数据挖掘导论

Introduction to Data Mining

Advanced Clustering

D I-LAB 数 据 智 能 实 验 室 DATA INTELLIGENCE LABORATORY
浙江大学 Zhejiang University

Agenda

- ☐ Soft Clustering
- ☐ Gaussian Mixture Model
- ☐ Spectral Clustering
- ☐ Bi-Clustering
- ☐ Summary

Soft Clustering

- In hard clustering (e..g, k-means), each point belongs to only one cluster
- In soft clustering, each point can belong to multiple classes
  - ✓ An apple could be red or green
  - ✓ In Marketing, each user may belong to multiple target audience groups
- There is a weight or probability between each point and each cluster

$$SSE = \sum_{j=1}^{k}\sum_{i=1}^{N} w_{ij}^{\mathrm{m}}\left(x_i - c_j\right)^2, \qquad \sum_{j=1}^{k} w_{ij} = 1$$

  - ✓ m is any real number greater than 1

Fuzzy C-means Clustering

- Extends k-means clustering with the following idea:
  - ✓ In the iterations, repeat the following steps:
    - Fix c_j and determine the best w_{ij}
    - Fix w_{ij} and recompute c_j

$$w_{ij} = \frac{1}{\sum_{k=1}^{C}\left(\frac{\|x_i - c_j\|}{\|x_i - c_k\|}\right)^{\frac{2}{m-1}}}$$

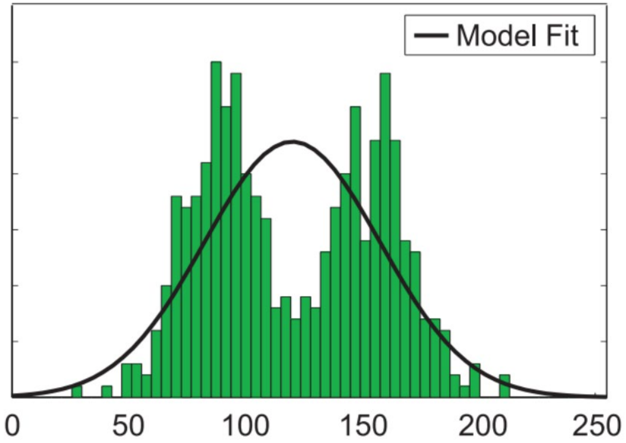
$$c_j = \frac{\sum_{i=1}^{N} w_{ij}^{m} \cdot x_i}{\sum_{i=1}^{N} w_{ij}^{m}}$$

