Spectral Clustering

Ira A. Fulton Schools of Engineering
Arizona State University

Spectral Clustering: Introduction

Ira A. Fulton Schools of Engineering
Arizona State University

Objective

Objective

Illustrate the key idea of spectral clustering

Objective

Define basic graph notations useful for spectral clustering

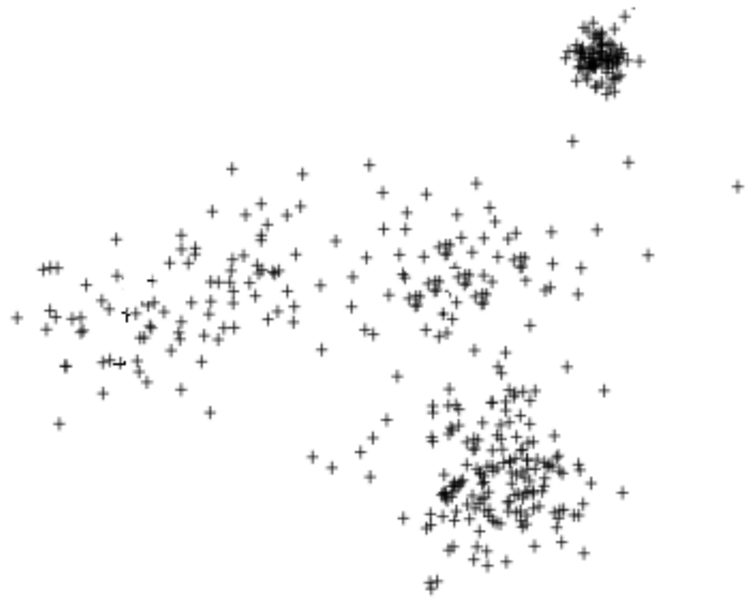
Revisiting k-means & mixture models

- K-means use “hard” membership while mixture models allow “soft” membership

- Both use feature/vector representation of the data as input ➔ E.g., Euclidean distance is one natural (dis)similarity measure.
  - What if the input data is NOT represented in feature/vector, format?
    - E.g., graph data.
    - E.g., objects with only pair-wise similarities (like individuals on a social network ➔ community detection)

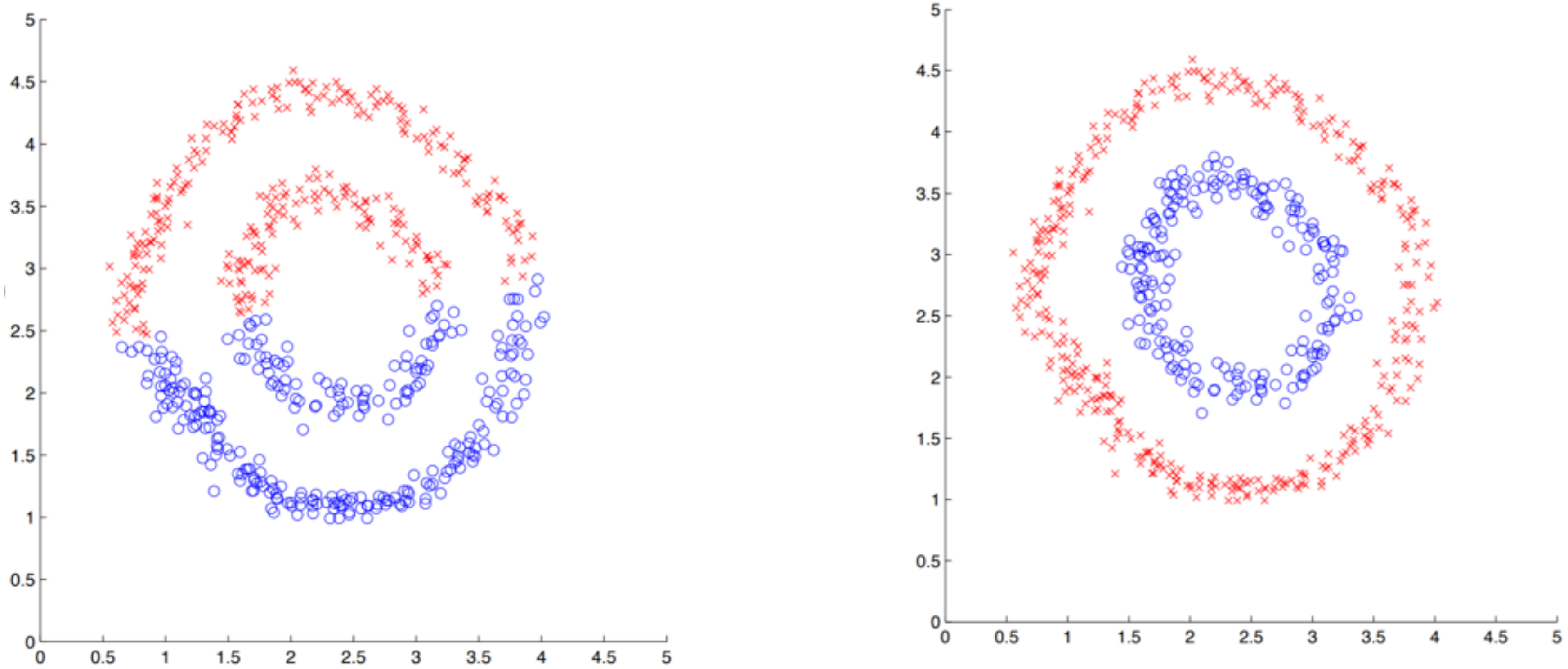
Revisiting k-means & mixture models

- In both k-means and mixture models, we look for compact clustering structures.



