Assignment 3

October 17, 2016

1 Tasks

For each data set, we assume the number of cluster is known and denoted by c. You are required to implement the following clustering algorithms:

- the standard k-medoids algorithm in chapter 6.3.4 of [1], and you need to use ℓ_1-norm to calculate the distance between two instances, e.g., the ℓ_1-norm distance of X_i and X_j is:

$$Dist(X_i, X_j) = \|X_i - X_j\|_1 = \sum_{k=1}^{d} |X_{ik} - X_{jk}|, \tag{1}$$

  the X_{ik} is the k feature of the instance X_i.

- The spectral clustering algorithm in [2]. To make life easier, you can follow Step 1(b), Step 2(b) and Step 3 in [2], where $m = c$ in Step 3. After that, you will get a c-dimensional representant for each data point, and then apply k-mediods you realized before to clustering the new data. Again, you set k = c in k-mediods. There is a parameter in Step 1(b), which is the number of nearest neighbors n. You can try n = 3, 6 and 9.

Note that optimization algorithms for k-medoids can only find local optimum. So, each time you solve k-medoids, you need to run its optimization algorithm at least 10 times with different initializations, and use the solution with smallest objective value.
To compare different methods quantitatively, you are required to calculate the Purity and Gini index of Section 6.9.2 of the textbook [1]. In the report, you may present the final results as follows:

References

[1] Charu C. Aggarwal. *Data Mining: The Textbook*. Springer, 2015.

	k-medoids	Spectral(n=3)	Spectral(n=6)	Spectral(n=9)
Dataset1				
Dataset2				
...				

Table 1: Purity of different algorithms

	k-medoids	Spectral(n=3)	Spectral(n=6)	Spectral(n=9)
Dataset1				
Dataset2				
...				

Table 2: Gini index of different algorithms

[2] Wei Xu, Xin Liu, and Yihong Gong. Document clustering based on non-negative matrix factor-ization. *In Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 267C273, 2003.