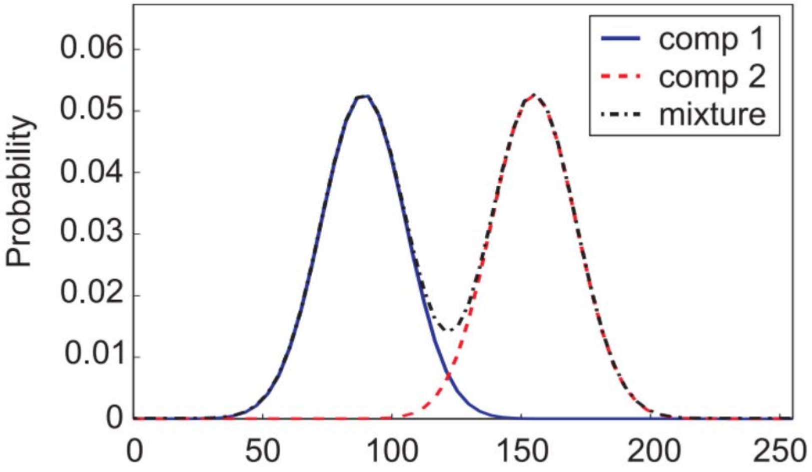
Agenda

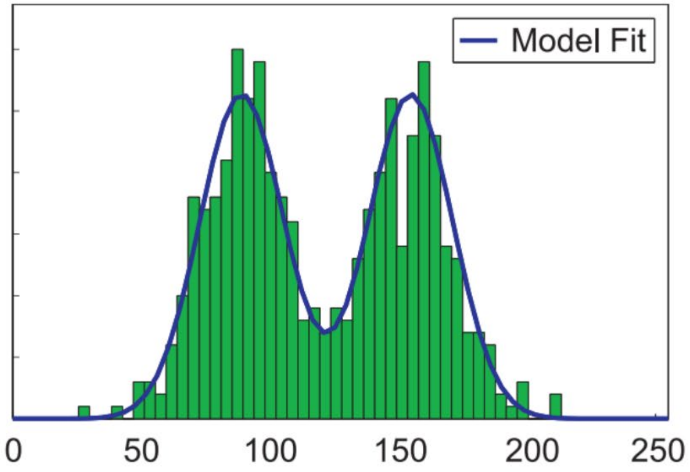
- Soft Clustering
- Gaussian Mixture Model
- Spectral Clustering
- Bi-Clustering
- Summary

