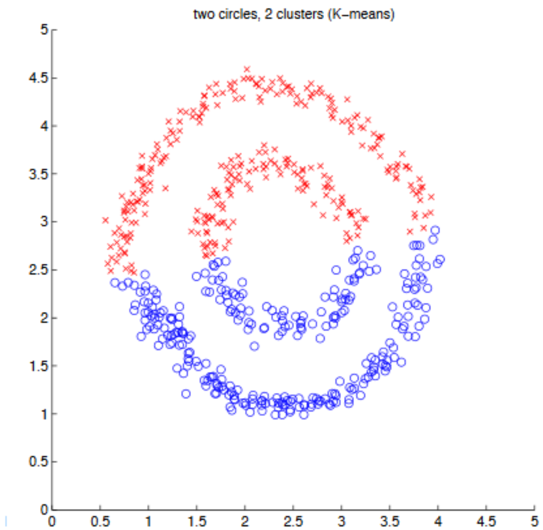
DATA, MSML, BIOI 602
Principles of Data Science

Heng Huang, Ph.D.
Department of Computer Science

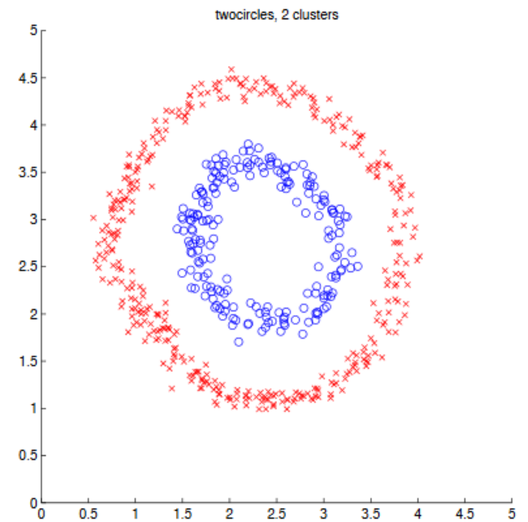
Existing Problems

K-means



(i)

Spectral clustering



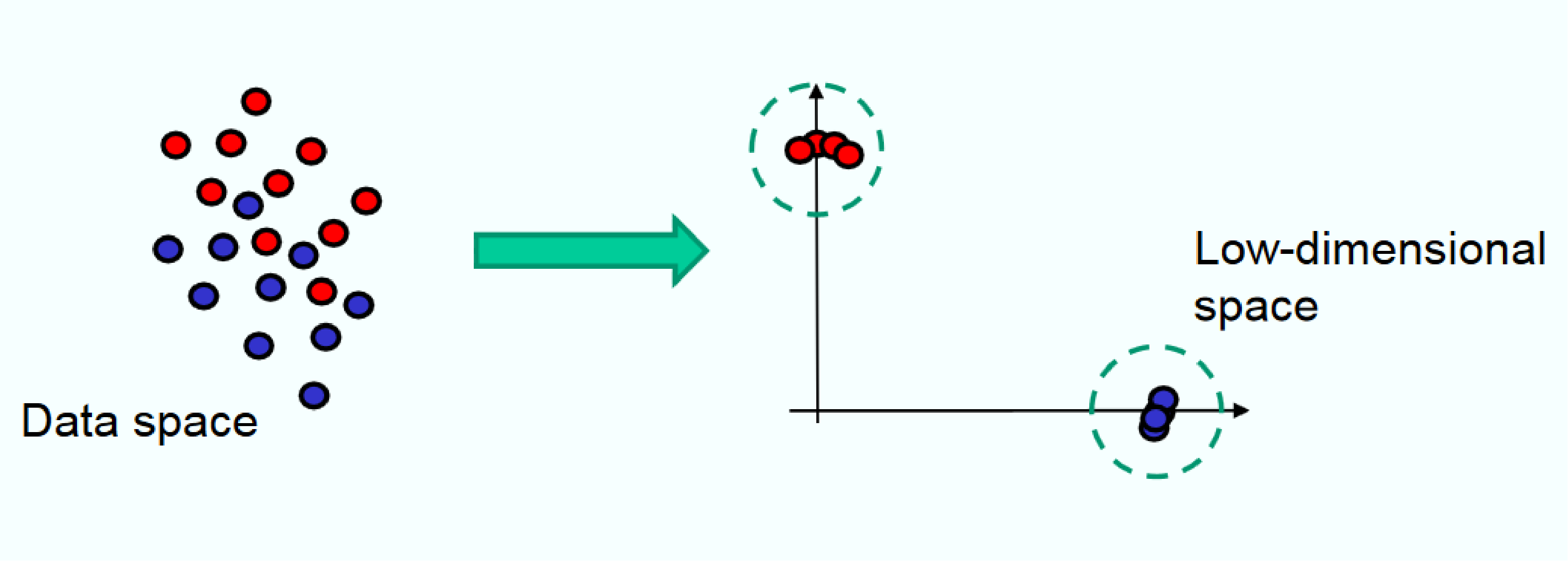
(e)

