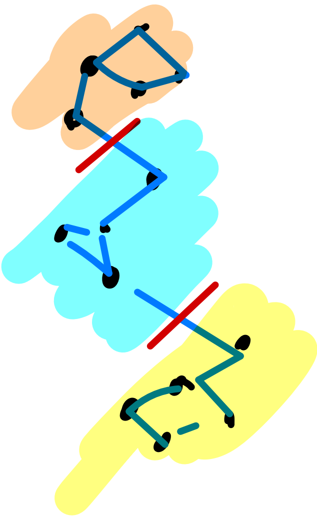
Lecture 9 — Modern Clustering Algorithms (I)

18/11/2024

Spectral Clustering



- transform our data points into a graph
- Find "optimal" cuts
- Clusters are the connected components in our graph

Undirected Weighted Graphs

$G(E, V)$

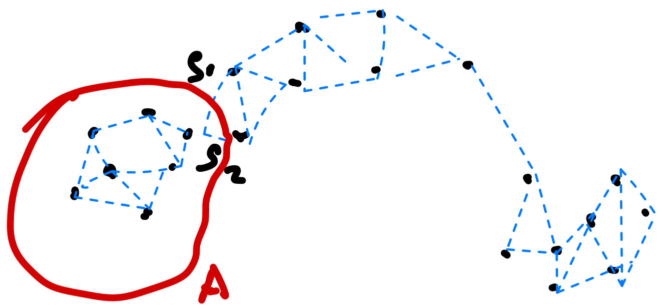
$E\{(i, l), S_{il} \geq \phi\}$

1. Obtain the (Similarity) Graph

- ε- ball graph: Join vertices within a radius ε
- k-NN: Join vertices that are the k-NN (Symmetric)
- Fully connected graph. $S_{ij} = e^{-\frac{d_{ij}^2}{2\sigma^2}}$

Made with Goodnotes

• "Optimal" cut of a graph

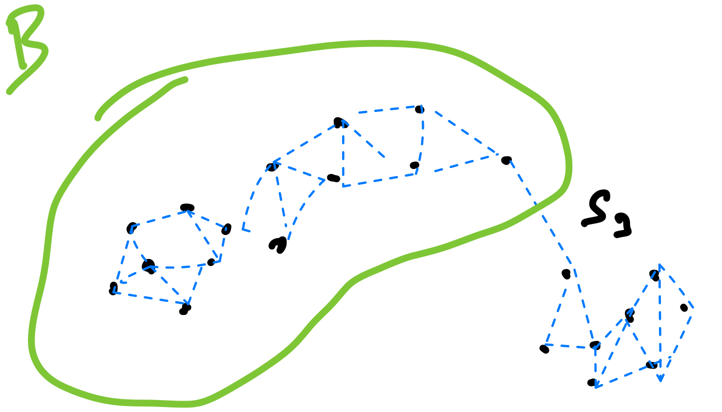


$C \{V_c\}$

$\bar{C} \{V_{\notin c}\}$

$$\mathrm{cut}(C, \bar{C}) = \sum_{i \in C} \sum_{l \in \bar{C}} S_{il}$$

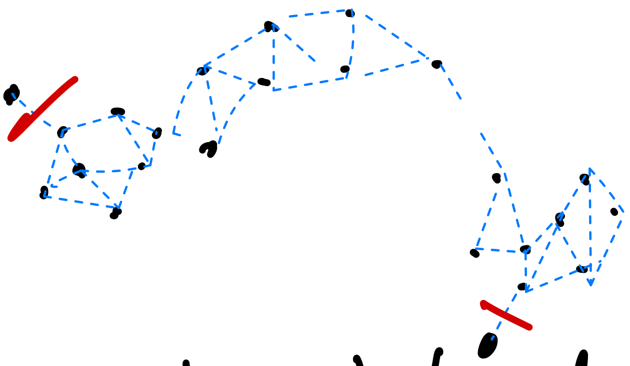
$$\mathrm{Cut}(A, \bar{A}) = S_1 + S_2$$



$$\mathrm{Cut}(B, \bar{B}) = S_3$$

k-way cut: $\mathrm{Cut}(C_1, C_2, C_3 \dots C_k) =$

$$\frac{1}{2} \sum_m \mathrm{Cut}(C_m, \bar{C}_m)$$



Does not take into account the population of the clusters.

