CSC4020 Assignment 4

Li Chenyi119010143

May 2021

1 PROBLEM

Given data set with 210 data points and 7 features, we need to assign clusters to it. The performance metrics include purity, rand index and normalized mutual information

2 EVALUATION METRICS

2.1 PURITY

$$Purity = \frac{1}{N}\sum_{i=1}^{k} max_j|c_i \cap t_j|$$

N is the number of data points. k is the number of clusters. c_i is the i th cluster. t_j is the j th class.

2.2 RAND INDEX

We first draw a matrix as follows. The horizontal axis is the class index, and the vertical axis is the cluster index. The value of matrix is the number of data points in i cluster and j class. For each row, we sum up the data points number with same cluster and get n_1, n_2 ...n_k. For each column, we sum up the data points number with same class, and get m_1, m_2 ...m_c.

In clustering problems, $TP = \sum_{i=1}^{k}\sum_{j=1}^{c}\binom{a_{ij}}{2}$, a_{ij} represent the number of data points in i cluster and j class.

$$TP + FN = \sum_{i=1}^{j}\binom{m_i}{2}$$

$$TN + FN + TP + FP = \binom{N}{2}$$

$$FN + TN = \binom{N}{2} - (TP + FP)$$

$$TN = (FN + TN) - (FN + TP) + TP$$

$$Randindex = \frac{TP + TN}{TP + TN + FP + FN}$$

2.3 NORMALIZED MUTUAL INFORMATION

$$NMI(\Omega, C) = \frac{2 \cdot I(\Omega, C)}{H(\Omega) \cdot H(C)}$$

$$I(\Omega, C) = \frac{1}{N} \sum_{i=1}^{k} \sum_{j=1}^{c} P(c_i \cap t_j) \log \frac{P(c_i \cap t_j)}{P(c_i)P(t_j)}$$

$$= \frac{1}{N} \sum_{i=1}^{k} \sum_{j=1}^{c} \frac{|c_i \cap t_j|}{N} \log \frac{N|c_i \cap t_j|}{|c_i||t_j}$$

$$H(C) = -\sum_{i=1}^{k} P(c_i) \log P(c_i)$$

$$H(\Omega) = -\sum_{j=1}^{c} P(t_j) \log P(t_j)$$

3 EXPERIMENTAL RESULTS

We set k to 3 clusters.

3.1 SENSITIVITY ANALYSIS

The k-means and soft k-means are very sensitive to the initialized value. Each time changing the initial value, the purity, mutual information and rand index will change.

3.2 K-MEANS

```
Each cluster has the following classes:

Cluster 0
[1.0, 2.0, 2.0, 2.0, 2.0, 2.0, 2.0, 2.0, 2.0, 2.0, 2.0, 2.0, 2.0, 2.0, 2.0,
For cluster 0, the maximum class 2.0 has 60 data points

Cluster 1
[1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 3.0, 3.0, 3.0,
For cluster 1, the maximum class 3.0 has 70 data points
```