[Shi & Malik ‘00; Ng, Jordan, Weiss NIPS ‘01]

- Distance based method (like k-means) fails to capture the complicated geometric structure

Spectral Clustering

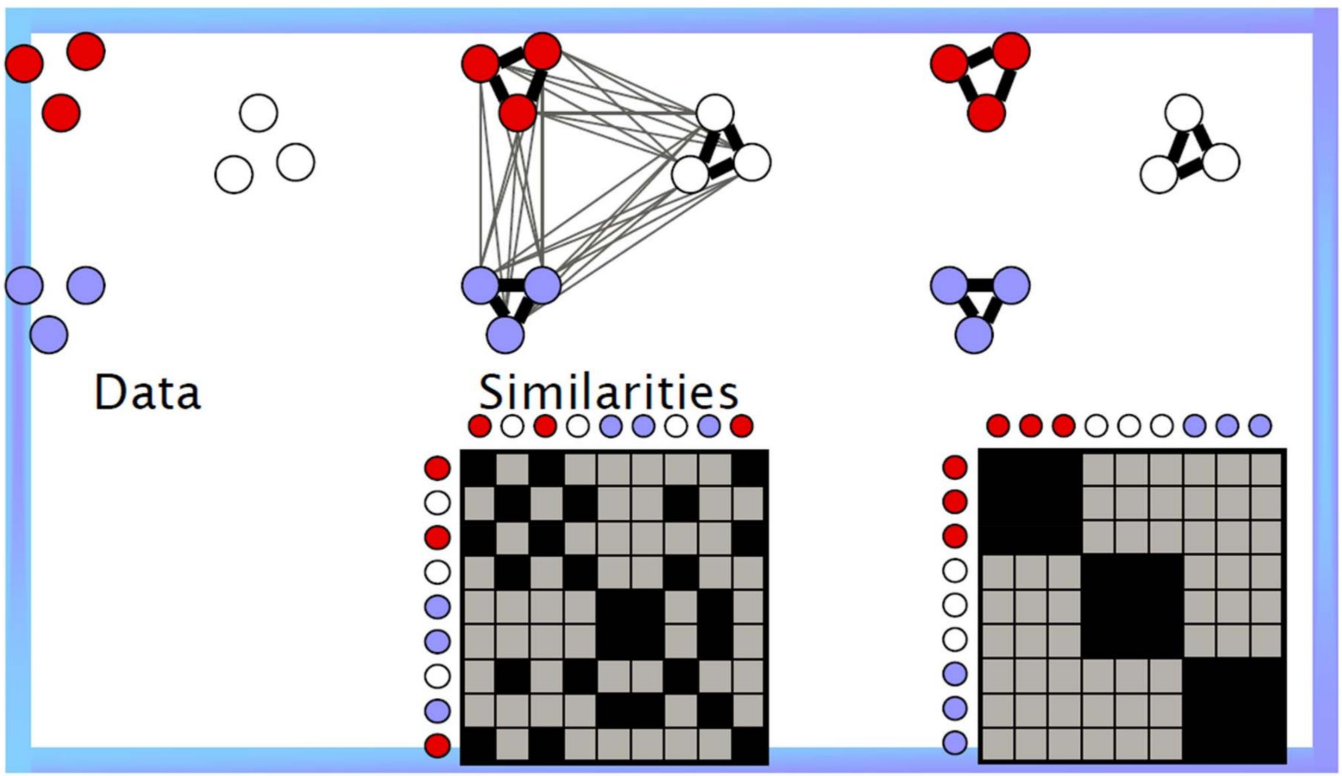
- Basic Idea: Low-dimensional embedding point of view
  - Obtain data representation in the low-dimensional space that can be easily clustered



How to learn this transformation?

