CSCI4360/6360 - HW4

Instructor: Dr. Ninghao Liu (ninghao.liu@uga.edu)

April 9, 2024

– Upload two files to eLC (failing to do so will receive 20% penalty):
1) a scanned handwritten solution or typed pdf file named "*YourID_HW4.pdf*" containing your answer to each question and code running results;
2) a zip file named "*YourID_HW4PQ.zip*" containing your programs for Question 1.

– Due Date: April 23, 2024

1 Outlier Detection (40pts)

1.1 Parametric Methods (15pts)

In this problem, we will detect outliers, using statistical methods, from a dataset with 600 instances, where each instance has 2 features. Assume that the data is generated by a Gaussian distribution. Return the top 3 outliers (report their coordinates), and the mean vector and the covariance matrix. The dataset could be found in the file "data_1.npy", which could be opened as below.

```
import numpy as np

with open('data_1.npy', 'rb') as f:
    X = np.load(f)
```

1.2 Isolation Trees (25pt)

In this problem, we will detect outliers using Isolation Forest on the "data_2.npy" dataset. The base codes could be found in "iForest.py". Complete the Isolation Forest algorithm in "iForest.py". Return the top 4 outliers (report their coordinates).

2 Recommender Systems (10pt × 2 = 20pt)

1) In a e-commerce platform with users and items, suppose the characteristics of items are available, and the historical records (purchasement of items) of users are available, briefly introduce how to build a content-based recommender system.
2) In a e-commerce platform with users and items, suppose we only have the historical ratings given by some users to some items, briefly introduce how to build a recommender system.

3 Recommendation Evaluation (10pt × 4 = 40pt)

Given the documents retrieval results as below:

The relevance scores of retrieved documents that are relevant with query 2 are given as shown in the figure.

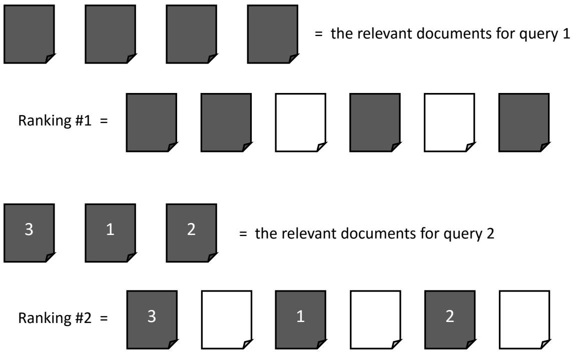
= the relevant documents for query 1
Ranking #1 =
3
1
2
= the relevant documents for query 2
Ranking #2 =
3
1
2

Figure 1: A toy example of documents retrieval

1) For query 1, calculate Precision@K for $K = 4, 5, 6$.
2) Calculate the MAP for the ranking system above.
3) For query 2, calculate the $NDCG_5$. Suppose $DCG_p = rel_1 + \sum_{i=2}^{p} \frac{rel_i}{log_2 i}$, where rel_i denotes the relevance score for the i^{th} returned document.
4) Explain when NDCG is more advantageous over Precision@K and MAP.