

Similarity-Based Clustering

175

- Similarity matrix W : $W_{i,j}$ is large if x_i and x_j are similar.
- Assumptions on W :
 - Symmetric: $W_{i,j} = W_{j,i}$
 - Positive entries: $W_{i,j} \geq 0$
 - Invertible: $\exists W^{-1}$

Examples of Similarity Matrix¹⁷⁶

$$W_{i,j} = W(\mathbf{x}_i, \mathbf{x}_j)$$

■ Distance-based:

$$W(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\gamma \|\mathbf{x}_i - \mathbf{x}_j\|^2)$$

■ Nearest-neighbor-based:

$W(\mathbf{x}_i, \mathbf{x}_j) = 1$ if \mathbf{x}_i is a k' -nearest neighbor of \mathbf{x}_j or \mathbf{x}_j is a k' -nearest neighbor of \mathbf{x}_i .
Otherwise $W(\mathbf{x}_i, \mathbf{x}_j) = 0$.

■ Combination of two is also possible.

$$W(\mathbf{x}_i, \mathbf{x}_j) = \begin{cases} \exp(-\gamma \|\mathbf{x}_i - \mathbf{x}_j\|^2) \\ 0 \end{cases}$$

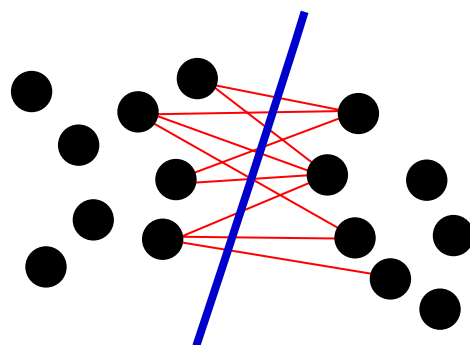
Cut Criterion

■ **Idea:** Minimize sum of similarities between samples inside and outside the class

■ For two classes:

$$\min_{\mathcal{C}_1, \mathcal{C}_2} \left[\sum_{x \in \mathcal{C}_1} \sum_{x' \in \mathcal{C}_2} W(x, x') + \sum_{x \in \mathcal{C}_2} \sum_{x' \in \mathcal{C}_1} W(x, x') \right]$$

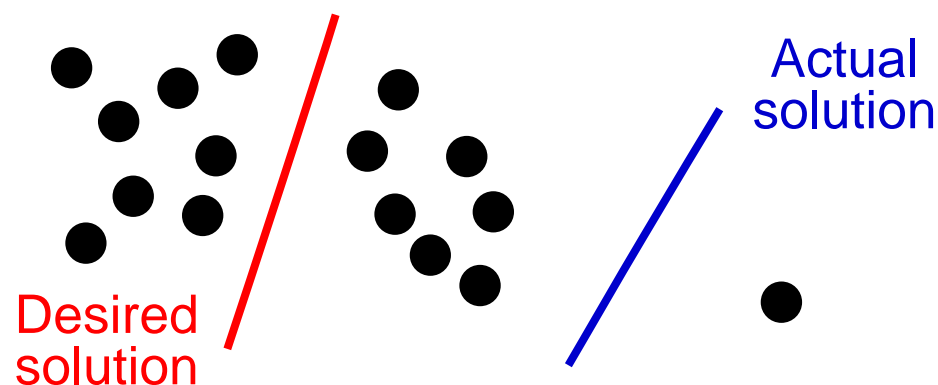
■ From a graph-theoretic viewpoint, this corresponds to finding **minimum cut**.



Cut Criterion (cont.)

$$\min_{\mathcal{C}_1, \mathcal{C}_2} \left[\sum_{x \in \mathcal{C}_1} \sum_{x' \in \mathcal{C}_2} W(x, x') + \sum_{x \in \mathcal{C}_2} \sum_{x' \in \mathcal{C}_1} W(x, x') \right]$$

- Mincut method tends to give a cluster with a **very small number of samples**.



