18/3/29

SPECTRAL CLUSTERING

Notation: let $G = (V, E)$ - undirected, weighted graph
$V = \{v_1, \dots, v_n\}$, $v_i \xrightarrow{w_{ij}} v_j$, $w_{ij} \geq 0$

- adjacency matrix: $W = (w_{ij})_{i,j=1,\dots,n}$ $[W = W']$

- degree matrix: $d_i = \sum_{j=1}^{n} w_{ij}$, $D = \mathrm{diag}(d_1, \dots, d_n)$

Similarity Graphs: Transform data $\{x_1, \dots, x_n\}$ with pairwise distance d_{ij} into graph

① ϵ - neighbourhood graph
-- connect all points whose pairwise ~~edges~~ distances are less than ϵ.
- since $w_{ij} \leq \epsilon$, can be considered un-weighted

② k-nearest-neighbour graph:
- connect v_i to v_j if v_i is among k-nearest neighbours of v_j and v_j is among k-nearest neighbors of v_i.
- weight the edges by similarity of v_i, v_j

③ fully connected graphs:
- connect all points with positive "similarity"
- eg: $s_{ij} = s(x_i, x_j) = \exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma^2}\right)$

GRAPH LAPLACIANS

① Unnormalized graph laplacian:

$$L = D - W$$

Proposition 1: ⓐ $f'Lf = \frac{1}{2}\sum_{i,j} w_{ij}(f_i - f_j)^2 \quad f \in \mathbb{R}^n$

ⓑ L - symmetric, p.s.d

ⓒ $\lambda_{min} = 0$, $u_{min} = \mathbb{1} \in \mathbb{R}^n$

ⓓ $0 \leq \lambda_{min} = \lambda_1 \leq \lambda_2 \leq \dots \leq \lambda_n$

Proof: ⓐ $x'Lx = x'Dx - x'Wx$

$$= \sum_i d_i x_i^2 - \sum_{i,j} x_i x_j w_{ij}$$

$$= \frac{1}{2}\left(\sum d_i x_i^2 - 2\sum_{i,j} x_i x_j w_{ij} + \sum d_j x_j^2\right)$$

$$= \frac{1}{2}\left(\sum_i \sum_j w_{ij} x_i^2 - 2\sum_{i,j} x_i x_j w_{ij} + \sum_j \sum_i w_{ji} x_j^2\right)$$

$$= \frac{1}{2}\sum_{i,j} w_{ij}(x_i - x_j)^2$$

ⓑ - D, W are symmetric $\Rightarrow$ L is symmetric

- $x'Lx \geq 0 \ \forall x \Rightarrow L$ is p.s.d

ⓒ - choosing $x = \mathbb{1} \Rightarrow \lambda_{min} = 0$

ⓓ follows from ⓐ, ⓑ, ⓒ. ☒

Proposition 2: The multiplicity 'k' of eigenvalue 0 of L equals the number of connected components $A_1, \dots, A_k$. The eigenspace of 0-eval is spanned by vectors $\mathbb{1}_{A_1}, \dots \mathbb{1}_{A_k}$

Proof: I: Let $k = 1$; $\Rightarrow$ graph is connected. Let x be eigenvector with eigenvalue 0. then

$$x' L x = \frac{1}{2} \sum w_{ij} (x_i - x_j)^2 = 0$$

$\Rightarrow$ x needs to be equal on all nodes which can be connected by a path in G. $\Rightarrow$ $x_i = \mathbb{1}$ $\forall i$

II w.l.o.g, let $L = \begin{pmatrix} L_1 & & 0 \\ & L_2 & \\ 0 & & \ddots & L_k \end{pmatrix}$

- L_i are graph laplacians of A_i
- eigenvectors of L = eigenvectors of L_i with 0's added appropriately
- eigenvalues of L_i = e-values of L. $\Rightarrow$ there are k eigenvalues 0, and the eigenvector is $\mathbb{1}_{A_i}$ [$\mathbb{1}_A = (x_1, \dots, x_n)^T$

$\Leftrightarrow$ $x_i = 1$ if $v_i \in A$]

② Normalized Graph Laplacians:

$$L_{sym} = D^{-1/2} L D^{-1/2} = I - D^{-1/2} W D^{-1/2}$$

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

Proposition 3 : The normalized laplacians satisfy following properties.

ⓐ $x' L_{sym} x = \frac{1}{2} \sum_{i,j} w_{ij} \left(\frac{x_i}{\sqrt{d_i}} - \frac{x_j}{\sqrt{d_j}} \right)^2$

ⓑ λ is an eigenvalue of L_{rw} with e-vector u if and only if λ is an eigenvalue of L_{sym} with e-vector $w = D^{1/2} u$

ⓒ λ is an eigenvalue of L_{rw} with e-vector u iff λ and u solve the generalised eigenproblem $Lu = \lambda D u$

ⓓ 0 is an eigenvalue of L_{rw} with the constant $\mathbb{1}$ as eigenvector. 0 is an eigenvalue of L_{sym} with e-vector $D^{1/2} \mathbb{1}$

ⓔ

Proposition 4 : The multiplicity k of eigenvalue 0 of both L_{rw} and L_{sym} equals the number of connected components $A_1, \dots, A_k$ in G. For L_{rw}, eigenspace of 0 is spanned by $\mathbb{1}_{A_i}$ and for L_{sym}, eigenspace of 0 is spanned by $D^{1/2} \mathbb{1}_{A_i}$

Proof of Prop 3, 4 are similar to that of Props. 1, 2 ⊠

SPECTRAL CLUSTERING ALGORITHMS

Un-normalized spectral clustering

Input : Similarity matrix $S \in \mathbb{R}^{n \times n}$ # clusters K

1. Construct similarity graph with adj. matrix W
2. Compute un-normalized Laplacian L
3. Compute top-k-eigenvectors of L as $V = [v_1, \dots, v_k] \in \mathbb{R}^{n \times k}$
4. Let y_i $(i = 1, \dots, n)$ be i^{th} row of V
5. cluster y_i into clusters $C_1, \dots, C_k$ using k-means.

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

Justification of S.C. ALGORITHM [Using Matrix Perturbation]

- ideally, "inter-cluster similarity" is 0

$$\Rightarrow \quad y_i \in \mathbb{R}^k \text{ are of the form } [0, 0, \dots, 0, \overset{j}{1}, 0, \dots, 0]$$

where j is s.t $y_i \in A_j \Rightarrow x_i \in A_j$

- in a nearly ideal case, "inter-cluster similarity" is ϵ

y_i will be of form $[0, \dots 0, 1, 0, \dots, 0] + \epsilon$

- using davis kahan if $\tilde{A} = A + H$, $S_1 \subset \mathbb{R}$ interval

$\sigma_{S_1}(A)$: set of all eig vals contained in S_1 and V_1 be eig. space corresponding to $\sigma_{S_1}(A)$ and same for $\tilde{A}, \tilde{V}_1$ then

$$\| \sin \theta (V_1, \tilde{V}_1) \| \leq \frac{\|H\|}{\delta}$$

$\delta = \min \{ |\lambda - s| ; \lambda - \text{eigval of } A, \lambda \notin S_1, s \in S_1 \}$