


# CIS 490 Machine Learning

## Lecture 16

Instructor: (Julia) Hua Fang


## Last time

- Unsupervised Learning: **Clustering**
  - ❖ Hierarchical clustering:
    - Types of hierarchical clustering
    - Similarity measures for hierarchical clustering
    - **Dendrogram**: How to interpret and derive Dendrogram.
    - Hierarchical clustering Algorithm
      - Gap statistic: choose optimal number of clusters
  - ❖ Run K-means in R Studio
- Written Homework 2



## Outline

- Advanced topic: Spectral Clustering
- HW3, Exam II, Final project and written report

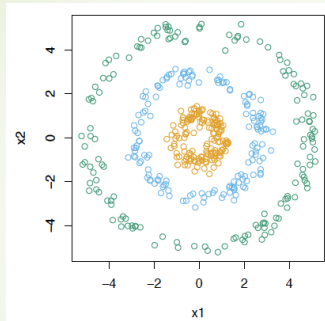


## ■ Overview of Advanced clustering: Spectral Clustering

Read 14.5.3 of "Elements of  
Statistical Learning"

Adapted from Hays, Fern, Jaakkola, Hon, Komber, Pei, James, Witten, Hastie, Tibshirani, Friedman, Howbert, Sontag, Ulrike von Luxburg

## Spectral clustering



Would clustering methods like K-means work? Why?

Answer: Traditional clustering methods like K-means use a spherical or elliptical metric to group data points.  
 --they will not work well when the clusters are non-convex, such as the concentric circles in the plot.

## Spectral Clustering

Spectral clustering is a **generalization** of standard clustering methods, and is designed for these situations, ie. when the clusters are non-convex, such as the concentric circles in the plot.

- The starting point is a  $N \times N$  matrix of pairwise similarities  $s_{ii'} \geq 0$  between all observation pairs.
- We represent the observations in a **similarity graph**.
  - The  $N$  **vertices/nodes**  $v_i$  represent the observations, and pairs of vertices are connected by an **edge (link)** if their similarity is positive (or exceeds some threshold).
  - The edges are weighted by the  $s_{ii'}$ .

Clustering is now rephrased as a **graph-partition problem**, where we identify connected components with clusters.

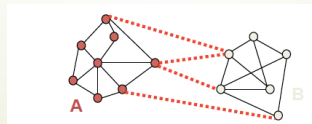
## Spectral clustering

- The goal is to partition the graph, such that edge (link) between different clusters have low weight (more different), and within a cluster have high weight (more similar).
- The idea in spectral clustering is to construct similarity graphs that represent the local neighborhood relationships between observations.
  - There are many ways to define a similarity/affinity matrix and its associated similarity graph that reflect local behavior. The most popular is the K-nearest-neighbor graph.

### Spectral clustering: Using K-nearest-neighbor graph

- Connect vertex  $v_i$  with vertex  $v_j$  if  $v_j$  is among the k-nearest neighbors of  $v_i$ .
- Ignore the directions of the edges (link), ie, connect  $v_i$  and  $v_j$  with an undirected edge if  $v_i$  is among the k-nearest neighbors of  $v_j$ , or if  $v_j$  is among the k-nearest neighbors of  $v_i$ .

The resulting graph is what is usually called the k-nearest neighbor graph.



Each node/vertex is only connected to its K nearest neighbors

## Suppl. Spectral clustering: graph Laplacian

- The main tools for spectral clustering are **graph Laplacian matrices**.
  - There exists a whole field dedicated to the study of those matrices, called spectral graph theory (e.g., see Chung, 1997).

Note: every author just calls "his" matrix the graph Laplacian. Care needed when reading literature on graph Laplacians.

## Suppl. Spectral clustering: graph Laplacian

Let  $G = (V, E)$  be an undirected graph with a vertex set  $V$  and an edge set  $E$ . Assume that the graph  $G$  is weighted (ie., each edge between two vertices carries a non-negative weight  $w_{ij} \geq 0$ )

- **Unnormalized graph Laplacian** matrix is defined as

$$\mathbf{L} = \mathbf{D} - \mathbf{W}.$$

**D** : the degree matrix, containing the degree of vertex  $i$ . The degree of a vertex  $v_i \in V$  is defined as the sum of the weights of the edges connected to it,

$$d_i = \sum_{j=1}^n w_{ij}$$

Note:  $D$  is the diagonal matrix with the degrees  $d_1, \dots, d_n$  on diagonal.

**W** : The matrix of edge weights  $(w_{ij})_{i,j=1,\dots,n}$ , called **weighted adjacency matrix** of the graph.

$w_{ij} = 0$ : means the vertices  $v_i$  and  $v_j$  are not connected by an edge.