Spectral Clustering

- Basic Idea: Graph cut point of view
  - Build similarity graph



Similarity Matrix

Similarity Graphs: Model local neighborhood relations between data points

- ε-neighborhood graph

  Connect all vertices whose pairwise distances are smaller than ε

- k-nearest neighbor graph

  Connect vertex v_i with vertex v_j if v_j is among the k-nearest neighbors of v_i.

- fully connected graph

  Connect all points with positive similarity with each other

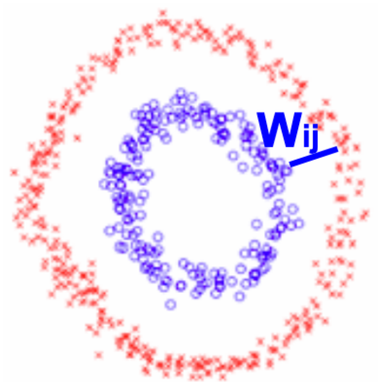
Similarity Matrix

E.g. epsilon-NN

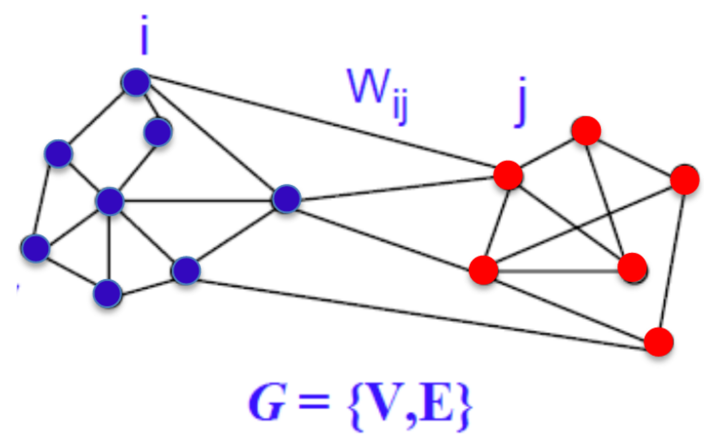
$$W_{ij} = \begin{cases} 1 & \|x_i - x_j\| \le \epsilon \\ 0 & \text{otherwise} \end{cases}$$

Controls size of neighborhood

or mutual k-NN graph (W_{ij} = 1 if x_i or x_j is k nearest neighbor of the other)



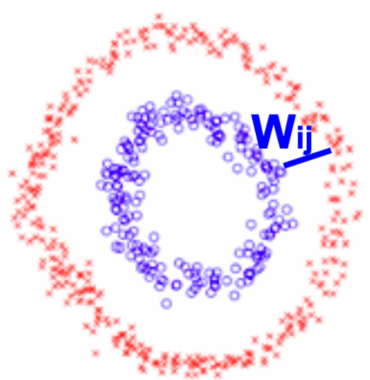
Data clustering



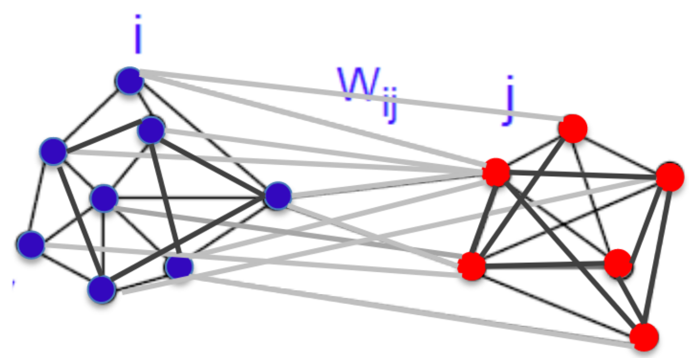
Similarity Matrix

E.g. Gaussian kernel similarity function

$$W_{ij} = e^{\frac{\|x_i - x_j\|^2}{2\sigma^2}} \longrightarrow$$ Controls size of neighborhood



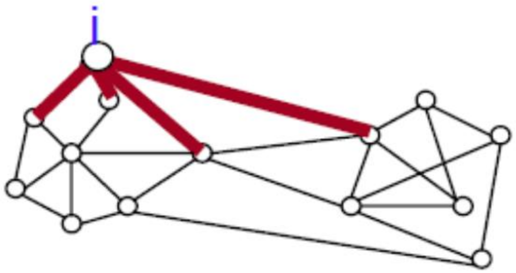
Data clustering



$G = \{V,E\}$

Degree

- $d_i = \sum_j w_{ij}$ degree of a vertex
- $D = diag(d_1, \ldots, d_n)$ degree matrix



	x_1	x_2	x_3	x_4	x_5	x_6
x_1	0	0.8	0.6	0	0.1	0
x_2	0.8	0	0.8	0	0	0
x_3	0.6	0.8	0	0.2	0	0
x_4	0	0	0.2	0	0.8	0.7
x_5	0.1	0	0	0.8	0	0.8
x_6	0	0	0	0.7	0.8	0

