19335030_陈至雪 _HW2

# Exercise 1

| **Ranking 1** | **d1** | **d2** | **d3** | **d4** | **d5** | **d6** | **d7** | **d8** | **d9** |
|---|---|---|---|---|---|---|---|---|---|
| recall | 0.17 | 0.17 | 0.33 | 0.5 | 0.5 | 0.67 | 0.83 | 0.83 | 0.83 |
| precision | 1.0 | 0.5 | 0.67 | 0.75 | 0.6 | 0.67 | 0.71 | 0.625 | 0.56 |
| **Ranking 2** | d3 | **d8** | d7 | d1 | d2 | d4 | d5 | **d9** | d10 |
| recall | 0.0 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 1.0 | 1.0 |
| precision | 0.0 | 0.5 | 0.33 | 0.25 | 0.2 | 0.167 | 0.14 | 0.25 | 0.22 |
| **Ranking 3** | d7 | d6 | **d5** | d3 | d2 | d1 | **d9** | d10 | d4 |
| recall | 0.0 | 0.0 | 0.33 | 0.33 | 0.33 | 0.33 | 0.67 | 0.67 | 0.67 |
| precision | 0.0 | 0.0 | 0.33 | 0.25 | 0.2 | 0.167 | 0.286 | 0.25 | 0.22 |

(a)

| | **AP@5** | **AP@10** | **RR@5** | **RR@10** |
|---|---|---|---|---|
| Ranking 1 | 0.704 | 0.6685 | 1.0 | 1.0 |
| Ranking 2 | 0.256 | 0.2257 | 0.5 | 0.5 |
| Ranking 3 | 0.156 | 0.2003 | 0.33 | 0.33 |

(b)

| **MAP@5** | **MAP@10** | **MRR@5** | **MRR@10** |
|---|---|---|---|
| 0.372 | 0.3648 | 0.61 | 0.61 |

# Exercise 2

| **Ranking** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** |
|---|---|---|---|---|---|---|---|---|---|---|
| recall | 0.14 | 0.29 | 0.29 | 0.43 | 0.57 | 0.71 | 0.71 | 0.86 | 0.86 | 1.0 |
| precision | 1.0 | 1.0 | 0.66 | 0.75 | 0.8 | 0.83 | 0.71 | 0.75 | 0.67 | 0.7 |

(a)

| P@5 | P@10 |
|---|---|
| 0.8 | 0.7 |

(b)

| R@5 | R@10 |
|---|---|
| 0.57 | 1.0 |

(c)

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

(d)

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

(e)

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

(f)

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

(g)

R-precision requires knowing all documents that are relevant to a query. The number of relevant documents R, is used as the cutoff for calculation, and this varies from query to query.

Set K = R.

(h)

$AP = \frac{(1.0+1.0+0.75+0.8+0.83+0.75+0.7)}{7} = 0.83285$

AP 是对一个排序结果中的相关文档的精度加和求平均。反应了此排序的精度。

MAP是对同个查询主题的多个排序的AP加和平均。反应了多个算法对此查询主题的查询精度。

(i)

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

(j)

| rank | discounted gain | DCG |
|---|---|---|
| 1 | 4 | 4 |
| 2 | 0.629 | 4.629 |
| 3 | 0 | 4.629 |
| 4 | 1.29 | 5.919 |
| 5 | 1.54 | 7.459 |
| 6 | 0.36 | 7.819 |
| 7 | 0 | 7.819 |
| 8 | 0.32 | 8.139 |
| 9 | 0 | 8.139 |
| 10 | 0.58 | 8.719 |

So DCG5 = 7.459

(k)

(i)

| rank | docID |
|---|---|
| 1 | 51 |
| 2 | 321 |
| 3 | 75 |
| 4 | 101 |
| 5 | 501 |

(ii)

| rank | docID | discounted gain | DCG |
|---|---|---|---|
| 1 | 51 | 4 | 4 |
| 2 | 321 | 1.156 | 5.156 |
| 3 | 75 | 1.5 | 6.656 |
| 4 | 101 | 0.862 | 7.518 |
| 5 | 501 | 0.386 | 7.904 |

IDCG5 = 7.904

(iii)

$$NDCG_5 = \frac{DCG_5}{IDCG_5} = \frac{7.459}{7.904} = 0.9437$$

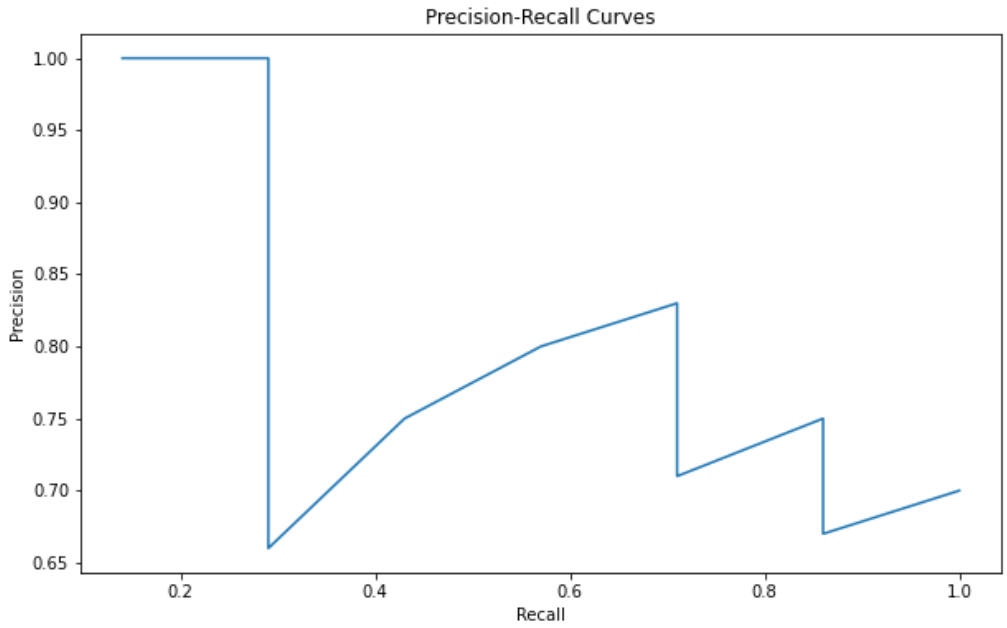
(l)