- **Normalized graph Laplacian matrix** is a symmetric matrix.

$$\mathbf{L}_{\text{sym}} := \mathbf{D}^{-1/2} \mathbf{L} \mathbf{D}^{-1/2} = \mathbf{I} - \mathbf{D}^{-1/2} \mathbf{W} \mathbf{D}^{-1/2}$$

Goal: finds the first  $k$  eigenvectors corresponding to the  $k$  smallest eigenvalues of  $\mathbf{L}$

## Suppl. Normalized Spectral Clustering: The Ng-Jordan-Weiss (NJW) Algorithm (Ng, Jordan, Weiss, 2002)

Dimension  
Reduction

**Input:** Similarity matrix  $S \in n \times n$ , number  $k$  of clusters to construct.

- Construct a similarity graph by, e.g.,  $k$ -nearest neighbor graph
- Let  $W$  be its **weighted adjacency matrix**.
- Compute the **normalized Laplacian**  $L_{\text{sym}}$ .
- Compute the first  $k$  eigenvectors  $u_1, \dots, u_k$  of  $L_{\text{sym}}$ .
- Let  $U \in \mathbb{R}^{n \times k}$  be the matrix containing the vectors  $u_1, \dots, u_k$  as columns.
- Form the matrix  $T \in \mathbb{R}^{n \times k}$  from  $U$  by normalizing the rows to norm 1, that is, set  $t_{ij} = u_{ij} / (\sum_k u_{ik}^2)^{1/2}$

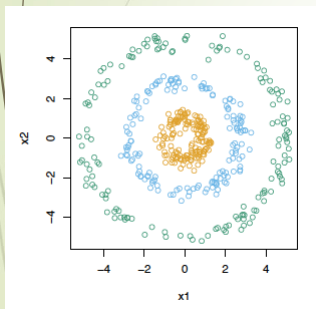
Clustering

- For  $i = 1, \dots, n$ , let  $y_i \in \mathbb{R}^k$  be the vector corresponding to the  $i$ -th row of  $T$ .
- Cluster the points  $(y_i)_{i=1, \dots, n}$  with the  $k$ -means algorithm into clusters  $C_1, \dots, C_k$ .

**Output:** Clusters  $A_1, \dots, A_k$  with  $A_i = \{j \mid y_j \in C_i\}$

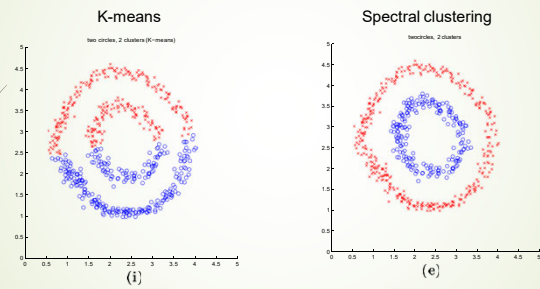
<https://papers.nips.cc/paper/2092-on-spectral-clustering-analysis-and-an-algorithm.pdf>

## Spectral clustering: General Steps

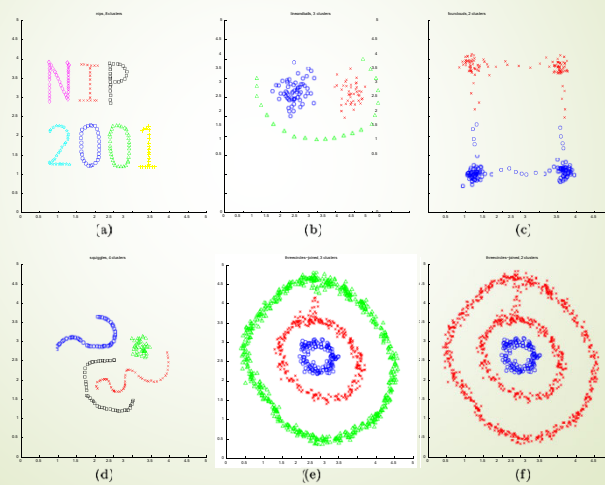


- General steps of spectral clustering:
  - finds the  $k$  eigenvectors  $T_{n \times k}$  corresponding to the  $k$  **smallest** eigenvalues of  $L_{\text{sym}}$
  - Using a standard method like  $K$ -means, cluster the rows of  $T$  to yield a clustering of the original data points.

## Spectral clustering: Graphical illustration



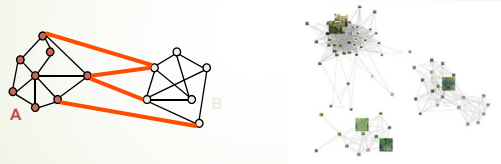
## Spectral clustering: Graphical illustration



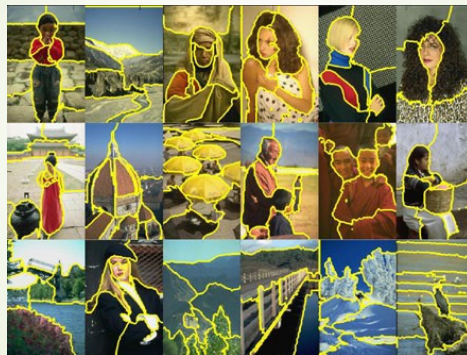


## Spectral clustering: Graphical illustration

Group points based on edges/links in a graph



## Application: Spectral clustering for segmentation





17

## Spectral Clustering: Pros and Cons

### Pros:

- Effective in tasks like image processing
- Can be combined with other clustering methods

### Cons:

- Scalability challenge: Computing eigenvectors on a large matrix is costly
- Must choose the type of similarity graph, eg. nearest neighbors, and associated parameters such as the number of nearest neighbors.
- Must also choose the number of eigenvectors to extract from  $L$  and finally, as with all clustering methods, the number of clusters.