Normalized Cut Criterion

- We penalize small clusters: "Normalized cut"
- For two classes,

$$\min_{\mathcal{C}_1, \mathcal{C}_2} \left[\frac{\sum_{\mathbf{x} \in \mathcal{C}_1} \sum_{\mathbf{x}' \in \mathcal{C}_2} W(\mathbf{x}, \mathbf{x}')}{n} + \frac{\sum_{\mathbf{x} \in \mathcal{C}_2} \sum_{\mathbf{x}' \in \mathcal{C}_1} W(\mathbf{x}, \mathbf{x}')}{n} \right]$$

$$\left[\frac{\sum_{\mathbf{x}'' \in \mathcal{C}_1} \sum_{j=1} W(\mathbf{x}'', \mathbf{x}_j)}{\sum_{\mathbf{x}'' \in \mathcal{C}_1} \sum_{j=1} W(\mathbf{x}'', \mathbf{x}_j)} + \frac{\sum_{\mathbf{x}'' \in \mathcal{C}_2} \sum_{j=1} W(\mathbf{x}'', \mathbf{x}_j)}{\sum_{\mathbf{x}'' \in \mathcal{C}_2} \sum_{j=1} W(\mathbf{x}'', \mathbf{x}_j)} \right]$$

- Denominator is a normalization factor, which is the sum of similarities between samples inside the class and all samples.

Normalized Cut Criterion (cont.)¹⁸⁰

- For k classes, normalized cut is defined as

$$\operatorname{argmin}_{\{\mathcal{C}_i\}_{i=1}^k} [J_{Ncut}]$$

$$J_{Ncut} = \sum_{i=1}^k \left[\frac{\sum_{\mathbf{x} \in \mathcal{C}_i} \sum_{\mathbf{x}' \notin \mathcal{C}_i} W(\mathbf{x}, \mathbf{x}')}{\sum_{\mathbf{x}'' \in \mathcal{C}_i} \sum_{j=1}^n W(\mathbf{x}'', \mathbf{x}_j)} \right]$$

Lemma

■ Let $A_{i,j} = \begin{cases} 1 & \text{if } x_j \in \mathcal{C}_i \\ 0 & \text{o.w.} \end{cases}$, $A = (a_1 | a_2 | \cdots | a_k)^\top$

■ Then

$$J_{Ncut} = \sum_{i=1}^k \frac{\langle La_i, a_i \rangle}{\langle Da_i, a_i \rangle} \quad \begin{aligned} L &= D - W \\ D &= \text{diag}(\sum_{j=1}^n W_{i,j}) \end{aligned}$$

■ Proof:

$$\sum_{x'' \in \mathcal{C}_i} \sum_{j=1}^n W(x'', x_j) = \langle W a_i, \mathbf{1} \rangle = \langle D a_i, a_i \rangle$$

$$\begin{aligned} \sum_{x \in \mathcal{C}_i} \sum_{x' \notin \mathcal{C}_i} W(x, x') &= \sum_{j \neq i} \langle W a_i, a_j \rangle = \langle W a_i, \mathbf{1} - a_i \rangle \\ &= \langle D a_i, a_i \rangle - \langle W a_i, a_i \rangle = \langle L a_i, a_i \rangle \end{aligned}$$

Equivalence (1)

■ Recall

$$J_{WS} = \sum_{i=1}^k \sum_{\mathbf{x} \in \mathcal{C}_i} d(\mathbf{x}) \|\phi(\mathbf{x}) - \mu_i\|^2$$

$$\mu_i = \frac{1}{s_i} \sum_{\mathbf{x}' \in \mathcal{C}_i} d(\mathbf{x}') \phi(\mathbf{x}') \quad s_i = \sum_{\mathbf{x} \in \mathcal{C}_i} d(\mathbf{x})$$

■ We have

$$\operatorname{argmin}_{\{\mathcal{C}_i\}_{i=1}^k} [J_{Ncut}] = \operatorname{argmin}_{\{\mathcal{C}_i\}_{i=1}^k} [J_{WS}]$$

with

$$d(\mathbf{x}) = \sum_{i=1}^n W(\mathbf{x}, \mathbf{x}_i) \quad K(\mathbf{x}_i, \mathbf{x}_j) = [\mathbf{D}^{-1} \mathbf{W} \mathbf{D}^{-1}]_{i,j}$$