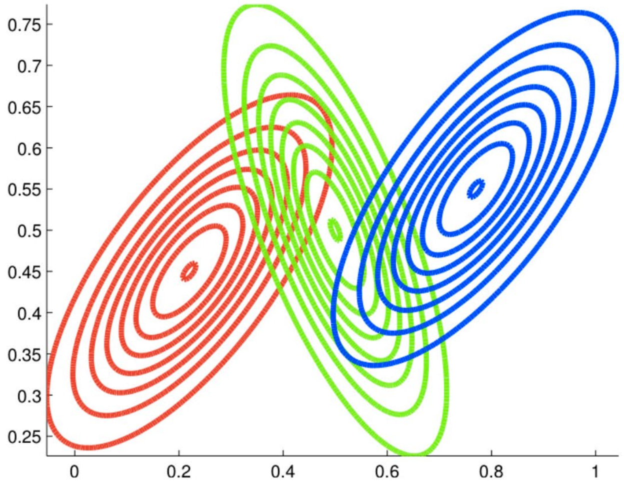
Clustering based on Mixture Model

- Fit data with mixture of Gaussian distributions









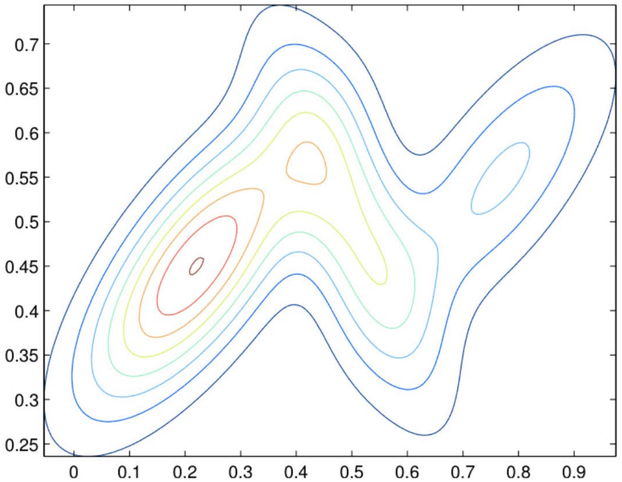
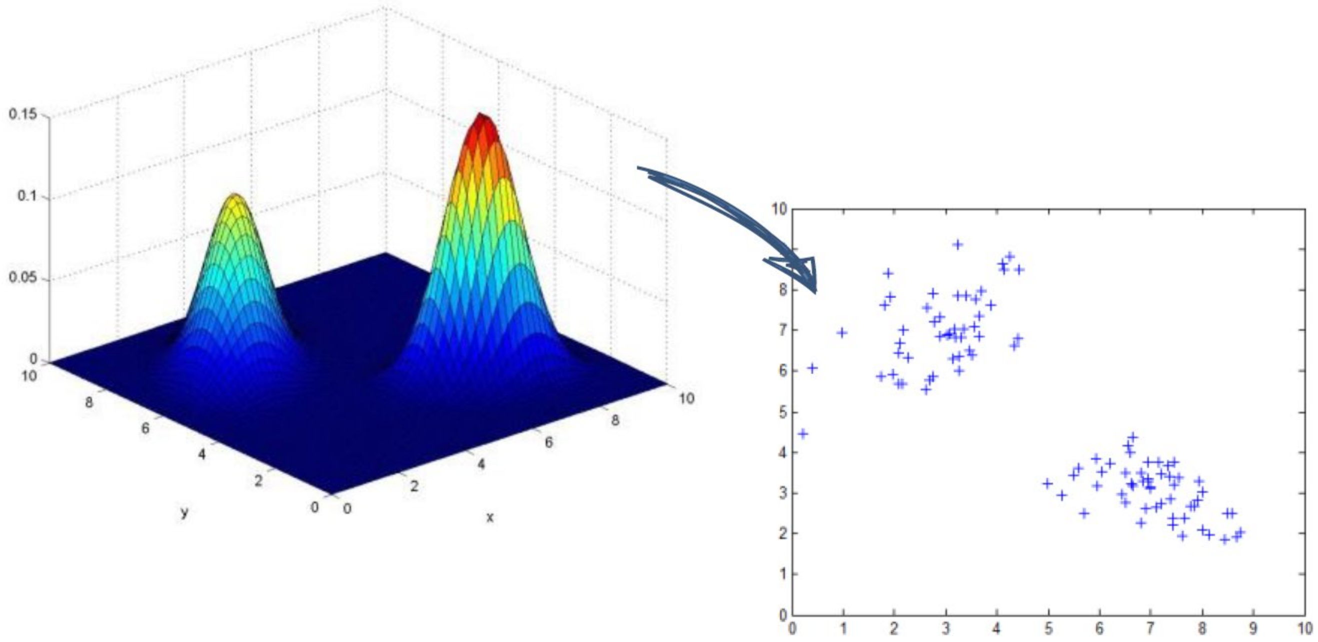


