Assignment 3

Lijun Zhang

October 20, 2015

1 Tasks

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

- The standard k-means algorithm, where $k = c$.
- Clustering by NMF, which can be found in [3]. Suppose you aim to decompose $X \in \mathbb{R}^{d\times n}$ as $U \times V^\top$. The sizes of U and V are $d \times c$ and $n \times c$, respectively.
- 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-means to clustering the new data. Again, you set $k = c$ in k-means.
  - 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-means and NMF can only find local optimum. So, each time you need to solve k-means or NMF, 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:

	k-means	NMF	Spectral ($n = 3$)	Spectral ($n = 6$)	Spectral ($n = 9$)
Data set 1					
Data set 2					
$\cdots$					

Table 1: Purity of different algorithms

	k-means	NMF	Spectral ($n = 3$)	Spectral ($n = 6$)	Spectral ($n = 9$)
Data set 1					
Data set 2					
$\cdots$					

Table 2: Gini index of different algorithms

References

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

[2] Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps and spectral techniques for embedding and clustering. In *Advances in Neural Information Processing Systems 14*, pages 585–591, 2002.

[3] Wei Xu, Xin Liu, and Yihong Gong. Document clustering based on non-negative matrix factorization. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 267–273, 2003.