Heavily unbalanced clusters!

Take into account the size

① $|C_m| \rightarrow$ Number of vertices of cluster C_m

$$\text{RatioCut}(C_1, \dots C_k) = \sum_m^k \frac{\text{Cut}(C_m, \bar{C}_m)}{|C_m|}$$

② Degree of a point i $\quad g^i = \sum_j S_{ij}$

$$\text{Vol}(C_m) = \sum_{i \in C_m} g^i$$

$$\text{Normalized Cut}(C_1, \dots C_k) = \sum_m^k \frac{\text{Cut}(C_m, \bar{C}_m)}{\text{Vol}(C_m)}$$

Matrices Involved

$\mathbb{S} \rightarrow$ Symm $\quad N \times N \quad$ weight matrix

$$\begin{pmatrix} 0 & & & S_{ij} & \\ & 0 & & & \\ & S_{ij} & 0 & & \\ & & & 0 & \\ & & & & 0 \end{pmatrix}$$

$\mathbb{D}$ Degree matrix $\quad$ Sym $\quad N \times N$

$\mathbb{D}_{ii} = g^i$

$\mathbb{D}_{ij} = \emptyset \quad i \neq j$

$$\begin{pmatrix} g^1 & & & \emptyset \\ & \ddots & & \\ & & g^i & \\ & \emptyset & & \ddots \\ & & & & g^N \end{pmatrix}$$

Laplacian Matrix $\mathbb{L} = \mathbb{D} - \mathbb{S}$

① Symmetric

② Positive semi-definite $\lambda_1 = \emptyset \leq \lambda_2 \leq \lambda_3 \dots$

③ v vector with dimension N

$$v^T \mathbb{L} v = v^T \mathbb{D} v - v^T \mathbb{S} v =$$

$$\underbrace{\sum_i v_i^2 g^i - \sum_{il} v_i v_l S_{il}}_{} = \frac{1}{2} \Big(\sum_i v_i^2 \underbrace{g^i}_{\sum_l S_{il}} -$$

$$2 \sum_{il} v_i v_l S_{il} + \sum_l v_l^2 g^l \Big) = \frac{1}{2} \Big(\sum_{il} \underline{v_i}^2 S_{il}$$

$$- 2 \sum_{il} \underline{v_i v_l} S_{il} + \sum_{il} \underline{v_l}^2 S_{il} = \frac{1}{2} \Big(\sum_{il} S_{il} (v_i - v_l)^2 \Big)$$

The Ratio Cut for $k=2$

Minimize $\dfrac{Cut(C,\bar{C})}{|C|}$

f indicator vector

$$\boxed{\begin{array}{l} f_i = \sqrt{\dfrac{|\bar{C}|}{|C|}} \\ \text{if } i \in C \\ \text{otherwise} \\ f_i = -\sqrt{\dfrac{|C|}{|\bar{C}|}} \end{array}}$$

$$f^T \mathbb{L} f = \sum_{il} \underbrace{S_{il}(f_i - f_l)^2}$$

$$\begin{cases} i \,\&\, l \in C = \phi \\ i \,\&\, l \in \bar{C} = \phi \\ i \in C \,\&\, l \in \bar{C} \ \cancel{} \\ i \in \bar{C} \,\&\, l \in C \ \cancel{} \end{cases}$$

$$f^T \mathbb{L} f = \frac{1}{2}\sum_{il} S_{il}(f_i - f_l)^2 = \frac{1}{2}\sum_{\substack{i \in C \\ l \in \bar{C}}} S_{il}\left(\sqrt{\frac{|\bar{C}|}{|C|}} + \sqrt{\frac{|C|}{|\bar{C}|}}\right)^2$$

$$+ \frac{1}{2}\sum_{\substack{i \in \bar{C} \\ l \in C}} S_{il}\left(-\sqrt{\frac{|\bar{C}|}{|C|}} - \sqrt{\frac{|C|}{|\bar{C}|}}\right)^2 = RatioCut(C,\bar{C})$$