Image source: https://www.cs.toronto.edu/~rgrosse/courses/csc411_f18/slides/lec16-slides.pdf

Main Idea

- The data points are generated by the underlying mixture of Gaussian models
- Treat each cluster as one component of the mixture distribution, **whose mean is the cluster center**
- Each point has a certain probability of belonging to each cluster



EM Algorithm

Expectation Maximization Algorithm

- ✓ **E-step:** Compute the posterior probability over z given our current model - i.e. how much do we think each Gaussian generates each datapoint
  - Estimate the probability of each point to a cluster

$$\gamma_{ik} = \frac{\pi_k \mathcal{N}(\mathbf{x}_i | \boldsymbol{\mu}_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}_i | \boldsymbol{\mu}_j, \Sigma_j)}$$

- ✓ **M-step:** Assuming that the data really was generated this way, change the parameters of each Gaussian to maximize the probability that it would generate the data it is currently responsible for
  - ✓ Update the parameters of Gaussian distributions

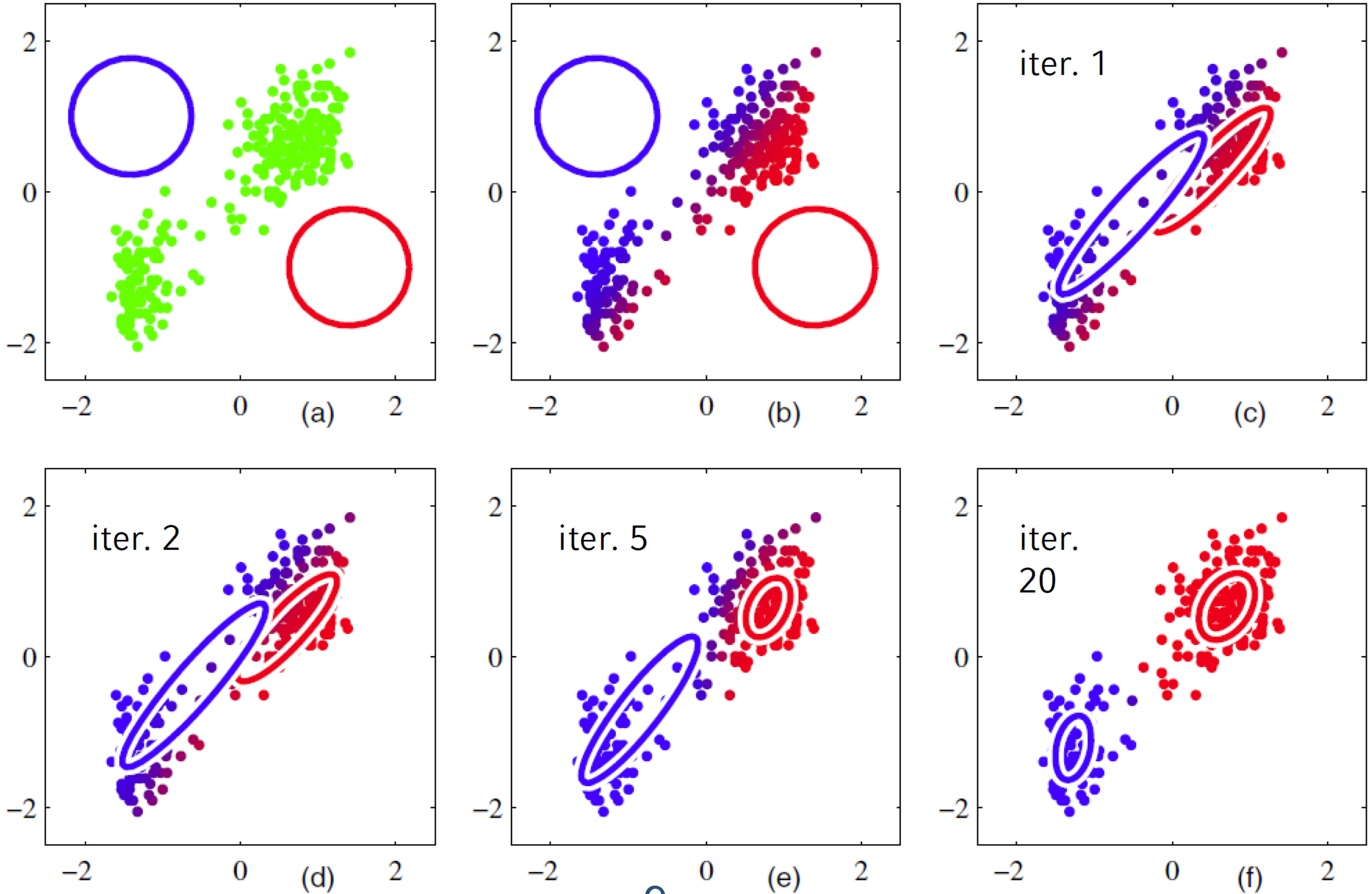
$$\boldsymbol{\mu}_k = \frac{1}{N_k} \sum_{i=1}^{N} \gamma_{ik} \mathbf{x}_i$$

$$\Sigma_k = \frac{1}{N_k} \sum_{i=1}^{N} \gamma_{ik} (\mathbf{x}_i - \boldsymbol{\mu}_k)(\mathbf{x}_i - \boldsymbol{\mu}_k)^\top$$

$$\pi_k = \frac{N_k}{N} \quad \text{where } N_k = \sum_{i=1}^{N} \gamma_{ik}$$

Main Idea

- Iteratively improve the parameters of each distribution until some quality threshold is reached



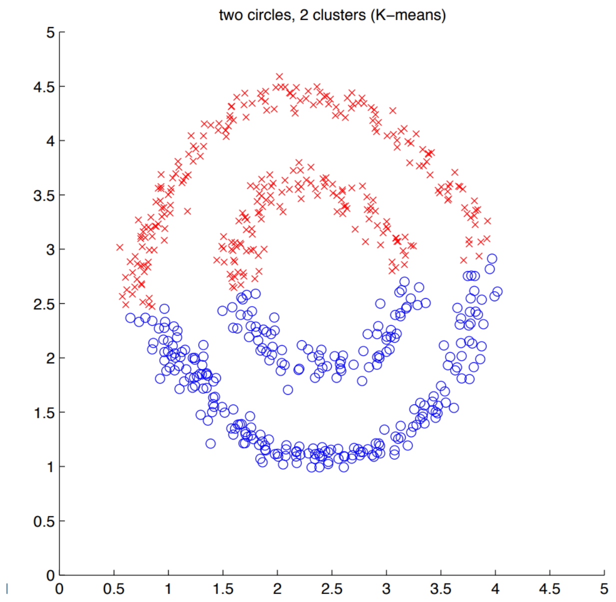
Agenda

- ☐ Soft Clustering
- ☐ Gaussian Mixture Model
- ☐ Spectral Clustering
- ☐ Bi-Clustering
- ☐ Summary