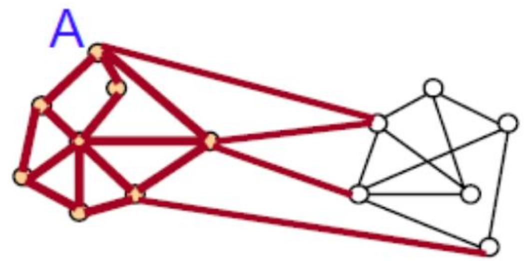
Similarity matrix

	x_1	x_2	x_3	x_4	x_5	x_6
x_1	1.5	0	0	0	0	0
x_2	0	1.6	0	0	0	0
x_3	0	0	1.6	0	0	0
x_4	0	0	0	1.7	0	0
x_5	0	0	0	0	1.7	0
x_6	0	0	0	0	0	1.5

Degree matrix

Volume of Set

- $|A|$ = number of vertices in A
- $\mathrm{vol}(A) = \sum_{i \in A} d_i$



		x_1	x_2	x_3	x_4	x_5	x_6
A	x_1	0	0.8	0.6	0	0.1	0
	x_2	0.8	0	0.8	0	0	0
	x_3	0.6	0.8	0	0.2	0	0
	x_4	0	0	0.2	0	0.8	0.7
B	x_5	0.1	0	0	0.8	0	0.8
	x_6	0	0	0	0.7	0.8	0

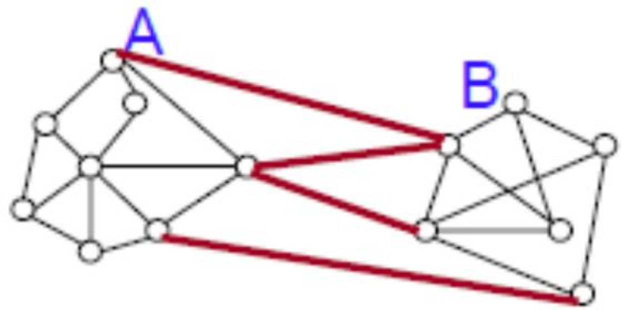
Similarity matrix

	x_1	x_2	x_3	x_4	x_5	x_6
x_1	1.5	0	0	0	0	0
x_2	0	1.6	0	0	0	0
x_3	0	0	1.6	0	0	0
x_4	0	0	0	1.7	0	0
x_5	0	0	0	0	1.7	0
x_6	0	0	0	0	0	1.5

