Data Mining Homework - 4

Graph Spectra

Group – 7

Xing Zeng <xingzeng@kth.se>   Sevket Melih Zenciroglu <smzen@kth.se>

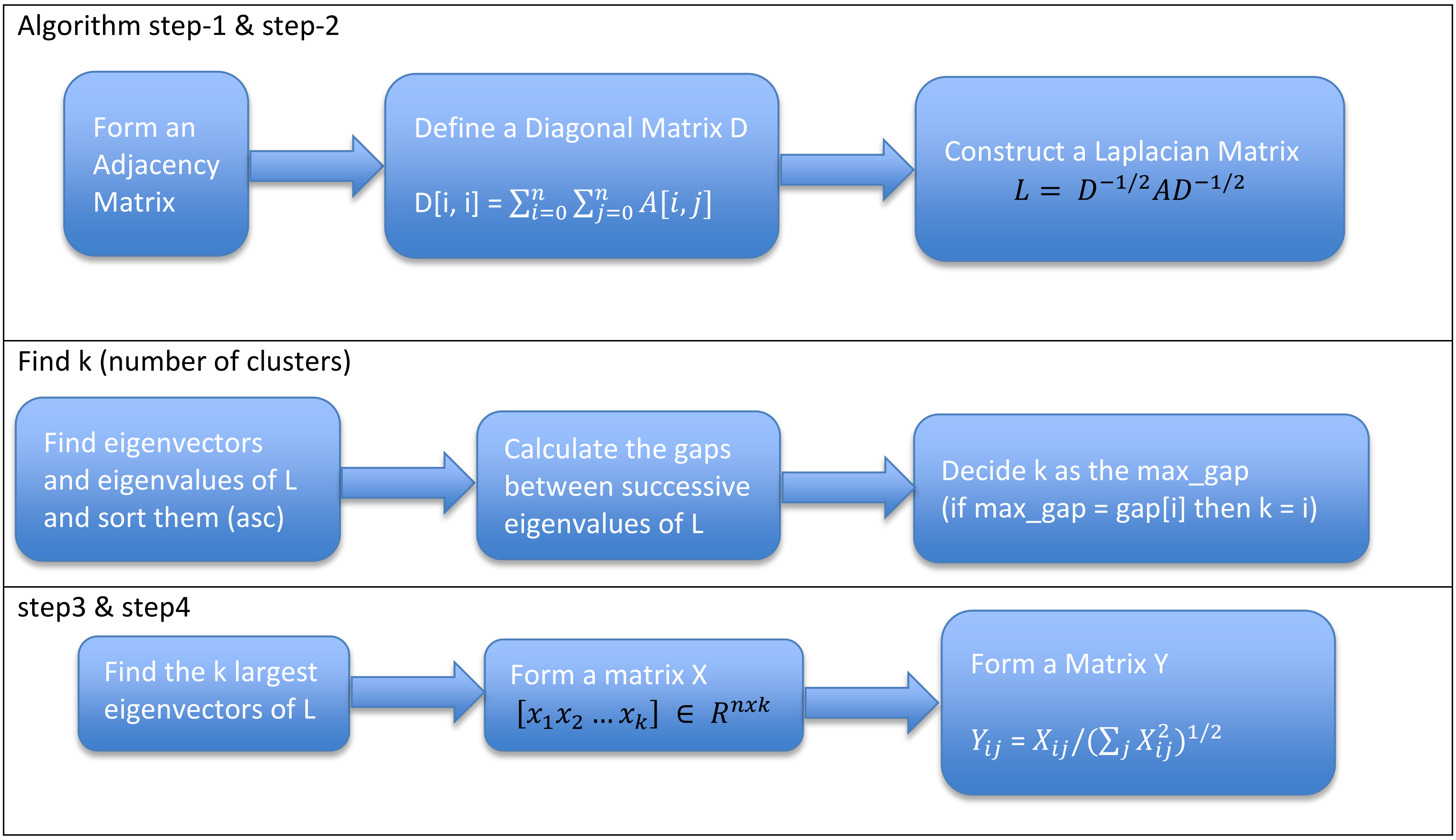
1. Dataset

1) A real graph “example1.dat” -- This data set was prepared by Ron Burt. He dug out the 1966 data collected by Coleman, Katz and Menzel on medical innovation. They had collected data from physicians in four towns Illinois, Peoria, Bloomington, Quincy and Galesburg.