Spectral Clustering

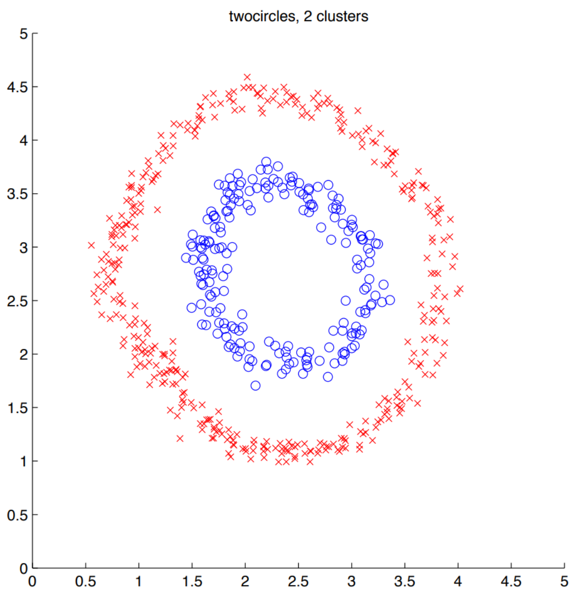
- Clusters are generated based on pairwise proximity/similarity/affinity
- Clusters do not have to be Gaussian or compact

K-means

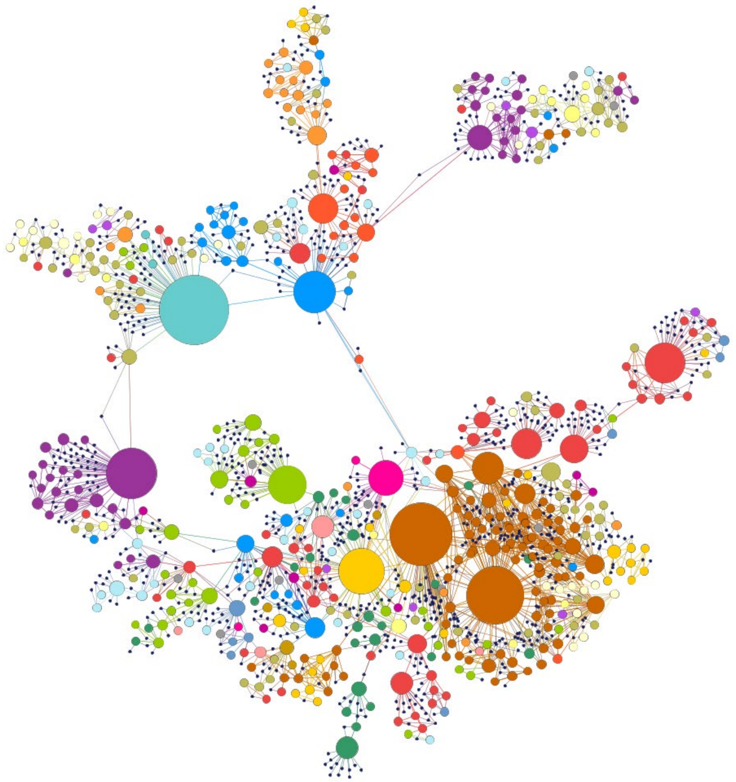


(i)