Minimizing the ratioCut is equivalent to minimize $f^T \mathbb{L} f$ given that $\|f\| = \sqrt{N}$

$\&\ f \perp \mathbb{1}$

RatioCut for arbitrary k

k indicator vectors $\rightarrow$ $\mathbb{H}$

$$\mathbb{H}_{il} = \frac{1}{\sqrt{|C_l|}} \text{ if } i \in C_l \, ; \; \mathbb{H}_{il} = \emptyset \text{ if } i \in \overline{C_l}$$

$i = 1 \ldots N$

$l = 1 \ldots k$

$\mathbb{H}^T \mathbb{H} = \mathbb{I}$ $\quad h_l \rightarrow l$ column of $\mathbb{H}$

$$h_l^T \mathbb{L} h_l = \frac{\text{Cut}(C_l, \overline{C_l})}{|C_l|}$$

$$\text{RatioCut}(C_1, \ldots C_k) = \sum_l \frac{\text{Cut}(C_l, \overline{C_l})}{|C_l|} =$$

$$\sum_l (\mathbb{H}^T \mathbb{L} \mathbb{H})_{ll} = \text{Trace}(\mathbb{H}^T \mathbb{L} \mathbb{H})$$

Relaxation :

* Relax the condition of discrete values of the indicators vectors

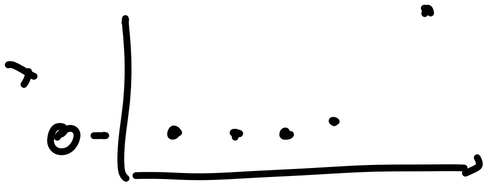
* Minimize $\text{Tr}(\mathbb{H}^T \mathbb{L} \mathbb{H})$ using lagrange multipliers to impose $\mathbb{H}^T \mathbb{H} = \mathbb{I}$.
  Eigenvalue - Eigenvector problem

* There's no warranty that you obtain the same solution that in the unrelax problem

Made with Goodnotes

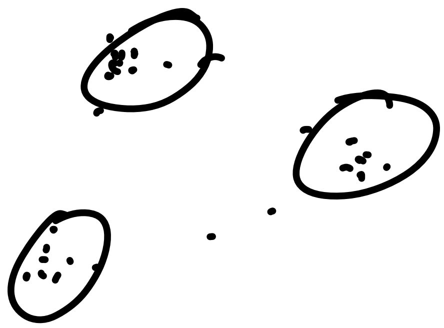
The Unnormalized Spectral clustering algorithm

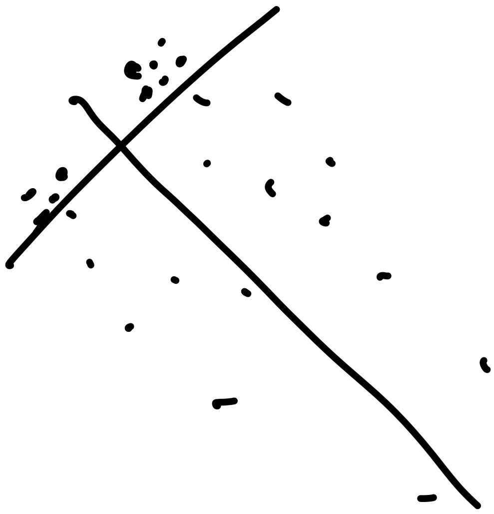
1. Construct the similarity Graph $\mathbb{S}$
2. Compute the laplacian $\mathbb{L} = \mathbb{D} - \mathbb{S}$
3. Compute the first k eigenvectors