2) A synthetic graph “example2.dat”

2. Workflow

Given a set of points $S = \{s_1, \dots, s_n\}$ in R^l, cluster in to k subsets



Algorithm step-1 & step-2

Form an Adjacency Matrix

Define a Diagonal Matrix D

D[i, i] = $\sum_{i=0}^{n} \sum_{j=0}^{n} A[i,j]$

Construct a Laplacian Matrix $L = D^{-1/2}AD^{-1/2}$

Find k (number of clusters)

Find eigenvectors and eigenvalues of L and sort them (asc)

Calculate the gaps between successive eigenvalues of L

Decide k as the max_gap (if max_gap = gap[i] then k = i)

step3 & step4

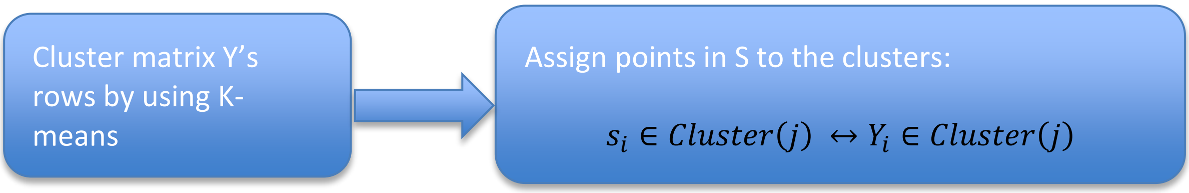
Find the k largest eigenvectors of L

Form a matrix X $[x_1 x_2 \dots x_k] \in R^{nxk}$

Form a Matrix Y

$Y_{ij} = X_{ij}/(\sum_j X_{ij}^2)^{1/2}$

step5 & step6



Cluster matrix Y's rows by using K-means

Assign points in S to the clusters:

$$s_i \in Cluster(j) \leftrightarrow Y_i \in Cluster(j)$$

3. Run the code:

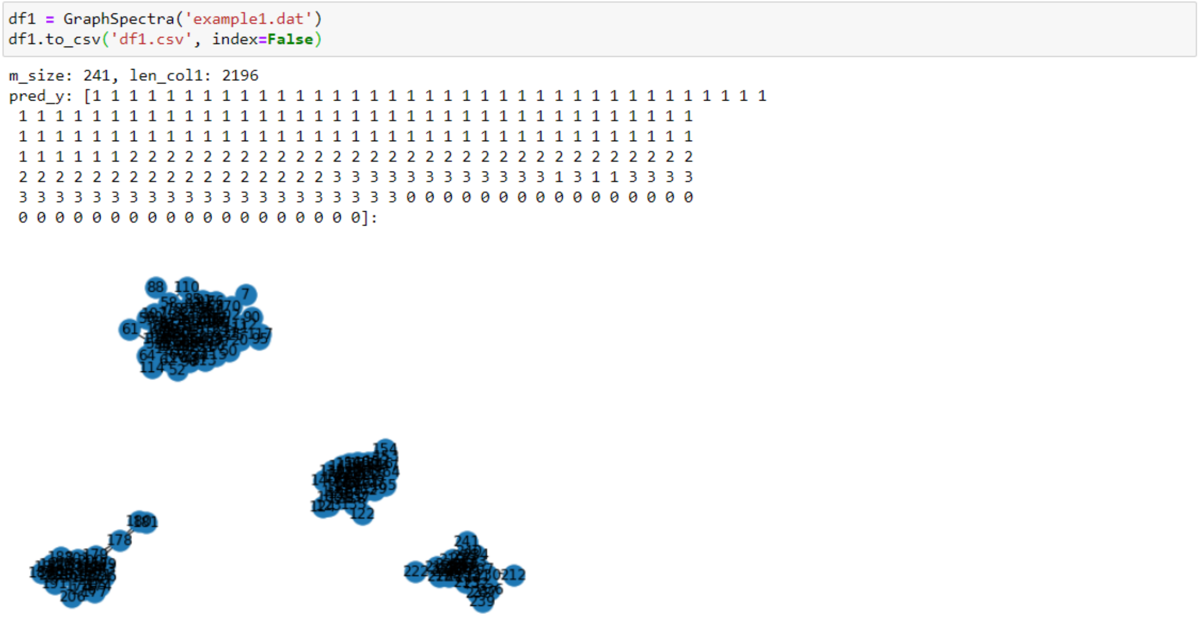
You can basically run the entire notebook (Group7_Homework4.ipynb) with the default settings on it.