Spectral clustering

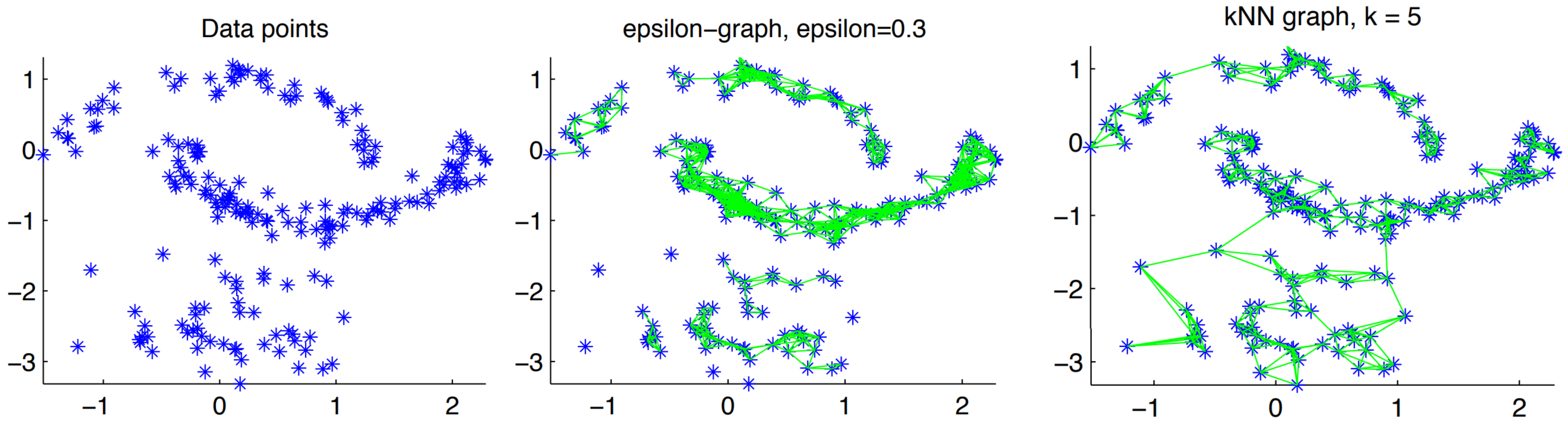


(e)



Similarity Graph Construction

- Compute the similarity between two objects
- Can use a threshold or kNN to reduce the number of edges

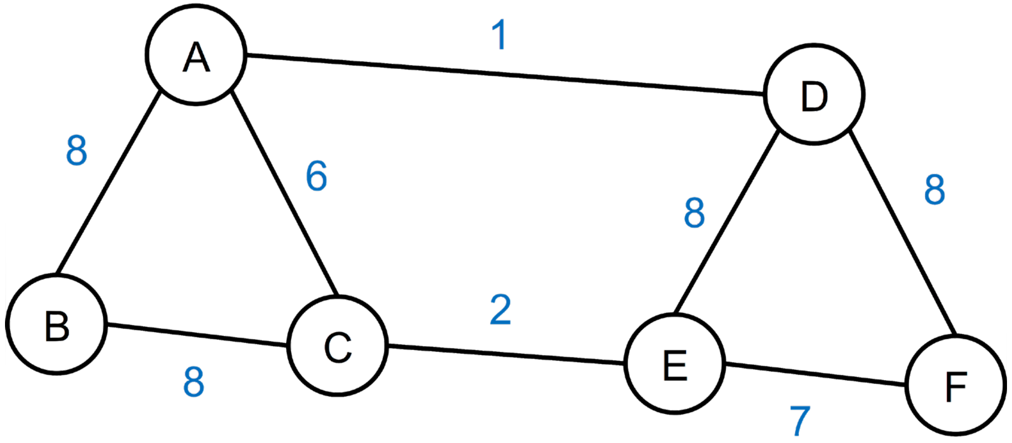


Spectral Clustering Algorithm

1. Create a similarity graph between our N objects to cluster.
2. Compute the first k eigenvectors of its Laplacian matrix to define a feature vector for each object.
3. Run k-means on these features to separate objects into k classes.

An Example of Spectral Clustering

Get Similarity Matrix



$$S = \begin{bmatrix} 0 & 8 & 6 & 1 & 0 & 0 \\ 8 & 0 & 8 & 0 & 0 & 0 \\ 6 & 8 & 0 & 0 & 2 & 0 \\ 1 & 0 & 0 & 0 & 8 & 8 \\ 0 & 0 & 2 & 8 & 0 & 7 \\ 0 & 0 & 0 & 8 & 7 & 0 \end{bmatrix}.$$