Degree matrix

Graph Cut

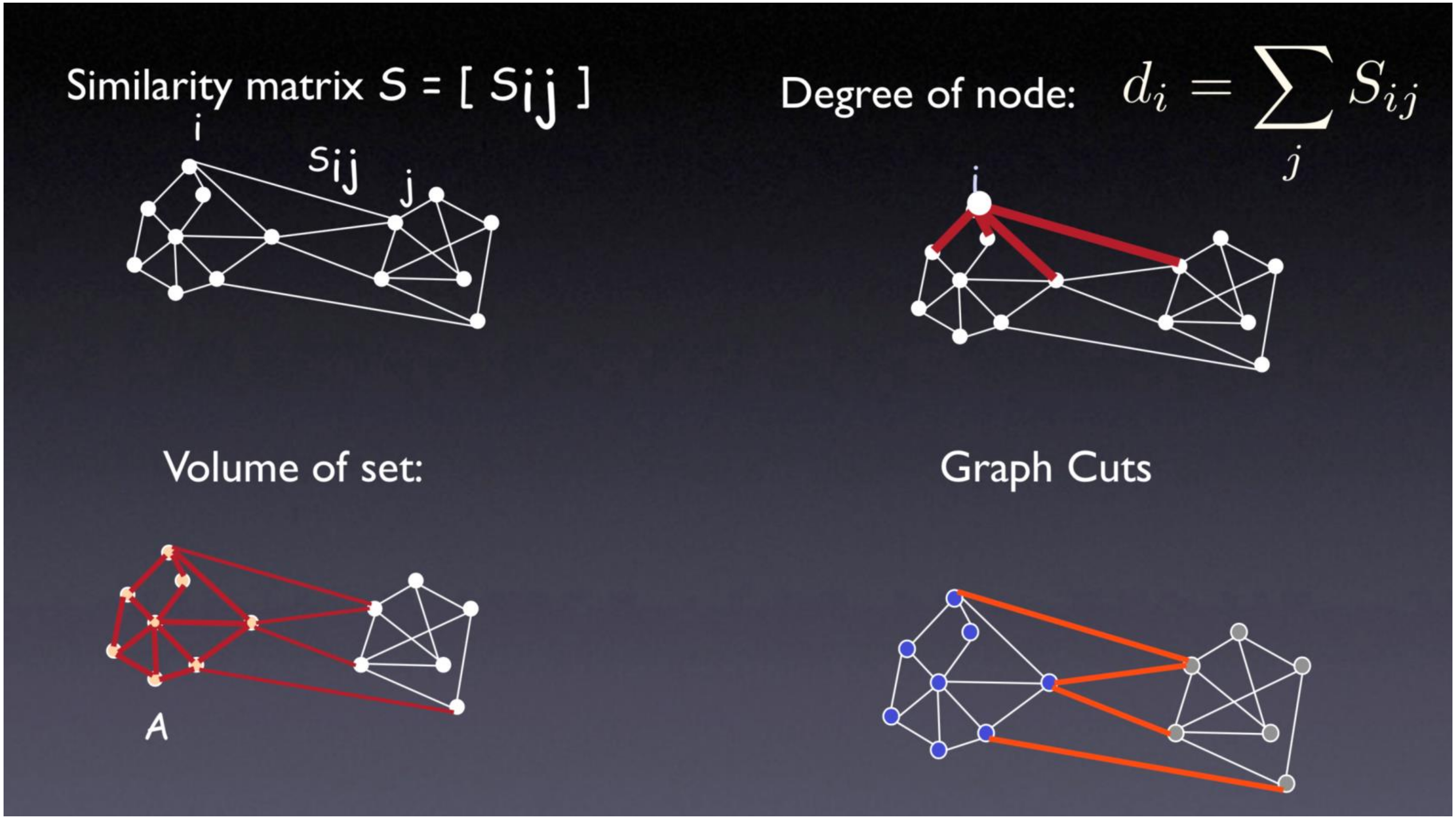
$$\text{cut}(A,B) := \sum_{i \in A, j \in B} w_{ij}$$



		x_1	x_2	x_3	x_4	x_5	x_6
A	x_1	0	0.8	0.6	0	0.1	0
	x_2	0.8	0	0.8	0	0	0
	x_3	0.6	0.8	0	0.2	0	0
	x_4	0	0	0.2	0	0.8	0.7
B	x_5	0.1	0	0	0.8	0	0.8
	x_6	0	0	0	0.7	0.8	0

Similarity matrix

Graph Terminology



Graph Laplacian matrix

- Un-normalized Graph Laplacian

$$L = D - W$$

	x_1	x_2	x_3	x_4	x_5	x_6
x_1	1.5	-0.8	-0.6	0	-0.1	0
x_2	-0.8	1.6	-0.8	0	0	0
x_3	-0.6	-0.8	1.6	-0.2	0	0
x_4	0	0	-0.2	1.7	-0.8	-0.7
x_5	-0.1	0	0	-0.8	1.7	-0.8
x_6	0	0	0	-0.7	-0.8	1.5

=

	x_1	x_2	x_3	x_4	x_5	x_6
x_1	1.5	0	0	0	0	0
x_2	0	1.6	0	0	0	0
x_3	0	0	1.6	0	0	0
x_4	0	0	0	1.7	0	0
x_5	0	0	0	0	1.7	0
x_6	0	0	0	0	0	1.5

−

	x_1	x_2	x_3	x_4	x_5	x_6
x_1	0	0.8	0.6	0	0.1	0
x_2	0.8	0	0.8	0	0	0
x_3	0.6	0.8	0	0.2	0	0
x_4	0	0	0.2	0	0.8	0.7
x_5	0.1	0	0	0.8	0	0.8
x_6	0	0	0	0.7	0.8	0

Graph Laplacian matrix

- Important properties:
  - Symmetric, positive semi-definite

    A symmetric matrix M is positive semidefinite (PSD) *if* $\forall x \in \mathbb{R}^n$,

    $$x^T M x \geq 0.$$

    - Eigenvalues are non-negative real numbers $0 = \lambda_i \leq \lambda_2 \leq \cdots \leq \lambda_n$
    - Eigenvectors are real and orthogonal
    - Smallest eigenvalue is 0, corresponding eigenvector is 1.