4. Results:

The first dataset contains 4 clusters. All those cluster are disconnected from each other, that's why we have 4 1s in L's eigenvalues.

The second dataset has 2 clusters but those clusters are not fully separated, they have connections between them. So, L has only 1 eigenvalue having the value of 1.

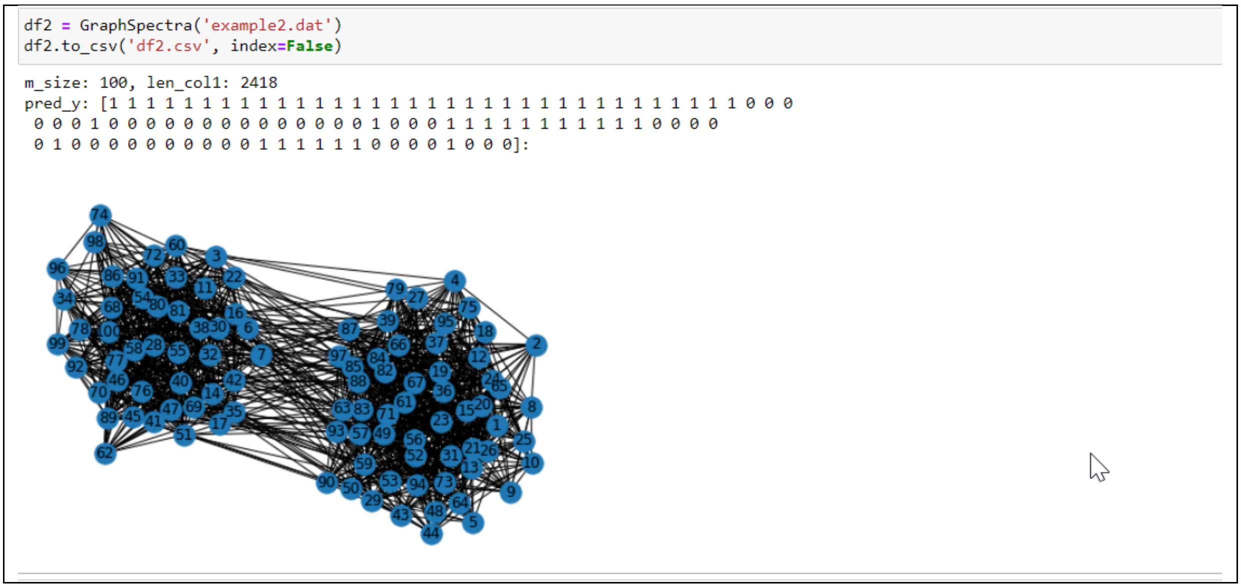
First Dataset



```
df1 = GraphSpectra('example1.dat')
df1.to_csv('df1.csv', index=False)
```

```
m_size: 241, len_col1: 2196
pred_y: [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 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 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 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 1 3 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 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0]:
```

Second Dataset



```
df2 = GraphSpectra('example2.dat')
df2.to_csv('df2.csv', index=False)
```

```
m_size: 100, len_col1: 2418
pred_y: [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 0 0 0
 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 1 1 1 1 1 1 1 1 1 1 1 0 0 0 0
 0 1 0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 0 0 0 0 1 0 0 0]:
```

df2.xlsx   df1.xlsx