Example from: https://changyaochen.github.io/spectral-clustering/

An Example of Spectral Clustering

Get Laplacian matrix

$$L = \begin{bmatrix} 15 & -8 & -6 & -1 & 0 & 0 \\ -8 & 16 & -8 & 0 & 0 & 0 \\ -6 & -8 & 16 & 0 & -2 & 0 \\ -1 & 0 & 0 & 17 & -8 & -8 \\ 0 & 0 & -2 & -8 & 17 & -7 \\ 0 & 0 & 0 & -8 & -7 & 15 \end{bmatrix}.$$

$$L_{\text{sym}} = D^{-1/2} L D^{1/2}$$

$$L_{\text{sym}} = \begin{bmatrix} 0.130 & -0.069 & -0.051 & -0.008 & 0. & 0. \\ -0.069 & 0.137 & -0.068 & 0. & 0. & 0. \\ -0.051 & -0.068 & 0.137 & 0. & -0.017 & 0. \\ -0.008 & 0. & 0. & 0.145 & -0.068 & -0.068 \\ 0. & 0. & -0.017 & -0.068 & 0.145 & -0.060 \\ 0. & 0. & 0. & -0.068 & -0.060 & 0.130 \end{bmatrix}$$

L=D-S

D is a simple diagonal matrix

Dii equals the sum of ith row of S

Example from: https://changyaochen.github.io/spectral-clustering/

An Example of Spectral Clustering

- Take the first k_1 eigenvectors, and to form a $N \times k_1$ matrix U
- Form a matrix T from U by normalizing the rows of U to norm 1
- Perform k-means clustering on the points in matrix T. Each row is considered as a point.

$$T = \begin{bmatrix} 0.706 & -0.708 \\ 0.677 & -0.735 \\ 0.738 & -0.674 \\ 0.710 & 0.703 \\ 0.740 & 0.672 \\ 0.677 & 0.735 \end{bmatrix}$$

Example from: https://changyaochen.github.io/spectral-clustering/

Agenda

- ☐ Soft Clustering
- ☐ Gaussian Mixture Model
- ☐ Spectral Clustering
- ☐ Bi-Clustering
- ☐ Summary

Motivation of Bi-Clustering

Simultaneous clustering of both rows and columns of a data matrix

- ✓ Bob is planning a housewarming party for his new 3-room house
- ✓ He owns 30 albums and wants to play different music in each room
- ✓ He has invited 50 guests and sent out a survey to each guest asking if they like each album
- ✓ He collects the data into a 50×30 binary matrix M, where M_{ij}=1if guest i likes album j
- ✓ Bob wants to distribute people and albums evenly among the rooms of his house

$$s(\boldsymbol{M}, \boldsymbol{r}, \boldsymbol{c}) = b(\boldsymbol{r}, \boldsymbol{c}) \cdot \sum_{i,j,k} M_{ij} r_{ki} c_{kj} \qquad b(\boldsymbol{r}, \boldsymbol{c}) = \exp\left(\frac{-(\max(\mathcal{S}) - \min(\mathcal{S}))}{\epsilon}\right)$$

- ✓ b(r,c) penalizes unbalanced solutions
- ✓ r_{ki}=1if guest i is assigned to room k ; c_{kj}=1if room k contains album j

Example from: http://www.kemaleren.com/post/an-introduction-to-biclustering/

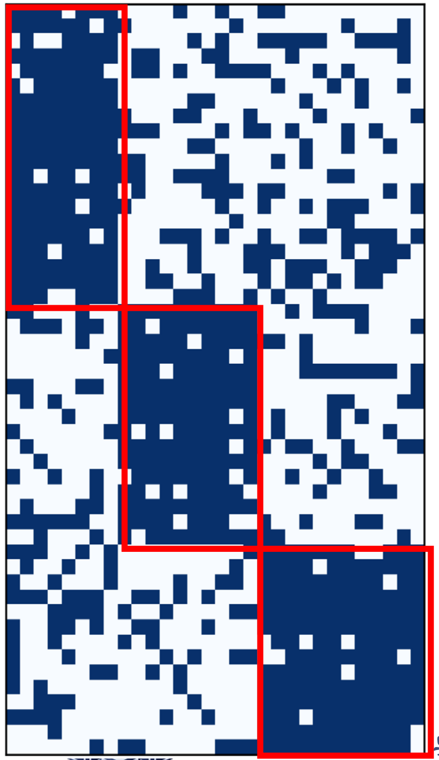
Motivation of Bi-Clustering

- Starting with a random assignment of rows and columns to clusters, Bob **reassigns row and columns to improve the objective function** until convergence
  - ✓ Can use heuristic algorithms (simulated annealing etc) in operation research

