IE529 Computational Assignment 1: Solutions

December 23, 2017

1 A scatter plot for each dataset

The scatter plots for each dataset are

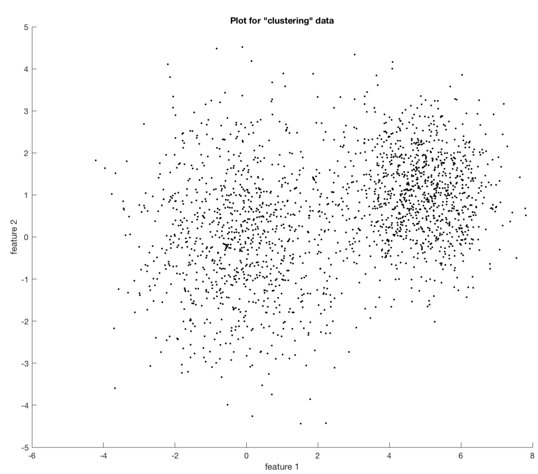

Figure 1: Plot for ”clustering” data

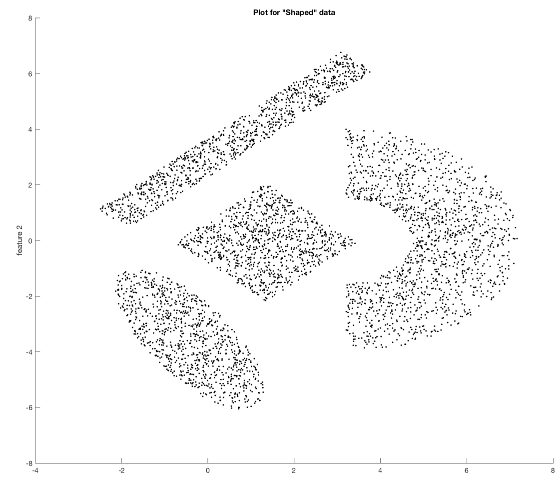

Figure 2: Plot for ”shaped” data

2 Results for K-means and Spectral Clustering

1. K-means output plots for best K.

   - $K = 2$ for mixture-of-gaussian dataset and $K = 4$ for ”shaped” dataset. From Fig.3 and Fig.4 below, we can conclude that K-means works for mixture-of-guassian data but doesn’t work well for ”shaped” data.

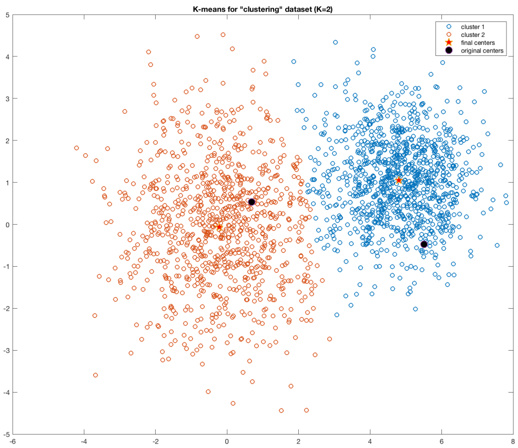

Figure 3: Clustering results for "clustering" data, K-means

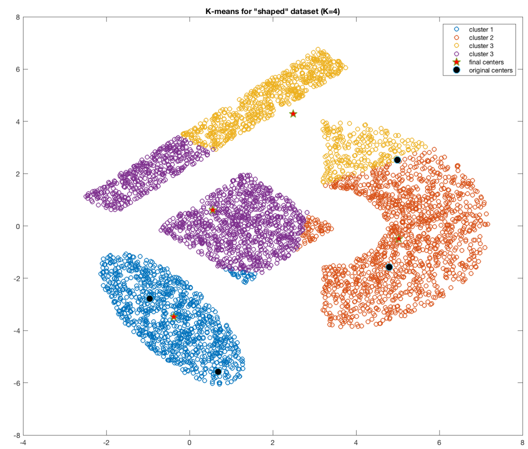

Figure 4: Clustering results for "shaped" data, K-means

Note: Result for the "shaped" dataset is not unique, it varys with different randomly selected initializations.