- In some cases, connected-component structures may be more desirable.

Example



Source: Ng, A.Y., Michael I.J., and Yair, W. "On spectral clustering: Analysis and an algorithm." *Advances in neural information processing systems*. 2002.

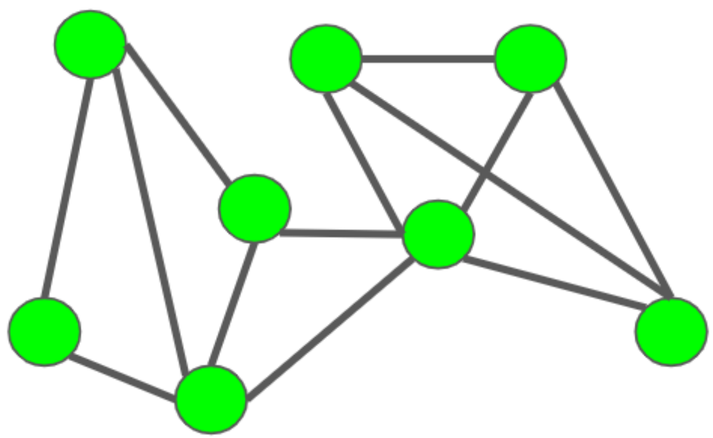
Spectral Clustering

- A family of methods for finding such similarity-based clusters
  - "Spectral": for using the eigenvalues (spectrum) of the *similarity matrix* of the data.
  - Graph clustering, similarity-based clustering
- The objects to be clustered are not in a vector space.
  - The primary feature is the similarity between objects.
  - For any pair of objects *i* and *j*, we have a value *s*(*i*,*j*) measuring their similarity; all such values form the similarity matrix.

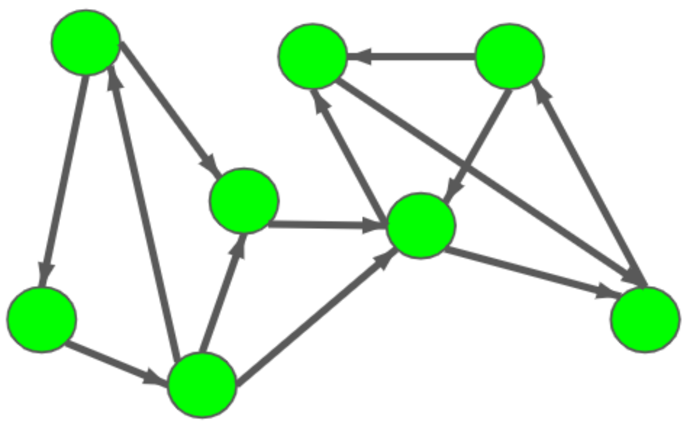
➔ **Graphs** are intuitive for representing/visualizing such data.

Graph Representation

- Definition: A graph $G = (V, E)$ is defined by V, a set of N *vertices*, and E, a set of *edges*.



Undirected graph



Directed graph

In spectral clustering, we consider undirected graphs.

Graph Representation (1/4)

Adjacency matrix **W** of undirected graph

- $\mathrm{N}\times N$ symmetric binary matrix
- The row and columns are indexed by the vertices and the entries represent the edges of the graph

$$\begin{cases} w_{i,j} = 0 & \textit{if vertices } i,j \textit{ are not connected} \\ w_{i,j} = 1 & \textit{if vertices } i,j \textit{ are connected} \end{cases}$$

- Simple graph = zero diagonal

Graph Representation (2/4)

- Weighted adjacency matrix (sometimes called affinity matrix)
  - Allow values other than 0 or 1
  - Each edge is weighted by pairwise similarity

$$\begin{cases} w_{i,j} = 0 & if\ i,j\ are\ not\ connected \\ w_{i,j} = s(i,j) & if\ i,j\ are\ connected \end{cases}$$

- $w_{i,j}$ may be defined through some kernel functions.

Graph Representation (3/4)

Degree matrix **D** of undirected graph

- $N \times N$ diagonal matrix that contains information about the degree of each vertex.
- Degree $d(v_i)$ of a vertex v_i: # of edges incident to the vertex.
  - Extended to sum of weights from edges incident to the vertex.
- So, we have:

$$\mathbf{D} = \begin{bmatrix} d(v_1) & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & d(v_N) \end{bmatrix}$$

Graph Representation (4/4)

- Laplacian matrix **L** of undirected graph
  - $\mathbf{L} = \mathbf{D} - \mathbf{W}$ (Degree-Affinity) (Unnormalized)
  - $\mathbf{L}$ is symmetric and positive semi-definite
  - N non-negative real-valued eigenvalues
  - The smallest eigen-value is 0, the corresponding eigenvector is the 1-vector (all elements being 1).
  - The smallest non-zero eigenvalue of **L** is called the spectral gap.