THU-70250043, Pattern Recognition (Spring 2018) **Homework: 12**

Clustering Methods

Lecturer: Changshui Zhang zcs@mail.tsinghua.edu.cn

Student: XXX xxx@mails.tsinghua.edu.cn

Problem 1

K-means

1.1 The set D contains n samples. Divide D into c disjoint subsets $D_1, D_2, \cdots, D_c$. The mean value of the samples of the subset D_i is m_i. If D_i is an empty set, m_i is no definition, and the sum of the error squares is only related to the nonempty subsets.

$$J_e = \sum_{D_i \neq \phi} \sum_{x \in D_i} \|x - m_i\|^2$$

Assuming that n is greater than c, prove that minimization of J_e means that the division of the D does not contain nonempty subsets.

Problem 2

Programming

Test the clustering algorithms **K − means**, **hierarhical clustering** and **spectral clustering** with different parameters on MNIST dataset or subsets of it when the scale is too large for the algorithm involved.

To compare the effectiveness of different clustering methods, *Normalized mutual information*(NMI) are widely used as a measurement. NMI is defined as following:

$$NMI = \frac{\sum_{s=1}^{K} \sum_{t=1}^{K} n_{s,t} log(\frac{n n_{s,t}}{n_s n_t})}{\sqrt{(\sum_s n_s log \frac{n_s}{n})(\sum_t n_t log \frac{n_t}{n})}} \tag{1}$$

Where n is the number of data points, n_s and n_t denote the numbers of the data in class s and class t, $n_{s,t}$ denotes the number of data points in both class s and class t. For more details and other measurements, google "evaluation of clustering".

2.1 Give a brief analysis of time complexity of each algorithm mentioned above (of standard implementation). Estimate how many samples each algorithm can manage with a reasonable time cost.

(Optional) Can you verify your estimations with experiments? Can you speed it up further?

2.2 Consider each data set, and use the true number of classes as the number of clusters.

- With K-means, will the initial partition affect the clustering results? How can you solve this problem? And do J_e and NMI match? Show your experiment results.
- When hierarchical clustering is adopted, the choice of linkage method depends on the problem. Give an analysis of linkage method's effects with experiments, and which is better in the sense of NMI? For more linkage methods, refer to the linkage function's help in matlab.
- As above, give an experimental analysis of the choice of similarity graph and corresponding parameters. Which one is better?

2.3 In practice, we may not know the true number of clusters much. Can you give a strategy to identify the cluster number automatically for each algorithm? Show your results.

2.4 According to the above analysis, which method do you prefer? Why?