- D vs K plots for mixture-of-gaussian dataset and "shaped" dataset.
  - If the distance metric D is defined as distortion, which is the sum-of-squared distance:

$$Distortion = \sum_{j=1}^{K} \sum_{i=1}^{N} \parallel x_i - m_j \parallel_2^2 u_{ij}$$

where $u_{ij} = 1$ if $x_i \in C_j$, and $u_{ij} = 0$, otherwise; $m_j = \frac{\sum_{i=1}^{n} x_i u_{ij}}{|C_j|}$, $|C_j| =$ number of points in C_j

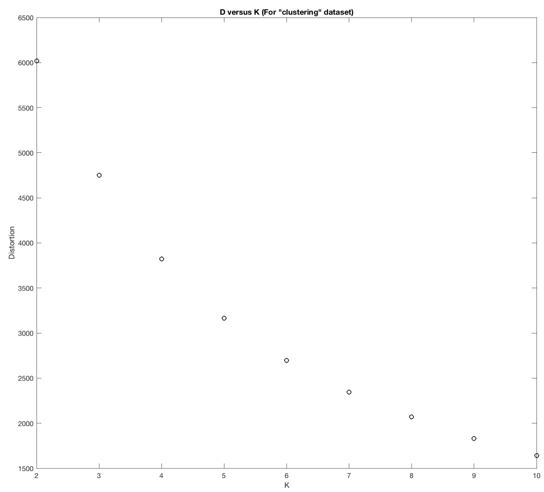

Figure 5: D(distortion) vs K results for ”clustering” data, K-means

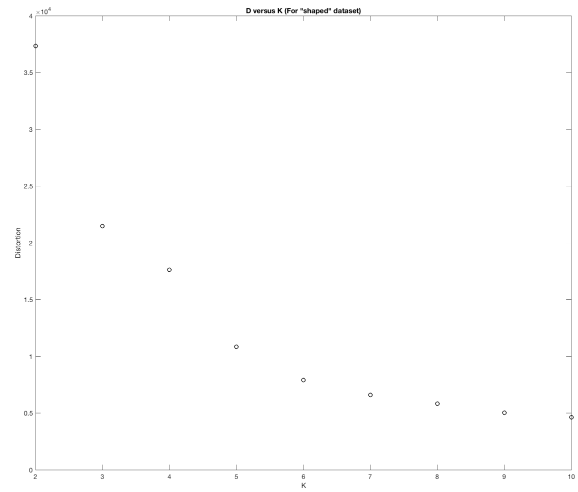

Figure 6: D(distortion) vs K results for ”shaped” data, K-means

– If the distance metric D is defined as averaged distance value.