Mean Reciprocal Rank.

$$The\ Reciprocal\ Rank\ score = \frac{1}{1} = 1$$

# Exercise 3

根据 Exercise 2 计算出的数据绘制 **Precision-Recall Curves**， 得到下图：




# Exercise 4

**分类问题**:

- 铰链损失
  - 铰链损失（Hinge loss）一般用来使“边缘最大化”（maximal margin）。
    铰链损失最开始出现在二分类问题中，假设正样本被标记为1，负样本被标记为-1，y是真实值，w是预测值，则铰链损失定义为：

$$L_{Hinge}(w, y) = max\{1 - wy, 0\} = |1 - wy|_+$$

- kappa系数
  - kappa系数用来衡量两种标注结果的吻合程度，标注指的是把N个样本标注为C个互斥类别。计算公式为

$$\kappa = \frac{p_o - p_e}{1 - p_e} = 1 - \frac{1 - p_o}{1 - p_e}$$

    两种标注结果完全相符时，K=1，越不相符其值越小，甚至是负的。

- 混淆矩阵

- 又被称为错误矩阵，通过它可以直观地观察到算法的效果。它的每一列是样本的预测分类，每一行是样本的真实分类（反过来也可以），顾名思义，它反映了分类结果的混淆程度。混淆矩阵i行j列的原始是原本是类别i却被分为类别j的样本个数，计算完之后还可以对之进行可视化。

**拟合问题**

- 平均绝对误差
  - 平均绝对误差MAE（Mean Absolute Error）又被称为l1范数损失（l1-norm loss）：

$$\text{MAE}(y, \hat{y}) = \frac{1}{n_{\text{samples}}} \sum_{i=1}^{n_{\text{samples}}} |y_i - \hat{y}_i|$$

- 决定系数
  - 又被称为R2分数：

$$R^2(y, \hat{y}) = 1 - \frac{\sum_{i=1}^{n_{\text{samples}}} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n_{\text{samples}}} (y_i - \bar{y})^2}$$

$$\text{其中} \bar{y} = \frac{1}{n_{\text{samples}}} \sum_{i=1}^{n_{\text{samples}}} y_i$$

当R2越接近1时，表示相关的方程式参考价值越高；相反，越接近0时，表示参考价值越低。

**聚类的评价指标**

- 兰德指数
  - 兰德指数（Rand index）需要给定实际类别信息C，假设K是聚类结果，a表示在C与K中都是同类别的元素对数，b表示在C与K中都是不同类别的元素对数，则兰德指数为：

$$\text{RI} = \frac{a+b}{C_2^{n_{\text{samples}}}}$$

$$\text{其中} C_2^{n_{\text{samples}}} \text{数据集中可以组成的总元素对数}$$

RI取值范围为[0,1]，值越大意味着聚类结果与真实情况越吻合。

  - 为了实现“在聚类结果随机产生的情况下，指标应该接近零”，调整兰德系数（Adjusted rand index）被提出，它具有更高的区分度：

$$\mathrm{ARI} = \frac{\mathrm{RI} - E[\mathrm{RI}]}{\max(\mathrm{RI}) - E[\mathrm{RI}]}$$

ARI取值范围为[−1,1]，值越大意味着聚类结果与真实情况越吻合。从广义的角度来讲，ARI衡量的是两个数据分布的吻合程度。

- 轮廓系数
  - 适用于实际类别信息未知的情况。对于单个样本，设a是与它同类别中其他样本的平均距离，b是与它距离最近不同类别中样本的平均距离，轮廓系数为：

$$s = \frac{b - a}{max\{a, b\}}$$

还原 S，有

$$S(i) = \begin{cases} 1 - \frac{a(i)}{b(i)} & a(i) < b(i) \\ 0 & a(i) = b(i) \\ \frac{b(i)}{a(i)} - 1 & a(i) > b(i) \end{cases}$$

可以发现，当a(i)<b(i)时，即类内的距离小于类间距离，则聚类结果更紧凑。S的值会趋近于1。越趋近于1代表轮廓越明显。

相反，当a(i)>b(i)时，类内的距离大于类间距离，说明聚类的结果很松散。S的值会趋近于-1，越趋近于-1则聚类的效果越差。

对于一个样本集合，它的轮廓系数是所有样本轮廓系数的平均值。
由此可得，轮廓系数取值范围是[−1,1]，**同类别样本越距离相近且不同类别样本距离越远，分数越高。**