Assignment 2, CS 5786

Group member: Yan Deng, yd256, Haonan Liu hl955, Nianze Liu nl443, Yunzhou Zhang yz824

Q1. Spectral clustering

- Data generation and cluster assignment

Initially, we generate 30 data point with two clusters. As shown in Fig.1, point 1, 2 , … 15 in color blue are located in cluster 1, point 16, 17, …, 30 in color red are located in cluster 2. For all the points within a cluster, there is a line (edge) connect with each other. The adjacency matrix A is shown in AspectralI.csv. If there is a edge between point i and j, Aij=1; if not Aij=0.

Then we modify the graph by adding three edges between point 1 – point 16, point 15 – point 30, and point 9 – point 23, as shown in figure 2. The corresponding adjacency matrix A' is shown in AspectralII.csv.

We expected the new datasets to result in significantly different clustering. In spectral clustering, the goal is to minimize the normalized cut, that is minimize the total cost while keep the balance of the number in each cluster. The optimal clustering assignment in the modify graph would be a cut between the edge of point 8-point 9, and edge of point 22-23. It is optimal because in this case, the total cut equals to two, and the two clusters has the same data points. Therefore, cluster 1 includes {1,2,3,…,8, 16,17,…22} with color in blue; cluster 2 includes {9, 10, …,15, 23, …30} with color in red. The spectral clustering code we developed verify the assignment.

- Difference in assignment

If data point is in the first cluster I, we identify it as 1 in cluster assignment vector c, 0 otherwise. By calculate the vary of spectral I and II, we check the difference of labels in cluster I and II, then flip the labels in cluster II, and check again for the difference with labels in cluster I, then get the minimum between the two values. The result shows cluster I varieties by 46.7% with cluster II, which meet the requirement.