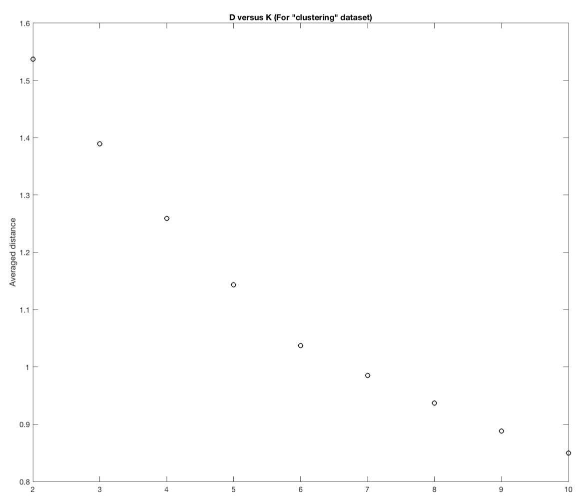

Figure 7: D(avg dis) vs K results for ”clustering” data, K-means

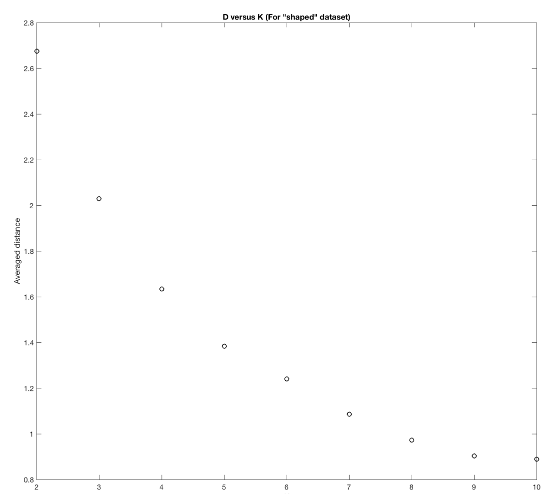

Figure 8: D(avg dis) vs K results for ”shaped” data, K-means

- Use GreedyKcenters to find initializations.

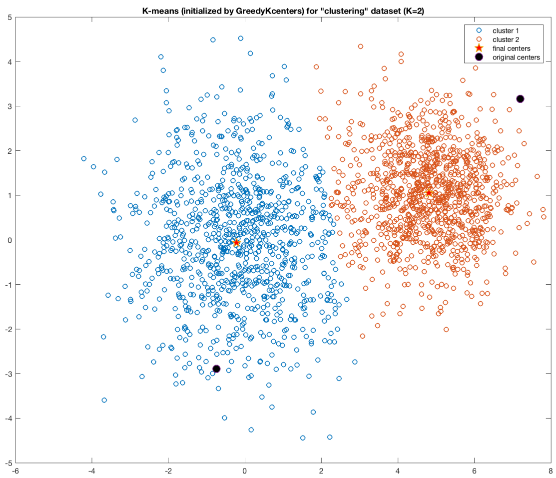

Figure 9: Clustering results for "clustering" data, K-means, initialized by GreedyKCenters

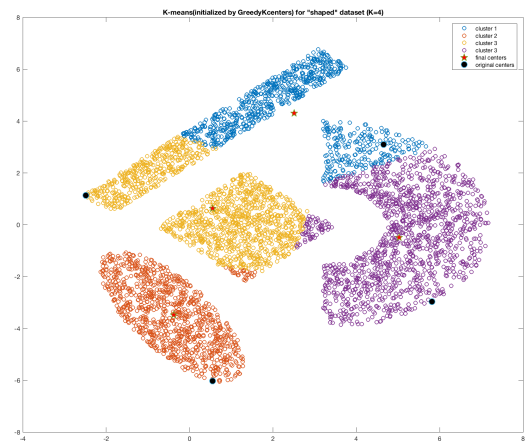

Figure 10: Clustering results for "shaped" data, K-means, initialized by GreedyKCenters

Without GreedyKcenters initializations, average number of iteration for K-means to converge is roughly 6 (for mixture-of-gaussian data). The number of iteration for K-means initialized by GreedyKCenters is 7.

2. Spectral clustering (unnormalized) output plots for best K.

- $K = 2$ for mixture-of-gaussian dataset and $K = 4$ for "shaped" dataset. Here results in Fig.11 are with parameters $\sigma = 2$ and $k = 0.5 * N$. And results in Fig.12 are with parameters $\sigma = 0.5$ and $k = 0.25 * N$. Slightly tuning these two values also yields good results. We can conclude that spectral clustering with appropriate parameters work for both datasets.

