

加权F1分数：对于加权F1分数，您需要计算每个类别的F1分数，然后对其进行加权。具体加权方式取决于类别的重要性。

3.时间序列分析：

均方误差（MSE）： MSE = (1/n) * Σ(真实值 - 预测值)² 这里，n 是时间序列数据点的数量。

平均绝对误差（MAE）： MAE = (1/n) * Σ|真实值 - 预测值| 这里，n 是时间序列数据点的数量。

自相关性和偏自相关性：这些是时间序列分析中的统计工具，需要使用专门的时间序列分析方法和工具库（例如，ARIMA模型）来计算。

4.异常检测：

Kappa统计量： Kappa统计量通常需要混淆矩阵，并计算观察到的一致性与预期一致性之间的差异。具体计算方式依赖于混淆矩阵和问题。

K-L散度： K-L散度是两个概率分布之间的相似性度量，通常用于计算实际分布与模型预测分布之间的差异。计算方法取决于具体的概率分布。