LECTURE 8 FUZZY C-MEANS & HIERARCHICAL CLUSTERING

11/11/2024

FUZZY C-MEANS

k-means loss

$$\boxed{L(z) = \sum_{l}^{K} \sum_{i}^{N} \delta(z_i, l) \, \|x^i - c^l\|^2}$$

$z^i \rightarrow$ cluster number of point i

$$\mu^i = (0, 0, 1, \ldots)$$

z^i position

$$L(\mu^i) = \sum_{l}^{K} \sum_{i}^{N} \underbrace{\cancel{\delta(z_i, l)}}_{\mu_l^i} \|x^i - c^l\|^2$$

one-hot encoding version of k-means

$\mu_l^i \in \mathbb{R} \rightarrow$ Fuzzy c-means

$$\sum_l \mu_l^i = 1$$

$$\boxed{L(\mu) = \sum_{l}^{K} \sum_{i}^{N} (\mu_l^i)^m \, \|x^i - c^l\|^2}$$

Fuzzy c-means loss

usually $m = 2$

$$c^l = \underbrace{\frac{\sum_i^N \delta(z^i, l)\, x^i}{\sum_i^N \delta(z^i, l)}}_{\text{k-means center}} \rightarrow \boxed{c^l = \frac{\sum_i^N (\mu_i^l)^m x^i}{\sum_{i}^{N} (\mu_i^l)^m}}$$

Fuzzy c-means centers

Made with Goodnotes

Fuzzy c-means algorithm

1. $\mu^i \rightarrow$ random vectors normalized to $1 = \sum_{\ell}^{K} \mu^i_\ell = 1$

2. Compute the centers for our clusters

$$c^\ell = \frac{\sum_i^N (\mu^i_\ell)^m \, x^i}{\sum_i^N (\mu^i_\ell)^m}$$

3. Update of μ^i

$$\mu^i_\ell = \frac{1}{\sum_n^K \left(\frac{\| x^i - c^\ell \|}{\| x^i - c^n \|} \right)^{2/m-1}}$$

4. max change $\mu^i_\ell < \sigma$

$\hookrightarrow$ small number ~ 0.001

Made with Goodnotes

$$\mu_l^i = \frac{1}{\sum_{n}^{K} \left(\frac{\|x^i - c^l\|}{\|x^i - c^n\|} \right)^{2/}}$$

- $x^i = (0.1, 0.5, 0.3)$

  $x^j = (0.5, 0.2, 0.1)$

- $c^1 = (0.5, 0.3, 0.2)$

- $c^2 = (0.1, 0.2, 0.8)$

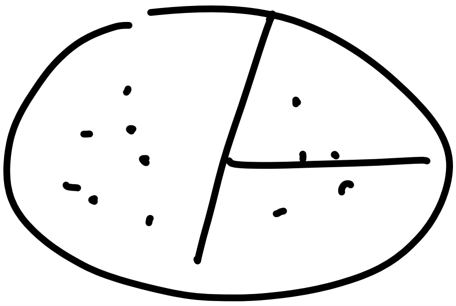
$$\mu_1^i = \frac{1}{\underbrace{\left(\frac{\|x^i - c^1\|}{\|x^i - c^1\|} \right)^2}_{1} + \left(\frac{\|x^i - c^1\|}{\|x^i - c^2\|} \right)^2}$$

$P(X)$

- ↳ Manifold learning
  - → Dim. Red.
  - → Intrinsic Dim.
- ↳ Density estimation
  - Histograms
  - KDE
  - k-nn
- ↳ Clustering
  - Flat (k-means)
  - Fuzzy (Fuzzy c-means)
  - Hierarchical