    $$L\mathbf{1} = D\mathbf{1} - W\mathbf{1} = \begin{bmatrix} d_1 \\ d_2 \\ \vdots \\ d_n \end{bmatrix} - \begin{bmatrix} \sum_j w_{1j} \\ \sum_j w_{2j} \\ \vdots \\ \sum_j w_{nj} \end{bmatrix} = 0$$

    - Not full rank:
      - if L is an $n \times n$ matrix, the rank of L plus the nullity of L is equal to n
      - *At most n-1*

Graph Laplacian matrix

- Important properties:
  - Symmetric, positive semi-definite
  - Eigenvalues and eigenvectors provide an insight into the connectivity of the graph
    - The dimension of the null space of L is exactly the number of connected components.
    - The number of eigenvalues equal to zero for the Laplacian is equal to the number of connected components in the graph.
    - The rank of Laplacian of a graph is number of vertices minus number of connected components.

Connected Component

- Connected A subset A of a graph is connected if any two vertices in A can be joined by a path such that all intermediate points also lie in A.

- Connected Component it is connected and if there are no connections between vertices in A and $\overline{A}$. The nonempty sets $A_1, \dots, A_k$ form a partition of the graph if $A_i \cap A_j = \emptyset$ and $A_1 \cup \cdots \cup A_k = V$.

Variants

- Normalized Graph Laplacian

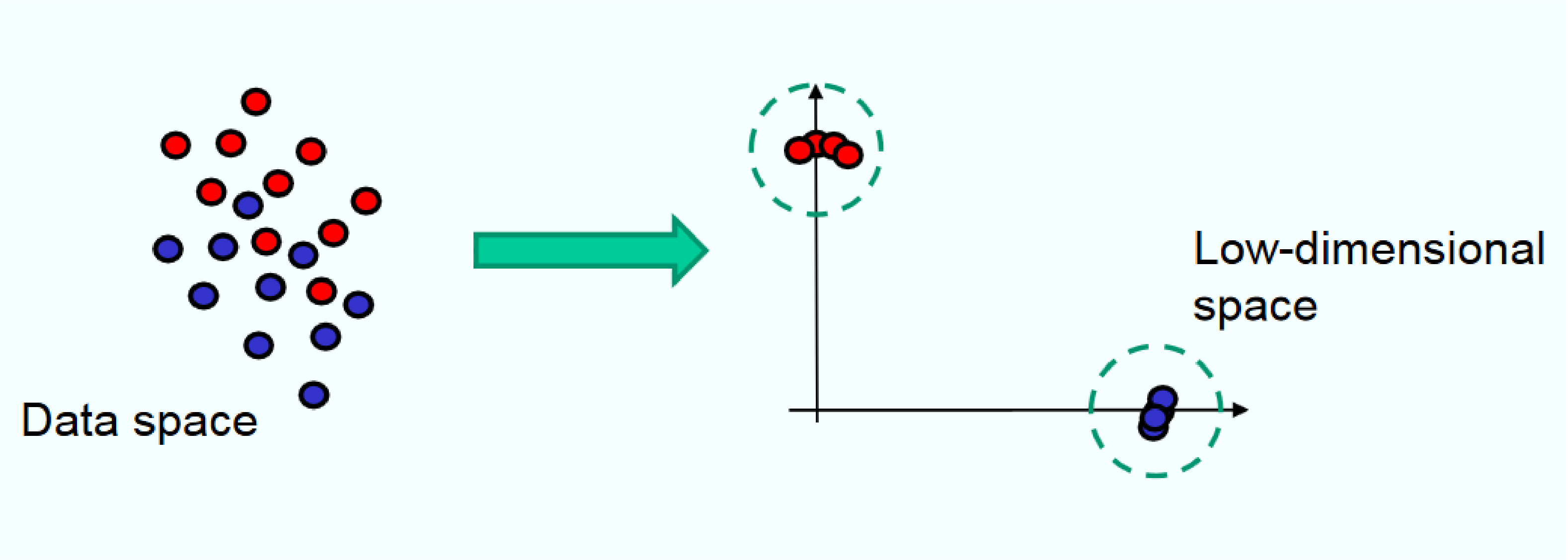
$$L_{sym} := D^{-\frac{1}{2}} L D^{-\frac{1}{2}} = I - D^{-\frac{1}{2}} W D^{-\frac{1}{2}}$$

$$L_{rw} := D^{-1} L = I - D^{-1} W$$

We denote the first matrix by L_{sym} as it is a symmetric matrix, and the second one by L_{rw} as it is closely related to a random walk.