Proof

$$\begin{aligned}
 J_{WS} &= \sum_{i=1}^k \sum_{\mathbf{x} \in \mathcal{C}_i} \left(d(\mathbf{x})K(\mathbf{x}, \mathbf{x}) \right. \\
 &\quad \left. - \frac{2}{s_i} d(\mathbf{x}) \sum_{\mathbf{x}' \in \mathcal{C}_i} d(\mathbf{x}')K(\mathbf{x}, \mathbf{x}') \right. \\
 &\quad \left. + \frac{1}{s_i^2} d(\mathbf{x}) \sum_{\mathbf{x}', \mathbf{x}'' \in \mathcal{C}_i} d(\mathbf{x}')d(\mathbf{x}'')K(\mathbf{x}', \mathbf{x}'') \right) \\
 &= \sum_{j=1}^n d(\mathbf{x}_j)K(\mathbf{x}_j, \mathbf{x}_j) \\
 &\quad - \sum_{i=1}^k \frac{1}{s_i} \sum_{\mathbf{x}', \mathbf{x}'' \in \mathcal{C}_i} d(\mathbf{x}')d(\mathbf{x}'')K(\mathbf{x}', \mathbf{x}'')
 \end{aligned}$$

Proof (cont.)

■ Using $\{\mathbf{a}_i\}_{i=1}^k$, we have

- $s_i = \langle \mathbf{D}\mathbf{a}_i, \mathbf{a}_i \rangle$
- $\sum_{\mathbf{x}', \mathbf{x}'' \in \mathcal{C}_i} d(\mathbf{x}')d(\mathbf{x}'')K(\mathbf{x}', \mathbf{x}'') = \langle \mathbf{D}\mathbf{K}\mathbf{D}\mathbf{a}_i, \mathbf{a}_i \rangle = \langle \mathbf{W}\mathbf{a}_i, \mathbf{a}_i \rangle$

■ Therefore,

$$\begin{aligned}
 \operatorname{argmin}_{\{\mathcal{C}_i\}_{i=1}^k} [J_{WS}] &= \operatorname{argmin}_{\{\mathcal{C}_i\}_{i=1}^k} \left[- \sum_{i=1}^k \frac{\langle \mathbf{W}\mathbf{a}_i, \mathbf{a}_i \rangle}{\langle \mathbf{D}\mathbf{a}_i, \mathbf{a}_i \rangle} \right] \\
 &= \operatorname{argmin}_{\{\mathcal{C}_i\}_{i=1}^k} \left[\sum_{i=1}^k \frac{\langle \mathbf{D}\mathbf{a}_i, \mathbf{a}_i \rangle - \langle \mathbf{W}\mathbf{a}_i, \mathbf{a}_i \rangle}{\langle \mathbf{D}\mathbf{a}_i, \mathbf{a}_i \rangle} \right] \\
 &= \operatorname{argmin}_{\{\mathcal{C}_i\}_{i=1}^k} [J_{Ncut}]
 \end{aligned}$$

Solution (1)

- Clustering based on the normalized cut criterion can be obtained by **weighted kernel k-means algorithm** with

$$d(\mathbf{x}) = \sum_{i=1}^n W(\mathbf{x}, \mathbf{x}_i) \quad K(\mathbf{x}_i, \mathbf{x}_j) = [\mathbf{D}^{-1} \mathbf{W} \mathbf{D}^{-1}]_{i,j}$$

1. Randomly initialize partition: $\{\mathcal{C}_i\}_{i=1}^k$
2. Update class assignments until convergence:

$$\mathbf{x}_j \rightarrow \mathcal{C}_t$$

$$t = \operatorname{argmax}_i \left[2s_i \sum_{\mathbf{x}' \in \mathcal{C}_i} d(\mathbf{x}') K(\mathbf{x}_j, \mathbf{x}') - \sum_{\mathbf{x}', \mathbf{x}'' \in \mathcal{C}_i} d(\mathbf{x}') d(\mathbf{x}'') K(\mathbf{x}', \mathbf{x}'') \right]$$

Equivalence (2)

$$\operatorname{argmin}_{\{\mathcal{C}_i\}_{i=1}^k} [J_{Ncut}] \quad J_{Ncut} = \sum_{i=1}^k \frac{\langle \mathbf{L} \mathbf{a}_i, \mathbf{a}_i \rangle}{\langle \mathbf{D} \mathbf{a}_i, \mathbf{a}_i \rangle}$$