Fig. 1 The initial graph

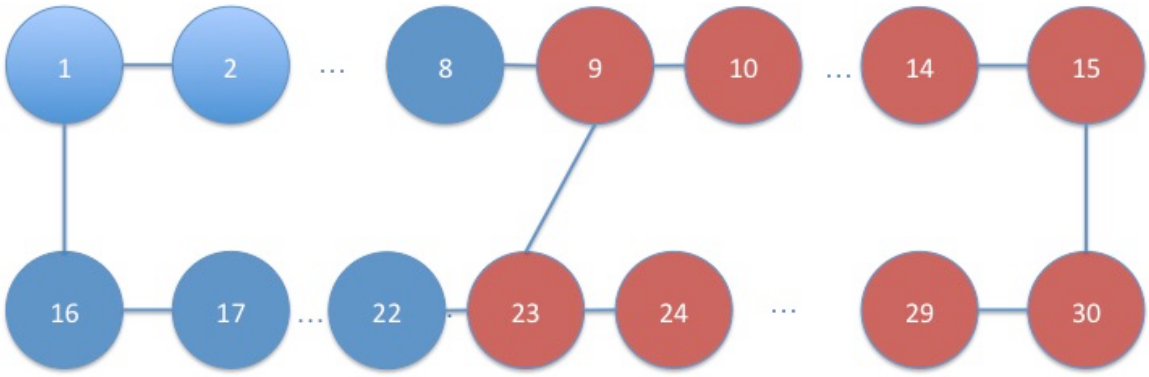

Fig. 2 The modified graph

Appendix: Matlab Code

```
%% AspectralI generation
% A is a matrix that 1-2-3-...-15 connected, 16-17-...-30 connected
data=zeros(30,30);
for i=1:29
    data(i,i+1)=1;
    data(i+1,i)=1;
end

data(15,16)=0;
data(16,15)=0;

csvwrite('AspetralI.csv',data);
A1=data;
D1=diag(sum(A1,2));
L1=eye(30,30)-D1^(-1/2)*A1*D1^(-1/2);
[vector1, value1]=eig(L1);
```

```
y1=vector1(:,1:2);

c1=kmean(y1);
csvwrite('cspectralI.csv', c1);

%% modify the data
% add a link for (1,16), (15,30),(9,23)
data(1,16)=1;
data(16,1)=1;
data(15,30)=1;
data(30,15)=1;
data(9,23)=1;
data(23,9)=1;

csvwrite('AspetralII.csv',data);

A2=data;
D2=diag(sum(A2,2));
L2=eye(30,30)-D2^(-1/2)*A2*D2^(-1/2);
[vector2, value2]=eig(L2);

y2=vector2(:,1:2);

c2=kmean(y2);
csvwrite('cspectralII.csv', c2);

%% difference
c3=~c2;
numdiff1=nnz(c1-c2);
numdiff2=nnz(c1-c3);
vary=min(numdiff1,numdiff2)/30;

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

function c=kmean(y)
mu1_old=[0,0];
mu2_old=[0,0];

mu1_new=[0.267,0];
mu2_new=[0,0.267];

c=zeros(length(y),1);% identify the cluster assignment

while max(mu1_old~=mu1_new)||max(mu2_old~=mu2_new)

    mu1_old=mu1_new;
    mu2_old=mu2_new;


    dis1=sum((y-repmat(mu1_old,length(y(:,1)),1)).^2,2);
    dis2=sum((y-repmat(mu2_old,length(y(:,1)),1)).^2,2);

    c1=0; % number of data in cluster 1
    c2=0; % number of data in cluster 2
```

```
    sum_c1=[0,0];
    sum_c2=[0,0];

    for i=1:30 %length(y2(:,1))
        if (dis1(i)-dis2(i))<=0
            c(i,:)=1;% i is in cluster 1
            c1=c1+1;
            sum_c1=sum_c1+y(i,:);
        else
            c(i,:)=0;% i is in cluster 2
            c2=c2+1;
            sum_c2=sum_c2+y(i,:);

        end
    end

    mu1_new=sum_c1/c1;
    mu2_new=sum_c2/c2;

end
```

Q2: EM algorithm

E-step:

$$Q_t^{(i)}(c_t) = P\left(c_t = k \middle| x_t, \theta^{(i-1)}\right)$$
$$\propto P\left(x_t \middle| c_t = k, \theta^{(i-1)}\right) \cdot P\left(c_t = k \middle| \theta^{(i-1)}\right)$$
$$\propto \Phi(x_t, \lambda) \cdot \pi_k^{(i-1)}$$
$$Q_t^{(i)}(c_t) = \frac{\lambda_k^{x_t} e^{-\lambda_k}}{\sum_{k=1}^{K} \lambda_k^{x_t} e^{-\lambda_k}}$$

M-step for π_k:

$$\pi_k = \underset{\pi_k}{argmax} \left(\sum_{t=1}^{n} \sum_{k=1}^{K} Q_t^{(i)}(k) \, log(\pi_k) \right)$$
$$\pi_k = \frac{\sum_{t=1}^{n} Q_t^{(i)}(k)}{n}$$

M-step for λ_k:

$$\lambda_k = \underset{\lambda_k}{argmax} \left(\sum_{t=1}^{n} \sum_{k=1}^{K} Q_t^{(i)}(k) \, log\left(P(x_t | c_t = k, \theta)\right) \right)$$
$$= \underset{\lambda_k}{argmax} \left(\sum_{t=1}^{n} \sum_{k=1}^{K} Q_t^{(i)}(k) \, log\left(\frac{\lambda_k^{x_t} e^{-\lambda_k}}{x_t!}\right) \right)$$
$$= \underset{\lambda_k}{argmax} \left(\sum_{t=1}^{n} \sum_{k=1}^{K} Q_t^{(i)}(k) (x_t log(\lambda_k) - \lambda_k - log(x_t!)) \right)$$

Take the derivative of the RHS and equate it to 0

Then we get:

$$\sum_{t=1}^{n} Q_t^{(i)}(k) \left(x_t \frac{1}{\lambda_k} - 1 \right) = 0$$

So:

$$\lambda_k = \frac{\sum_{t=1}^{n} Q_t^{(i)}(k) \, x_t}{\sum_{t=1}^{n} Q_t^{(i)}(k)}$$