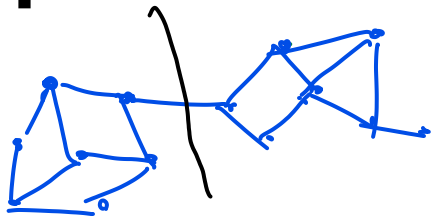
Spectral Clustering

Data →  → Laplacian → Transformed data

Similarity Graph

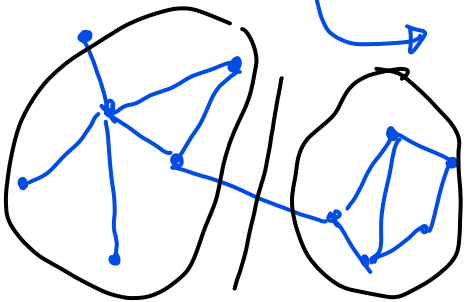
K-mean → $C_1, \ldots C_K$ clusters

Problem of Spectral Graph clustering

Data: $x_1, x_2, \ldots, x_n$

Similarity Notion: $s_{ij} \geq 0$ between two points x_i & x_j

Similarity Graph: $G(V, E)$



$V = \{x_1, \ldots, x_n\} \rightarrow$ Vertex set

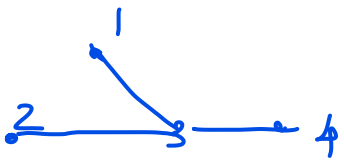
$E = \{E_{12}, \ldots, E_{np}\} \rightarrow$ Edge set

$W = \{W_{12}, \ldots W_{np}\} \rightarrow$ Weight set

Clustering: Find a partition of the graph such that the edges between different groups have very low weights and within a group have high weights.

* Adjacency matrix: $A_{ij} = 1$ if $E_{ij} \in E$

$(n \times n)$ $\qquad = 0$ if $E_{ij} \notin E$



$$\begin{array}{c|cccc} & 1 & 2 & 3 & 4 \\ \hline 1 & 0 & 0 & 1 & 0 \\ 2 & 0 & 0 & 1 & 0 \\ 3 & 1 & 1 & 0 & 1 \\ 4 & 0 & 0 & 1 & 0 \end{array}$$

* weighted adjacency matrix $W_{ij} = w_{ij}$ if $E_{ij} \in E$

$= 0$ if $E_{ij} \notin E$

Note: Undirected graph: $w_{ij} = w_{ji}$

* Degree of a vertex $v_i \in V$, $d_i = \sum_{j=1}^{n-1} w_{ij}$

* Degree matrix D: Diagonal matrix

$n \times n$ $$D = \begin{bmatrix} d_1 & & 0 \\ & d_2 & \\ 0 & & \ddots \; d_n \end{bmatrix}$$

* $A \subset V$: a subset of vertices
* $\bar{A} \subset V$: a subset of vertices V/A

* Define an Indicator vector: for A

$$1_A = \begin{bmatrix} f_1 \\ \vdots \\ f_n \end{bmatrix} \qquad \begin{aligned} f_i &= 1 \text{ if } v_i \in A \\ &= 0 \text{ if } v_i \notin A \end{aligned}$$

* For $A, B \subset V$

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

* Measuring the "size" of a subset $A \subset V$

$$|A| = \# \text{ of vertices in } A$$

$$\mathrm{vol}(A) = \sum_{i \in A} d_i$$

* A subset A is <u>connected if any</u> two vertices in A can be joined by a path such that all intermediate points also lie in A.

* Connected component: A if it is connected & No connections between vertices in A & $\bar{A}$.

* Partition of the graph:
For $A_1, \ldots A_k \neq \phi \subset V$ if $A_i \cap A_j = \phi$ and $\bigcup_{i=1}^{k} A_i = V$ then, $A_1 \ldots A_k$ form a partition of V.

* Similarity Graphs:
  - To model the local neighborhood relationships between the data points
* The ϵ-neighborhood Graphs:
  * connect all points whose distance $x_{ij} < \epsilon$
  * An unweighted graph

· K-nearest Neighbor Graphs:

* connect vertices v_i & v_j if v_j is among the k-nearest neighbors of v_i.

* Weight the edges by similarity of their endpoints.

* Graph Laplacians

$G(V, E, W) \rightarrow$ Undirected Graph

* Unnormalized graph Laplacian matrix

$$\boxed{L = D - W}$$

* Properties: For every $f \in \mathbb{R}^n$,

$$f^T L f = \frac{1}{2} \sum_{i,j=1}^{n} w_{ij} (f_i - f_j)^2$$

L: symmetric and positive semi definite

* The smallest eigenvalue of $L = 0$ & corresponding eigen vector $\mathbf{1}$

* L has n- non-negative real-valued Eigen values $0 = \lambda_1 \leq \dots \leq \lambda_n$

* For G with $w_{ij} \geq 0$, the multiplicity K of the eigenvalue 0 of Laplacian graph matrix = # of connected components

* The eigenspace corresponding to 0 EVs. spanned by the indicator vectors $\mathbb{1}_{A_1}, \ldots, \mathbb{1}_{A_k}$

Eigenvector

* $$\begin{bmatrix} L_1 & & \\ & L_2 & \\ & & L_3 \end{bmatrix} \longrightarrow \begin{bmatrix} 1 \\ \vdots \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ \vdots \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ \vdots \\ 1 \end{bmatrix}$$

* Normalized Graph Laplacian:

$$L_{sym} = D^{-1/2} L D^{-1/2} ; \quad L_{rw} = D^{-1} L$$

$D^{1/2} 1_{A_i}$

* Spectral clustering Algorithm

Input: Similarity matrix $S \in \mathbb{R}^{n \times n}$, $k = \#$ of clusters

output: clusters $A_1, \dots, A_k$, $A_i = \{ j \mid y_j \in C_i \}$ y_j : Transformed data.

step 1: $S \longrightarrow S_G$ Similarity graph, ϵ-NBH or KNN Graph

Denote W : Weighted adjacency matrix

step 2: Compute Laplacian L, L_{sys}, or L_{rw}

step 3: Compute the first k eigenvectors $u_1, \dots u_k$ of L

step 4: For $i = 1, \dots, n$ $y_i \in \mathbb{R}^k$, cluster the points $\{y_i\}_{i=1}^{n}$ with k-means algorithm into clusters $C_1, \dots, C_k$

step 5: Assign points to clusters $A_1, \dots, A_k$ for $A_i = \{ j \mid y_j \in C_i \}$