■ Solution is given by

$$\operatorname{argmin}_{\mathbf{A} \in \mathbb{R}^{n \times k}} \left[\operatorname{tr}(\mathbf{A} \mathbf{L} \mathbf{A}^\top) \right]$$

$$\text{subject to } \mathbf{A} \mathbf{D} \mathbf{A}^\top = \mathbf{I}_k$$

$$A_{i,j} = \begin{cases} 1 & \text{if } \mathbf{x}_j \in \mathcal{C}_i \\ 0 & \text{o.w.} \end{cases}$$

Proof

$$J_{Ncut} = \sum_{i=1}^k \frac{\langle \mathbf{L} \mathbf{a}_i, \mathbf{a}_i \rangle}{\langle \mathbf{D} \mathbf{a}_i, \mathbf{a}_i \rangle}$$

- J_{Ncut} is invariant under scale of \mathbf{a}_i .
- Let us rescale \mathbf{a}_i as

$$\mathbf{a}_i \longleftarrow \frac{\mathbf{a}_i}{\sqrt{\langle \mathbf{D} \mathbf{a}_i, \mathbf{a}_i \rangle}}$$

which is equivalent to impose $\langle \mathbf{D} \mathbf{a}_i, \mathbf{a}_i \rangle = 1$.

- Then $J_{Ncut} = \text{tr}(\mathbf{A} \mathbf{L} \mathbf{A}^\top)$
- Since $\langle \mathbf{D} \mathbf{a}_i, \mathbf{a}_j \rangle = 0$ for $i \neq j$, we have

$$\mathbf{A} \mathbf{D} \mathbf{A}^\top = \mathbf{I}_k$$

Relation to Laplacian Eigenmap¹⁸⁹

- Let us allow A to take real values.
- Then relaxed problem is given as

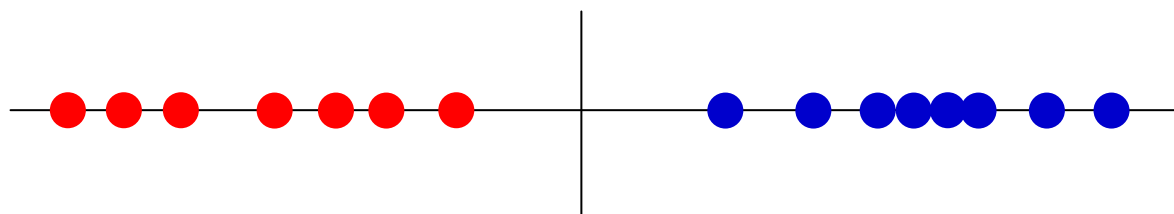
$$\min_{A \in \mathbb{R}^{n \times k}} \left[\text{tr}(A L A^\top) \right]$$

subject to $A D A^\top = I_k$

- Equivalent to Laplacian eigenmap criterion!
- Therefore, Laplacian eigenmap embedding may have a clustering property!

Solution (2)

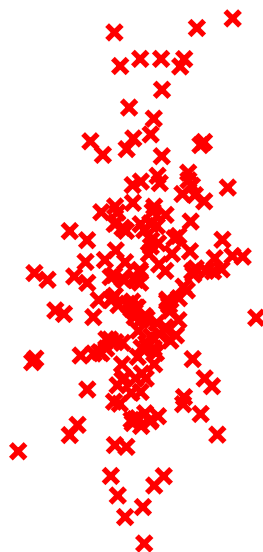
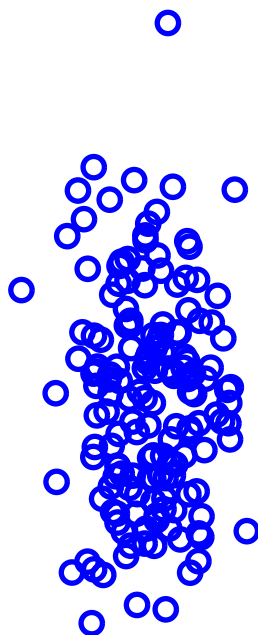
- Clustering into 2 classes:
 1. Embed the data into one-dimensional space by Laplacian eigenmap
 2. Cluster the embedded data by thresholding