$\mathbb{H}$

4. Use the eigenvectors as coordinates of input for k-means clustering





The normalized spectral clustering

minimize $\quad NCut = \sum \frac{Cut(C_\ell, \overline{C_\ell})}{Vol(C_\ell)}$

$\mathbb{H}_{i\ell} = \frac{1}{Vol(C_\ell)}$ if $i \in C_\ell$ otherwise $\mathbb{H}_{i\ell} = \emptyset$

$$\min N_{cut} = \frac{\sum Cut(C_\ell, \overline{C_\ell})}{Vol(C_\ell)} = Trace(\mathbb{H}^T \mathbb{L} \mathbb{H})$$

given $(\mathbb{H}^T \mathbb{D} \mathbb{H}) = \mathbb{I}$

$\underline{\mathbb{T} = (\mathbb{D}^{1/2} \mathbb{H})}$ $\quad \mathbb{T}^T = (\mathbb{D}^{1/2} \mathbb{H}^T)^T$

$$\min Tr(\mathbb{H}^T \mathbb{L} \mathbb{H}) = \min (\mathbb{T}^T \mathbb{D}^{-1/2} \mathbb{L} \mathbb{D}^{-1/2} \mathbb{T})$$

$$\boxed{\mathbb{L}_{sym} = \mathbb{D}^{-1/2} \mathbb{L} \mathbb{D}^{-1/2}}$$

minimize the trace of

$(\mathbb{T}^T \mathbb{L}_{sym} \mathbb{T})$ given $\mathbb{T}^T \mathbb{T} = \mathbb{I}$

- It's closely related to kernel k-means
- If instead of $\mathbb{L}_{sym} = \mathbb{D}^{-1/2} \mathbb{L} \mathbb{D}^{-1/2}$
  - $\mathbb{L}_{rw} = \mathbb{D}^{-1} \mathbb{L}$ → Random walk matrix
- We can infer K from the spectrum

Problems:

1. It depends strongly on the graph construction
2. What happens if the spectrum is not informative?
3. Scaling
4. Noise Sensitivity

Affinity propagation

Based on Message Passing
It determines the exemplars for the clusters

- Responsibility
- Availability

Responsibility : I want to be your exemplar
Availability : I can go with you

Responsibility.

Accumulated evidence that sample k should be the exemplar for sample i

$$r(i,k) \leftarrow s_{ik} - \max_{k' \neq k} \left[s_{ik'} + a(i,k') \right]$$

Availability:
Accumulated evidence that sample i should choose sample k to be its exemplar

$$\text{if } (i \neq k) \quad a[i,k] \leftarrow \min \left[0, \; r(k,k) + \sum_{i' \notin \{k,i\}} \max(0, r(i',k)) \right]$$

$$\text{if} \quad a[k,k] \leftarrow \sum_{i' \neq k} \max\left(0, r(i',k)\right)$$

Algorithm

1. $s_{ik} = -\| x^i - x^k \|^2$

$s_{kk} \equiv$ preference

$s_{kk} \sim \min(s_{ik}) \rightarrow$ few clusters

$s_{kk} \sim \max(s_{ik})$ many clusters

default $s_{kk} = \text{median}(s_{ik})$

2. r & $a = \emptyset$

3. Start iterating by first updating responsabilities

$$r(i,k) \leftarrow s_{ik} - \max_{k' \neq k} [s_{ik'}]$$

$$a(i,k) \leftarrow \min\left[0, r(k,k) + \sum_{i' \notin (i,k)} \max(0, r(i',k))\right]$$

$$r(i,k)_t = s_{ik} - \max_{k' \neq k} (s_{ik'} + a(i,k'))$$

4. Repeat until you reach the number of iterations

Modificate by adding a damping factor:

$$r(i,k)_{t+1} = \lambda \cdot r(i,k)_t + (1-\lambda) \underbrace{r'(i,k)_{t+1}}_{\text{original formula}}$$

Made with Goodnotes