Low-dimensional Embedding Point of View

- Basic Idea: Low-dimensional embedding point of view
  - Obtain data representation in the low-dimensional space that can be easily clustered



Low-dimensional Embedding Point of View

- Given the similarity matrix W,

$$\min \sum_{i,j}^{n} w_{ij}^2 (f_i - f_j)^2$$

  - f_i is the low-dimensional representation
  - Larger w_{ij} enforces f_i and f_j more similar
  - Two similar points in the original space will be similar in the low-dimensional space

Low-dimensional Embedding Point of View

- Reformulation:

$$f'Lf = f'Df - f'Wf = \sum_{i=1}^{n} d_i f_i^2 - \sum_{i,j=1}^{n} f_i f_j w_{ij}$$

$$= \frac{1}{2}\left(\sum_{i=1}^{n} d_i f_i^2 - 2\sum_{i,j=1}^{n} f_i f_j w_{ij} + \sum_{j=1}^{n} d_j f_j^2\right) = \frac{1}{2}\sum_{i,j=1}^{n} w_{ij}\left(f_i - f_j\right)^2$$

$$\min \sum_{i,j}^{n} w_{ij}^2 \left(f_i - f_j\right)^2 \Rightarrow \min f'Lf$$

  - Where un-normalized graph Laplacian $L = D - W$

How to optimize f?

Low-dimensional Embedding Point of View

- How to obtain low-dimensional embedding?

$$L = D - W$$

Multiplicity of eigenvalue 0 = number k of connected components $A_1, \ldots, A_k$ of the graph.

Ky Fan's Theorem $$\sum_{i=1}^{c} \sigma_i(L_S) = \min_{F \in \mathbb{R}^{n \times c}, F^T F = I} Tr(F^T L_S F)$$

The optimal solution F to the problem is formed by the c eigenvectors of L corresponding to the c smallest eigenvalues.

Low-dimensional Embedding Point of View

- Un-normalized Graph Laplacian

$$L = D - W$$

Unnormalized spectral clustering

Input: Similarity matrix $S \in \mathbb{R}^{n \times n}$, number k of clusters to construct.

- Construct a similarity graph by one of the ways described in Section 2. Let W be its weighted adjacency matrix.
- Compute the unnormalized Laplacian L.
- **Compute the first k eigenvectors $u_1, \ldots, u_k$ of L.**
- Let $U \in \mathbb{R}^{n \times k}$ be the matrix containing the vectors $u_1, \ldots, u_k$ as columns.
- For $i = 1, \ldots, n$, let $y_i \in \mathbb{R}^k$ be the vector corresponding to the i-th row of U.
- Cluster the points $(y_i)_{i=1,\ldots,n}$ in $\mathbb{R}^k$ with the k-means algorithm into clusters $C_1, \ldots, C_k$.

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

Variants

Main trick is to change the representation of the abstract data points x_i to points $y_i \in \Re^k$: unfold the manifold

1. Un-normalized Spectral Clustering
2. Normalized Spectral Clustering 1
3. Normalized Spectral Clustering 2

Variants

- Normalized Graph Laplacian $L_{rw} := D^{-1}L = I - D^{-1}W$

Normalized spectral clustering according to Shi and Malik (2000)

Input: Similarity matrix $S \in \mathbb{R}^{n \times n}$, number k of clusters to construct.

- Construct a similarity graph by one of the ways described in Section 2. Let W be its weighted adjacency matrix.
- Compute the unnormalized Laplacian L.
- **Compute the first k generalized eigenvectors $u_1, \dots, u_k$ of the generalized eigenproblem $Lu = \lambda Du$.**
- Let $U \in \mathbb{R}^{n \times k}$ be the matrix containing the vectors $u_1, \dots, u_k$ as columns.
- For $i = 1, \dots, n$, let $y_i \in \mathbb{R}^k$ be the vector corresponding to the i-th row of U.
- Cluster the points $(y_i)_{i=1,\dots,n}$ in $\mathbb{R}^k$ with the k-means algorithm into clusters $C_1, \dots, C_k$.

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

Proposition 3 (Properties of L_{sym} and L_{rw}) *The normalized Laplacians satisfy the following properties:*

3. λ is an eigenvalue of L_{rw} with eigenvector u if and only if λ and u solve the generalized eigenproblem $Lu = \lambda Du$.