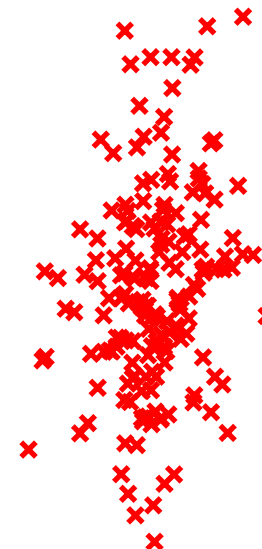
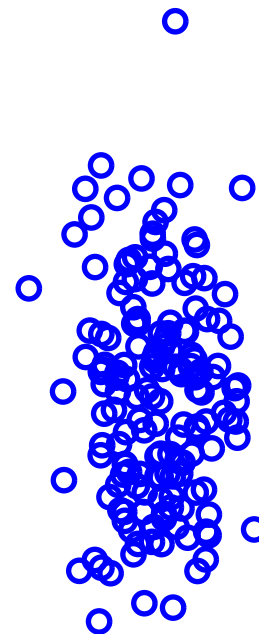
Solution (2)

- For more than 2 classes, cluster assignment may be obtained as follows.
 1. Embed the samples into k-dimensional space by Laplacian eigenmap.
 2. Run ordinary k-means algorithms for the embedded samples.
- This method is called **spectral clustering**.

Examples (1)

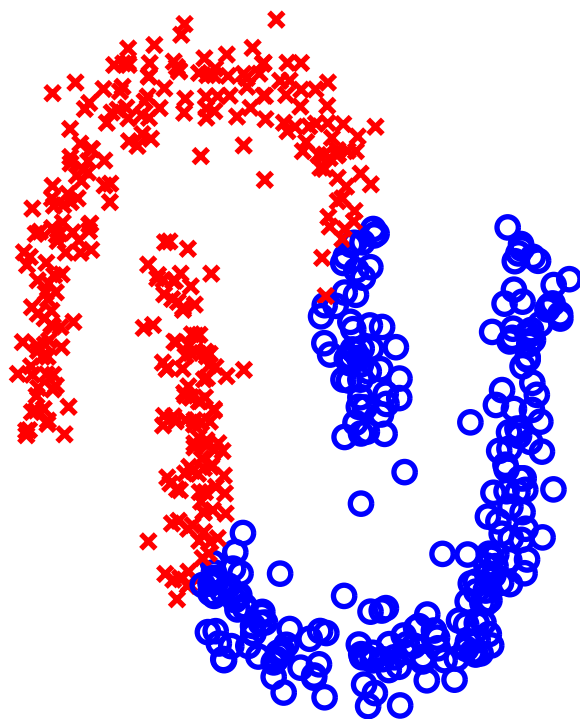


K-means clustering

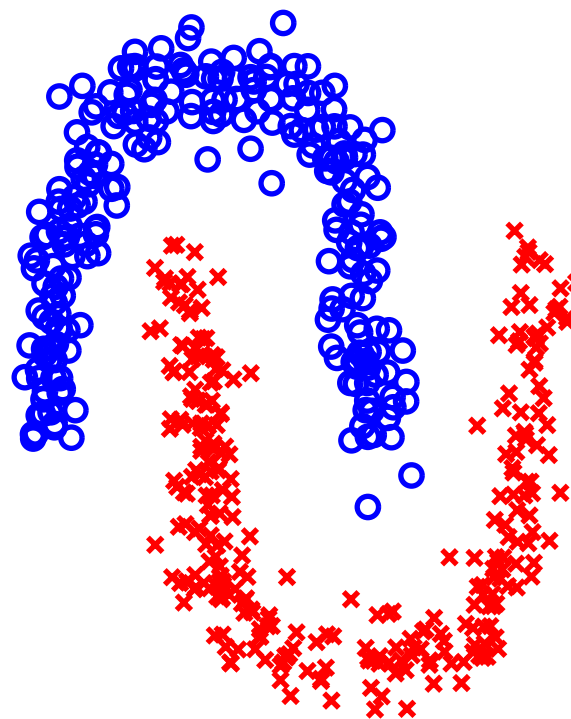


Spectral clustering

Examples (2)