Bi-clustering

Re-organize the order of guests and albums

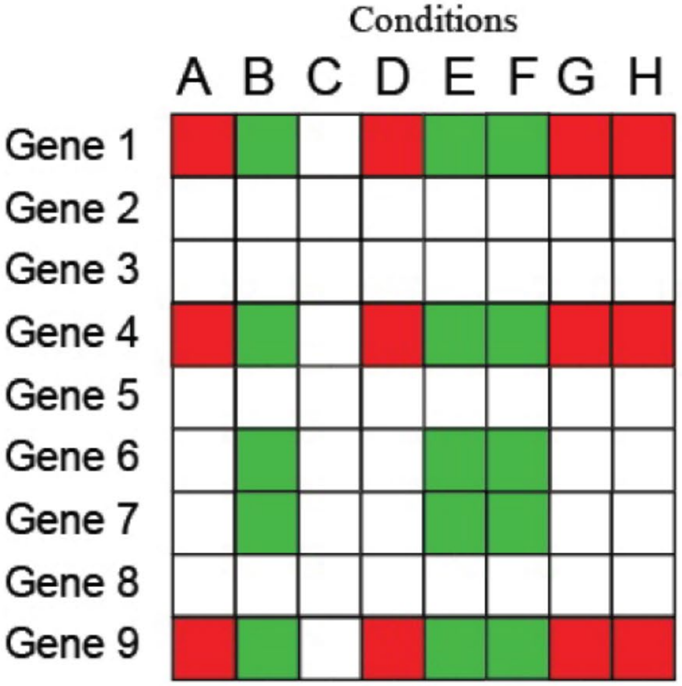
Room 1



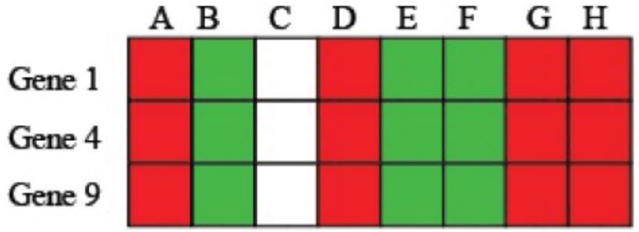
Room 3

Example from: http://www.kemaleren.com/post/an-introduction-to-biclustering/

Bi-clustering on Gene Data

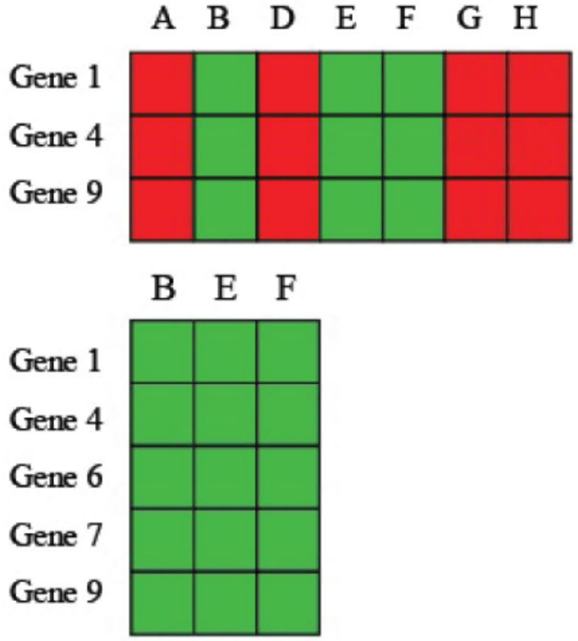


Clustering



Biclustering

Biclustering discovers local coherences over a subset of conditions



Agenda

- Soft Clustering
- Gaussian Mixture Model
- Spectral Clustering
- Bi-Clustering
- Summary