Variants

- Normalized Graph Laplacian

$$L_{sym} := D^{-\frac{1}{2}} L D^{-\frac{1}{2}} = I - D^{-\frac{1}{2}} W D^{-\frac{1}{2}}$$

Normalized spectral clustering according to Ng, Jordan, and Weiss (2002)

Input: Similarity matrix $S \in \mathbb{R}^{n \times n}$, number k of clusters to construct.

- Construct a similarity graph by one of the ways described in Section 2. Let W be its weighted adjacency matrix.
- Compute the normalized Laplacian L_{sym}.
- Compute the first k eigenvectors $u_1, \ldots, u_k$ of L_{sym}.
- Let $U \in \mathbb{R}^{n \times k}$ be the matrix containing the vectors $u_1, \ldots, u_k$ as columns.
- Form the matrix $T \in \mathbb{R}^{n \times k}$ from U by normalizing the rows to norm 1, that is set $t_{ij} = u_{ij} / (\sum_k u_{ik}^2)^{1/2}$.
- For $i = 1, \ldots, n$, let $y_i \in \mathbb{R}^k$ be the vector corresponding to the i-th row of T.
- Cluster the points $(y_i)_{i=1,\ldots,n}$ with the k-means algorithm into clusters $C_1, \ldots, C_k$.

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

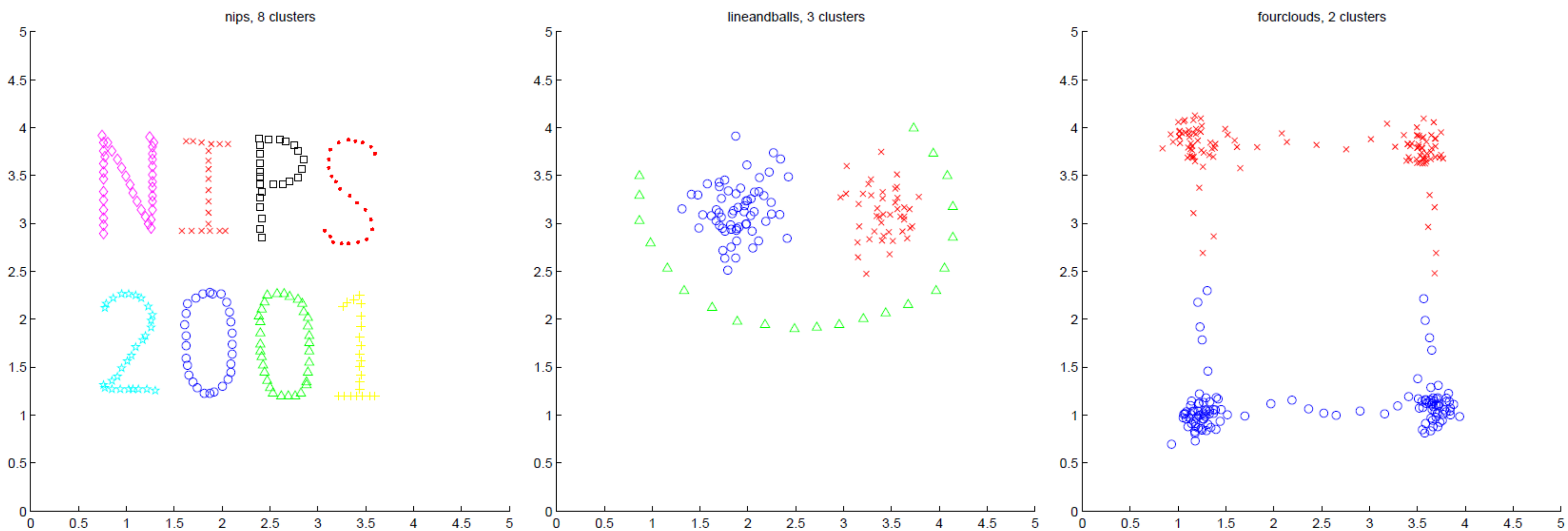
Variants

On Spectral Clustering: Analysis and an algorithm

Andrew Y. Ng
CS Division
U.C. Berkeley
ang@cs.berkeley.edu

Michael I. Jordan
CS Div. & Dept. of Stat.
U.C. Berkeley
jordan@cs.berkeley.edu

Yair Weiss
School of CS & Engr.
The Hebrew Univ.
yweiss@cs.huji.ac.il



(a) (b) (c)