Homework 2

Song Jiecheng 13300180032

Question 1 : Interpret the relationship between Spectral Clustering , Normalized Spectral Clustering and Graph Cut.

We assume that the elements can be divided into two clusters. We establish the adjacent matrix and the Laplace matrix, we suppose the Laplace matrix is W and every element is w_{ij}. The two clusters are A and $\bar{A}$. So in Graph Cut, we calculate the “cut” of the two cluster. The formula is

$$\mathrm{cut}(A,\bar{A}) = \sum_{i\in A, j\notin A} w_{ij},$$

and if we establish a vector

$$\mathrm{q} = [q_1, q_2, \dots, q_n]^T \ \ ,q_i = \begin{cases} c_1 & i \in A \\ c_2 & i \notin A \end{cases}$$

we can find that

$$\mathrm{Cut}(A,\bar{A}) = \sum_{i\in A, j\notin A} w_{ij} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(q_i - q_j)^2}{8}$$

And,

$$\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(q_i - q_j)^2 = \sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(q_i^2 - 2q_iq_j + q_j^2)$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{n} -2w_{ij}2q_iq_j + \sum_{i=1}^{n} q_i^2 \left(\sum_{j=1}^{n} w_{ij}\right)$$

$$= 2q^T L q$$

And in Spectral Clustering, we find the Fiedler vector and do k-means clustering with the vector, and the Fiedler vector is the solution of

$$\min_q \ q^T L q$$
$$\text{s.t.} \ \ q^T q = 1 \ \ q^T 1 = 0$$

so we can calculate the Fiedler vector to calculate the vector q in Graph Cut.

And we can extend the conclusion to the Multi-dimensions, if we divided the points into k clusters (k > 2), in Graph Cut, we can establish a n*(k-1) matrix $\mathrm{F} = (f_1, f_2, \dots, f_{k-1})$

$$f_{ij} = \begin{cases} 1 & if v_i \in A_j \\ 0 & otherwise \end{cases}$$

Then, we calculate F to minimize the $\mathrm{Tr}(F'LF)$. So, in Spectral Clustering we calculate the smallest k eigenvalues (the first one is 0) and eigenvectors, we make F with eigenvectors from the second-smallest to k-smallest eigenvectors, and $\mathrm{Tr}(F'LF)$ is the sum of the smallest k eigenvalues of L.

And in Normalized Cut, we minimize the value of $\mathrm{Ncut}(A,\bar{A}) = \frac{\sum_{i\in A, j\notin A} w_{ij}}{vol(A)} + \frac{\sum_{i\in A, j\notin A} w_{ij}}{vol(\bar{A})}$,

$vol(A)$ is the sum of the weight of the elements in cluster A. We can rewrite it,

$$\min_f f'Lf \qquad s.t. Df \perp 1, f'Df = vol(V)$$

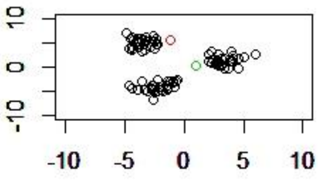
D is a diagonal matrix the k-th element on its diagonal line is the sum of the weight of the k-th elements in the data. And we can calculate the vector g $= D^{-1/2}f$,

$$\min_g g'D^{-1/2}LD^{-1/2}g \qquad s.t. g \perp D^{-1/2}1, g'g = vol(V)$$

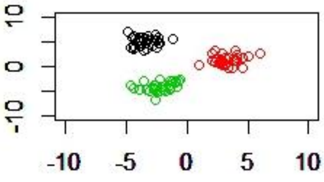
We can find g is the Fiedler vector of $D^{-1/2}LD^{-1/2}$, and we can also extend the conclusion to the Multi-dimensions.

Question 2 : Do clustering with Data1\3.csv with Spectral Clustering (using 3 type of adjacent matrix), and compare the consequence with k- means.

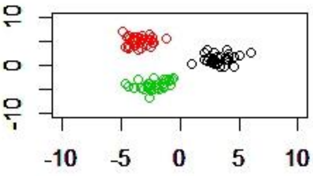
Data1.csv



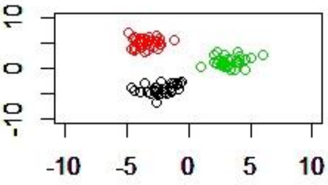
ε-neighborhood(ε=1)



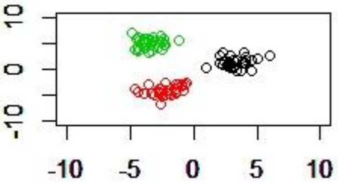
ε-neighborhood(ε=1.5)



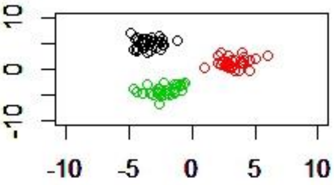
ε-neighborhood(ε=2)