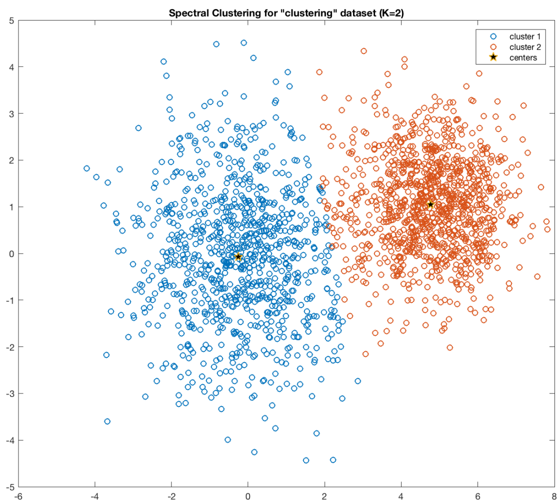

Figure 11: Clustering results for "clustering" data, Spectral Clustering

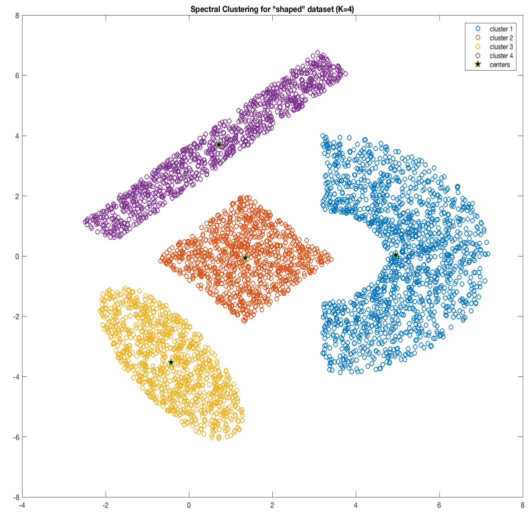

Figure 12: Clustering results for "shaped" data, Spectral Clustering

- D vs K plots for mixture-of-gaussian dataset and "shaped" dataset.
  - If the distance metric D is defined as distortion,