## Suppl.: Spectral Clustering

### ➤ R:

<https://cran.r-project.org/web/packages/kernlab/kernlab.pdf>

## Final Project Written Report Instruction:

Using standard IEEE conference template (e.g, font and font size are all set), posted under Final Project Written Report at myCourses

Note: Fill into the sections on the template as appropriate.

**No more than 8 pages. (Do not paste coding in this written report)**

Required:

1. Title: "CIS490 Project: XXXXXXXX" and Authors
2. Abstract
3. Introduction: Motivation and Brief summary of why, what and how you are going to accomplish in this final project.
4. Literature review & Methods: describe your data and proposed **machine learning (ML) techniques** and **evaluation metrics** in detail

Suggestions: use equations, pseudocodes and illustrative graphs, explain those parameters in your model within the context of your selected data, **NOT in a general sense**.

Continued...

## Final Project Report Instruction:

5. Results: Compare and interpret results from the ML techniques and datasets you used for your project.

Suggestion: eg., using tables, graphs, and corresponding text to discuss your methods, evaluation results, and interpret your findings **within the context of your application area and data**.

6. Conclusion/Discussion: Summary of your project, discuss your findings and limitations. Highlight the part you are most proud of.
7. References: A must.

Statement of team members' project contribution on the last page

## Final Project Report Instruction:

Statement of team members' project contribution on the last page

**Summary of your group meet time and duration need to be included in your HW:**

In person or Zoom:

- . Group meet time and duration (e.g., 5pm-7pm, Feb 1<sup>st</sup>):
- . Average time in communication and discussion regarding assigned group work (via email or other social media, e.g. What's app.):
- . Participants (Print and sign your names):

**Contribution report need to be included in your HW:**

- . If your team members contribute equally to this project, please make this statement "Each member contributes equally" on your last page, so that each of you will receive the same score.
- . If your team members do not contribute equally to this project, please note your team members' names, and mark the percentage of effort each member makes (e.g., Sukumar: 80% then if your group receives a project score of 30, then this member with 80% effort will only get 24).
- . Participants: Print and sign your names

## Coding, Final Project Written report submission: Instruction

- Comment the dataset attributes, the main functions and evaluation methods in your code.
- Submit two files:
  - Written report in Word or PDF format: Give your file name as "CIS490\_FinalProject\_Group#.docx" e.g. "CIS490\_FinalProject\_Group1.docx"
  - Source code file in R, saved as \*.R: include the link to your data.

## Presentation slides presentation and submission: Instruction

- ~15-20 min for each group:
- A smaller version of final project report, covering
  1. Motivation and Brief summary of why, what and how you accomplish in this final project.
  2. Data/Literature review and Methods: describe your data and proposed machine learning (ML) techniques and evaluation metrics
  3. Results from the ML techniques and dataset you used for your project.
  4. Conclusion/Discussion: Summary of your project, discuss your findings and limitations. Highlight the part you are most proud of.
  5. References: A must.
  6. Statement of team members' project contribution

## Grading Rubric: Final project, 100 points

- Slides and presentation: 25 points
- Coding: 35 points
- Written report: 40points

**ATTN: Attach the following on the last page and the last slide**

**Summary of your group meet time and duration need to be included:**

In person or Zoom:

Group meet time and duration (e.g., 5pm-7pm, Feb 1<sup>st</sup>):

Average time in communication and discussion regarding assigned group work (via email or other social media, e.g. What's app.):

Participants (Print and sign your names):

**Contribution report need to be included:**

If your team members contribute equally to this project, please make this statement "Each member contributes equally" on your last page, so that each of you will receive the same score.

If your team members do not contribute equally to this project, please note your team members' names, and mark the percentage of effort each member makes (e.g., Sukumar: 80% then if your group receives a project score of 30, then this member with 80% effort will only get 24).

Participants: Print and sign your names

25

## Exam II: April 26, Tuesday

- ▶ Exam II guideline posted at MyCourses
- ▶ Grading Rubric: 100 points
- ▶ Open Notes; bring your calculator; No smart phones are allowed. Inter-personal exchange is NOT allowed in any format.
- ▶ Exam Place: **Online** ;
- ▶ Date and Time: April 26, 11:00pm ~ 12:15pm; 10:45 start to download the exam paper and upload your exam paper by 12:45pm.
- ▶ Print your name on the first page after reading exam rules.

26

## Due Dates

- ❖ April 12, April 14: Instructor Q&A sessions available both in class and on Zoom. Study group's project preparation.
- ❖ April 19, April 21: Final project presentations
- ❖ April 26: Exam II
- ❖ April 28: Final Written Project Report and Coding Due