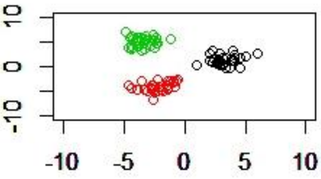
k-nearest(k=5)



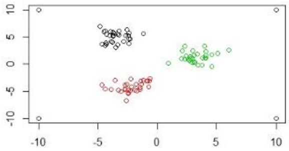
k-nearest(k=10)



k-nearest(k=20)



full connected



k-means(last homework)

Data3.csv

```
1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
1 1 2 1 1 1 1 1 1 1 1 1 1 1 1
1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
3 1 1 1 1 1 1 1 1 1 1 1 1 1 1
```

ε-neighborhood(ε=1)

```
1 1 1 1 1 1 1 2 1 1 1 1 1 1 1
1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
1 1 3 1 1 1 1 1 1 1 1 1 1 1 1
1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
```

ε-neighborhood(ε=1.5)

```
2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
```

ε-neighborhood(ε=3)

```
2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
```

k-nearest(k=5)

```
2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
```

k-nearest(k=10)

```
1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
```

k-nearest(k=20)

```
2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
```

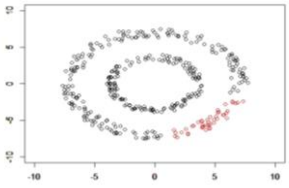
full-connected

```
1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
```

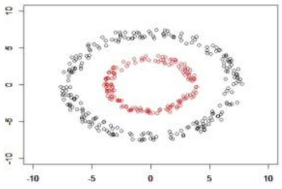
k-means(last homework)

Question 3 : Do clustering with Circles\Curves.csv with Spectral Clustering (using 3 type of adjacent matrix)

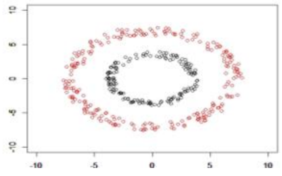
Circles.csv



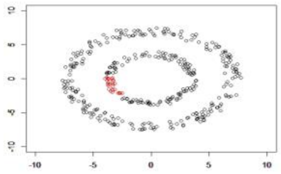
ε-neighborhood(ε=1)



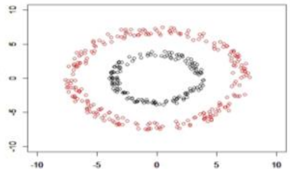
ε-neighborhood(ε=1.5)



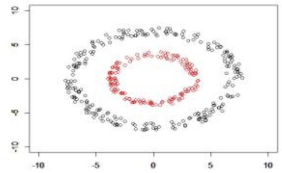
ε-neighborhood(ε=2)



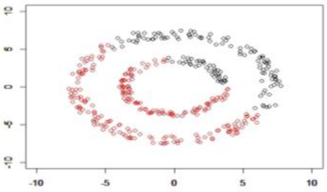
k-nearest(k=5)



k-nearest(k=10)

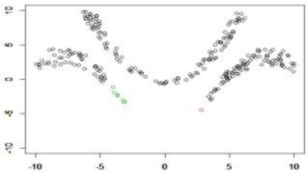


k-nearest(k=20)

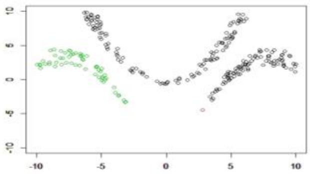


full-connected

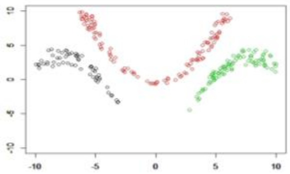
Curves.csv



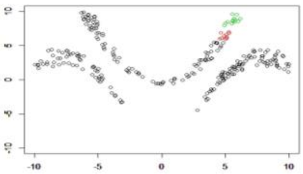
ε-neighborhood(ε=1)



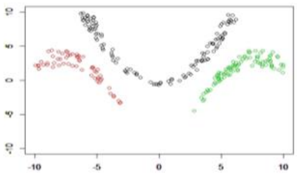
ε-neighborhood(ε=1.5)



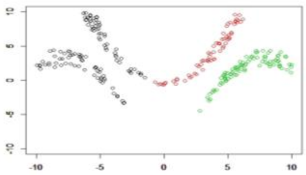
ε-neighborhood(ε=2)



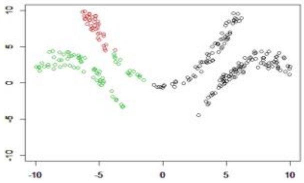
k-nearest(k=5)



k-nearest(k=10)



k-nearest(k=20)



full-connected

Thinking question : Can we use gap statistic to determine the number of clusters in Spectral Clustering.

I think we cannot use gap statistic, we calculate the distance in one cluster and divide the number of elements in this cluster. But in Spectral Clustering, the data is the eigenvectors of L, cannot be used in gap sstatistic.