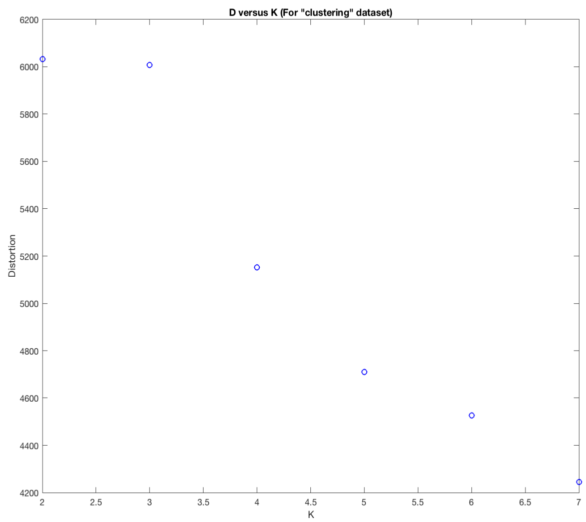

Figure 13: D(distortion) vs K results for "clustering" data, Spectral Clustering

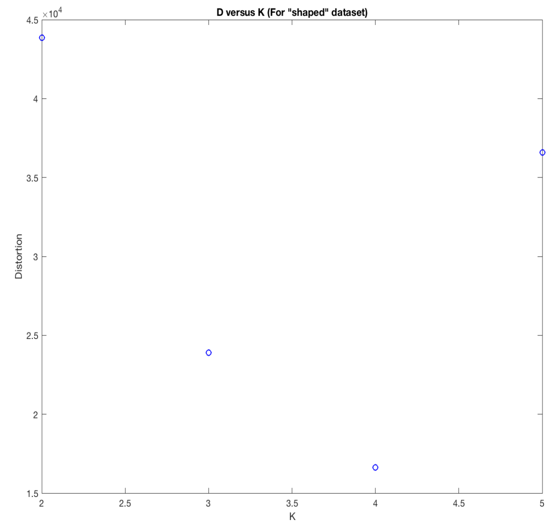

Figure 14: D(distortion) vs K results for "shaped" data, Spectral Clustering

– If the distance metric D is defined as averaged distance value,

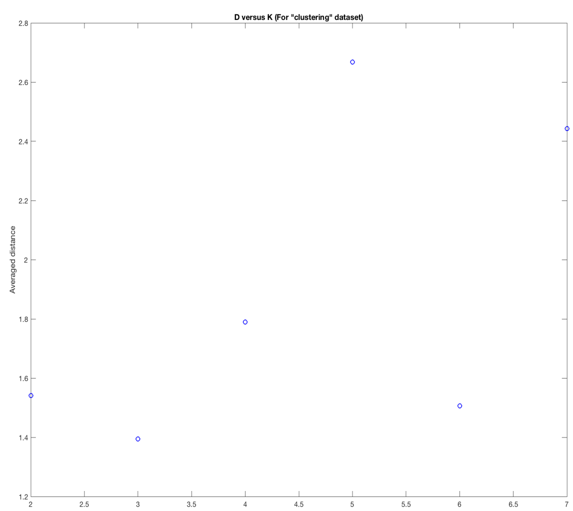

Figure 15: D(avg dis) vs K results for "clustering" data, Spectral Clustering

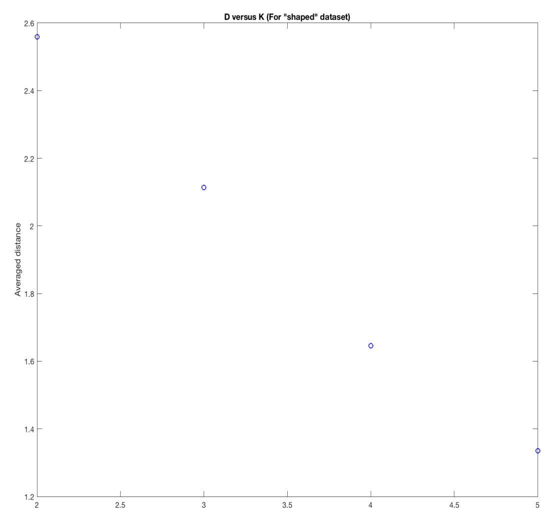

Figure 16: D(avg dis) vs K results for "shaped" data, Spectral Clustering

3 Convergence criterion for K-means

Results from Fig.3 to Fig.10 are obtained with $\| T_{p+1} - Y_p \| < tol$ and $tol = 1 \times 10^{-5}$. For mixture-of-guassian data, *tol* can even be set 0.9 to get good results. For "shaped" dataset, no good results can be obtained no matter how small the *tol* is.

4 Comparison between K-means and Spectral clustering

Spectral clustering works on both dataset (see Fig.11 and Fig.12). But K-means algorithm only works on misture-of-gaussian dataset (see Fig.3 and Fig.4).

5 "Natural" clusters

Mixture-of-gaussian dataset has two "natural" clusters and "shaped" dataset has 4 "natural" clusters.

6 Computational effort

- K-means
  Each iteration has complexity $O(NKd)$. Worst case complexity is $O(N^{dk+1} \log n)$.

- Spectral Clustering The overall computational complexity of spectral clustering algorithm is $O(N^3)$. (The most expensive step is the computation of the eigenvalues/eigenvectors of Laplacian matrix, which has complexity $O(N^3)$. The construction of similarity matrix has time complexity $O(N^2)$ and the application of k-means in the results of eigenvalue decomposition costs $O(NldK)$, where N is the number of input data points, l is the number of k-means iterations, d is the dimensionality of the input data and K is the number of final clusters).

- Runing time (Matlab, MacPro (2.7 GHz Intel i7, 16 GB RAM))

Algorithm	dataset	running time
K-means (K=2)	mixture-of-gaussian data	0.026s
K-means (K=4)	shaped data	0.36s
Spectral clustering (K=2)	mixture-of-gaussian data	8.27s
Spectral clustering (K=4)	shaped data	74.91s