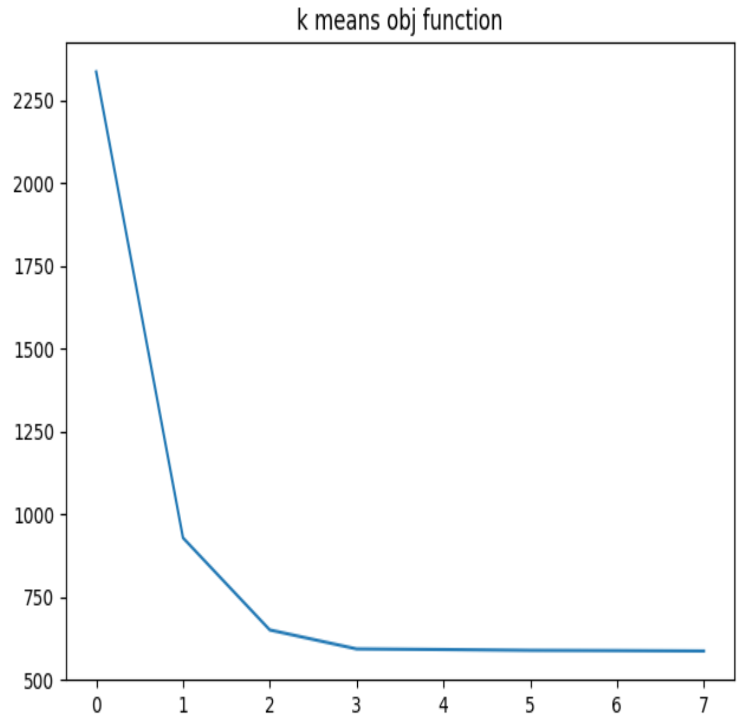

FIGURE 1: K-MEANS OBJECTIVE FUNCTION CURVE

```
Cluster 2
[1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0,
For cluster 2, the maximum class 1.0 has 57 data points

Purity: 0.8904761904761905
Rand index: 0.8713602187286398
NMI: 0.7100637577760452
k-means time cost: 0.4058678150177002 s
```

3.3 SOFT K MEANS

```
Cluster 0
[1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 3.0, 3.0, 3.0, 3.0, 3.0, 3.0,
For cluster 0, the maximum class 3.0 has 68 data points

Cluster 1
[1.0, 2.0, 2.0, 2.0, 2.0, 2.0, 2.0, 2.0, 2.0, 2.0, 2.0, 2.0, 2.0, 2.0, 2.0,
For cluster 1, the maximum class 2.0 has 60 data points

Cluster 2
```

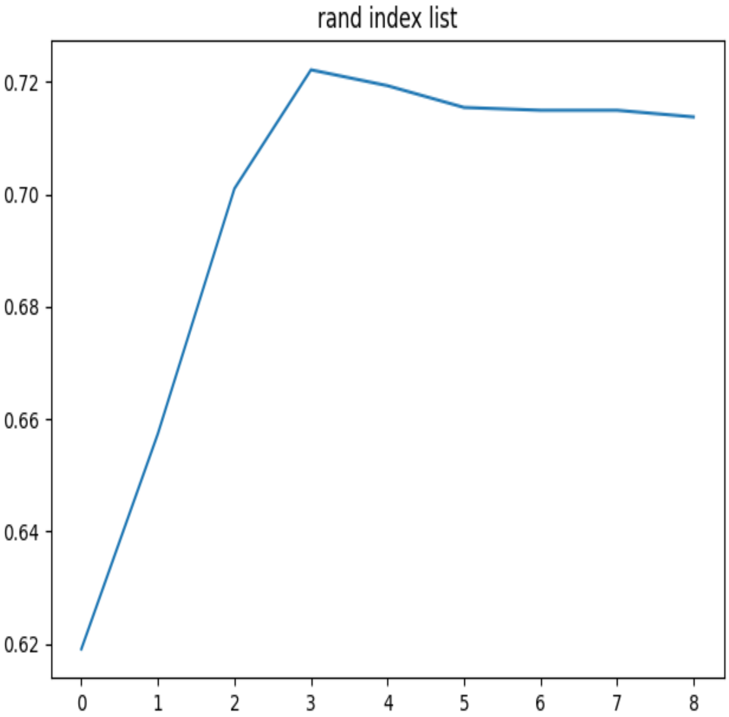

FIGURE 2: RAND INDEX

```
[1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0,
For cluster 2, the maximum class 1.0 has 60 data points

Purity:  0.8952380952380953
Rand index:  0.8743677375256322
NMI:  0.6949250270680581
soft-k-means time cost:  0.2184159755706787 s
```

3.4 EM ALGORITHM

```
Cluster 0
[1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0,
For cluster 0, the maximum class 2.0 has 58 data points

Cluster 1
[2.0, 2.0, 2.0, 2.0, 2.0, 2.0, 2.0, 2.0, 2.0, 2.0, 2.0, 2.0]
For cluster 1, the maximum class 2.0 has 12 data points

Cluster 2
[1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0,
For cluster 2, the maximum class 3.0 has 70 data points
```