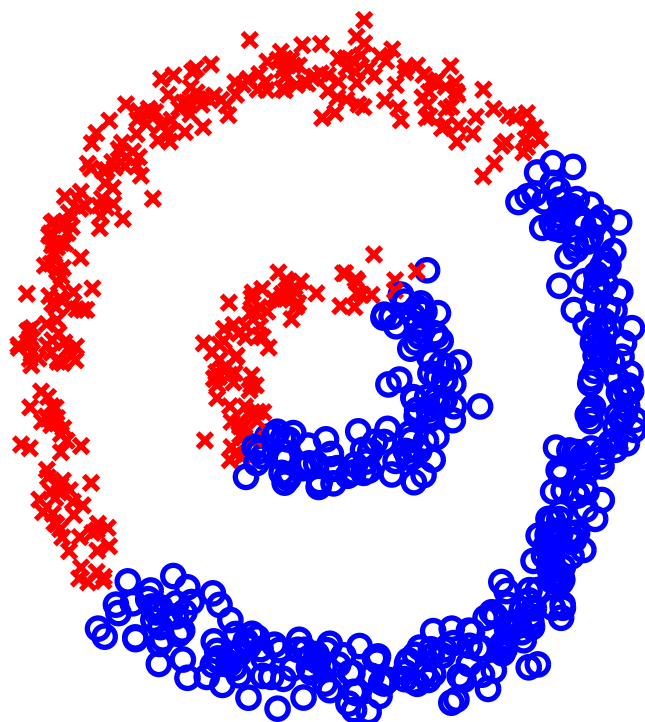


K-means clustering

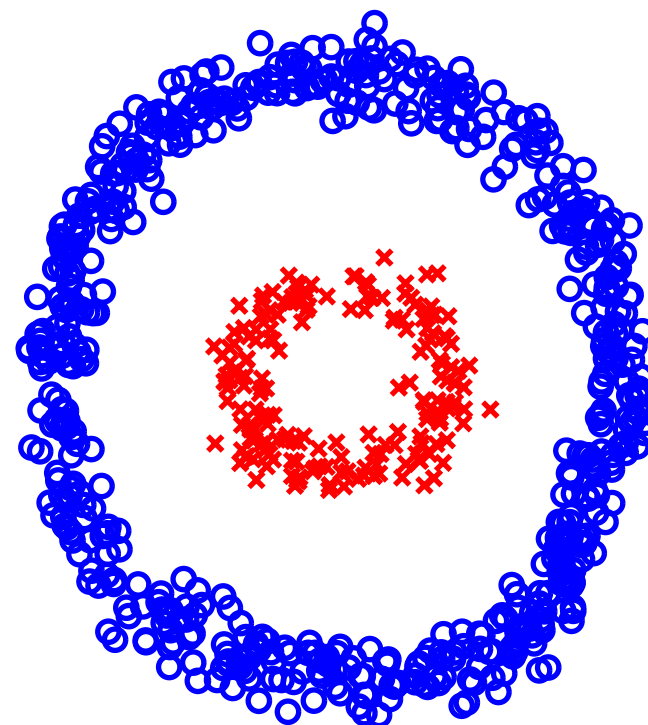


Spectral clustering

Examples (3)



K-means clustering



Spectral clustering

Spectral Graph Theory

- **Spectral graph theory** studies relationships between the properties of a graph and its adjacency matrix.
- **Graph**: A set of vertices and edges
- **Adjacency matrix** \mathbf{W} : $\mathbf{W}_{i,j}$ is the number of edges from i -th to j -th vertices.
- **Vertex degree** d_i : Number of connected edges
- **Graph Laplacian** \mathbf{L} :

$$\mathbf{L}_{i,j} = \begin{cases} d_i & (i = j) \\ -1 & (i \neq j \text{ \& } \mathbf{W}_{i,j} > 0) \\ 0 & (\text{o.w.}) \end{cases}$$

Relation to Spectral Graph Theory¹⁹⁸

- Suppose our similarity matrix W is defined by nearest neighbors.
- Consider the following graph:
 - Each vertex corresponds to each point x_i
 - Edge exists if $W_{i,j} > 0$
- W is the adjacency matrix.
- D is the diagonal matrix of vertex degrees.
- L is the graph Laplacian.

Suggestion

- If you are interested in spectral graph theory, the following book would be interesting.

Chung, F. R. K., *Spectral Graph Theory*, American Mathematical Society, Providence, R.I., 1997.