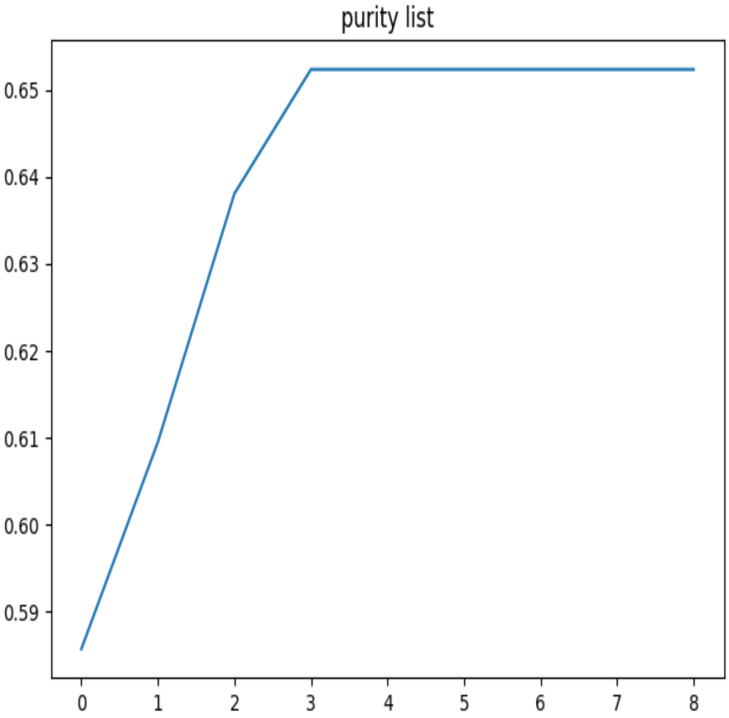

FIGURE 3: PURITY

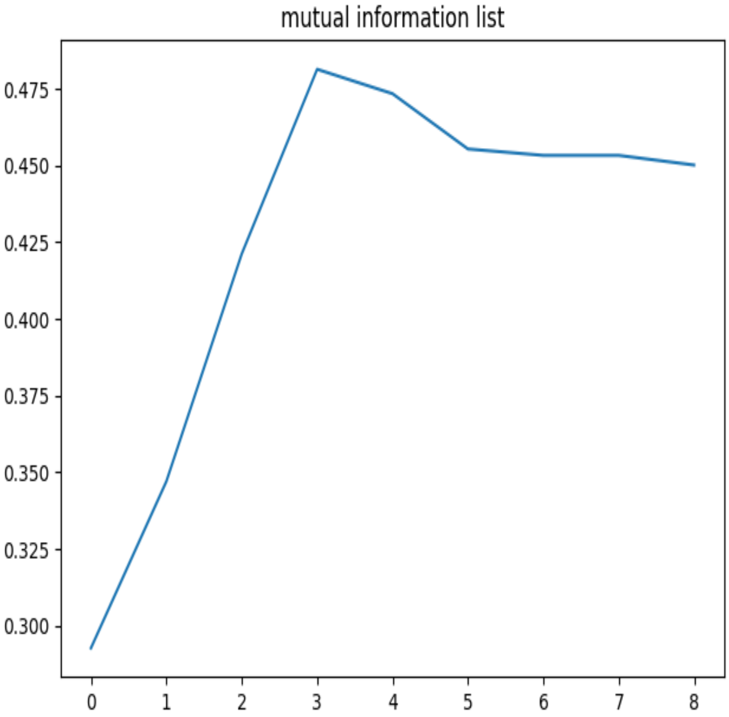

FIGURE 4: MUTUAL INFORMATION

```
Purity:  0.6666666666666667
Rand index:  0.7267714741398952
NMI:  0.5251223332076916
EM time cost:  1.55379319190979 s
```

3.5 NOTE

For accelerated k means, I could not understand the parameter r in each iterations. Also due to the time limit, I am not able to realize it.