Hierarchical Clustering

Divisive Clustering

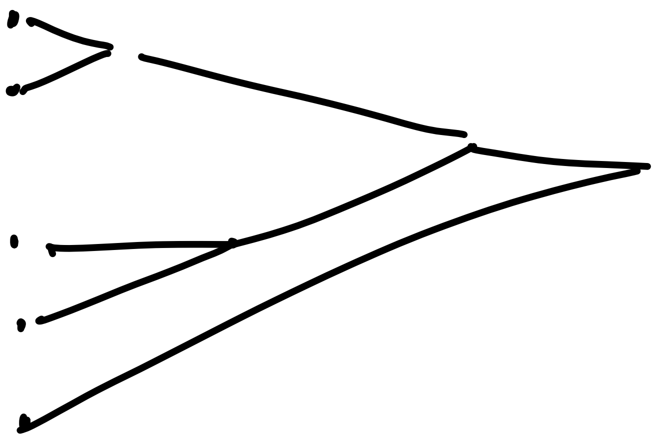


→ start considering all our points a single cluster and divide until each data point belongs to a different cluster

DIANA

Made with Goodnotes

Agglomerative Clustering



• Start considering each data point a cluster and you stop once all the data points belong to the same cluster

$2^{N-1} - 1$ possible divisions

Agglomerative clustering algorithm

① Compute the distance matrix between all your clusters

② Repeat :

2.a) Merge the two closest clusters

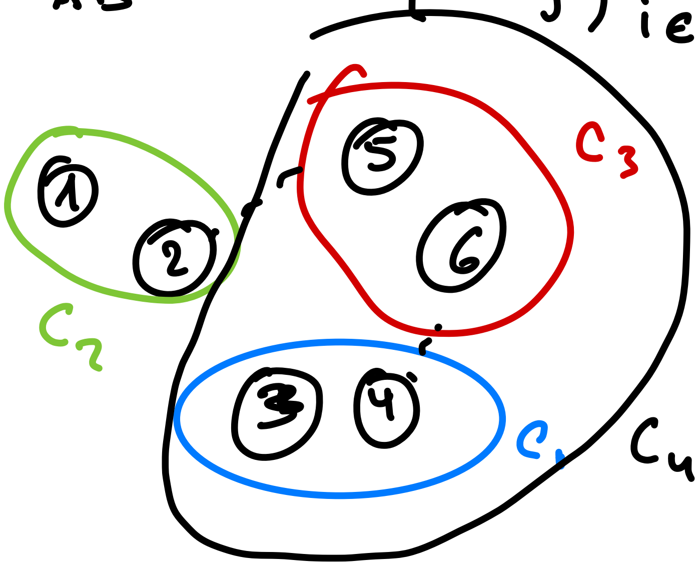
2. b) Update the distance matrix

③ Finish when all the points belong to a single cluster

Made with Goodnotes

Single Linkage distance:

$$d_{AB} = \min\{d_{ij}\} \quad i \in A;\ j \in B$$

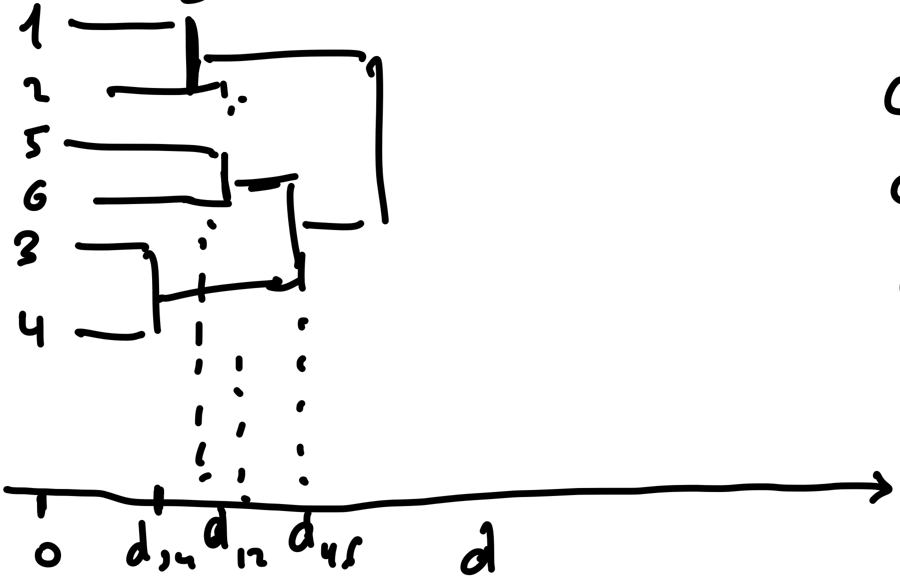


$$\begin{array}{c|cccccc} & 1 & 2 & 3 & 4 & 5 & 6 \\ \hline 1 & \emptyset & d_{12} & d_{13} & \cdots & & \cdot \\ 2 & & \emptyset & & & & \cdot \\ 3 & / & & \emptyset & & & \cdot \\ 4 & / & / & & \emptyset & & \\ 5 & & / & & & \emptyset & \\ 6 & & / & & / & & \emptyset \end{array}$$

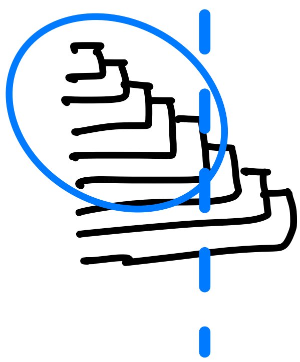
1 2 C_1 5 6

$$\begin{array}{c|ccc} & C_1 & C_2 & C_3 \\ \hline C_1 & & d_{23} & d_{46} \\ C_2 & & & d_{25} \\ C_3 & & & \end{array}$$

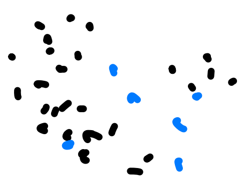
Dendrogram

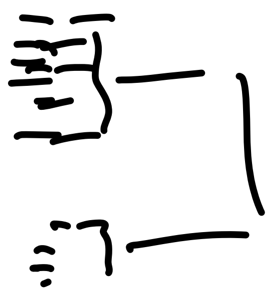


Advantge: It follows the cluster shape



Disadvantage:



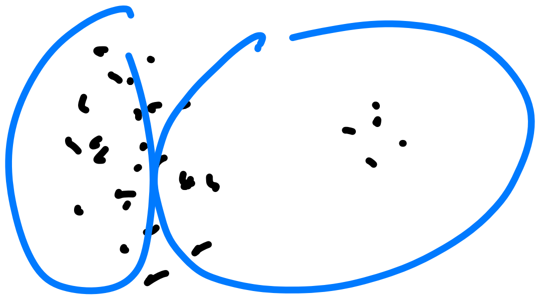


Complete Linkage

$$d_{AB} = \max [d_{ij} \; ; \; i \in A, j \in B]$$

Advantages: → noise unsensitive
→ It tends to generate balanced clusters

Disadvantages:
→ It forces the balanced structure



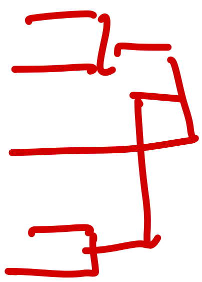
- Group-Average linkage

$$d_{AB} = \frac{1}{|A| \, |B|} \sum_{\substack{i \in A \\ j \in B}} d_{ij}$$

$|C|$ = number of points in cluster

- Centroid linkage

$d_{AB} = \| c^A - c^B \|$ distance between centroids



Made with Goodnotes

Ward's method

$$d_{AB}^2 = \frac{2\,|A|\,|B|}{|A| + |B|} \cdot \|C^A - C^B\|^2$$

$$d_{A\cup B, C} = \frac{|A| + |C|}{|A| + |B| + |C|} d_{AC} + \frac{|B| + |C|}{|A| + |B| + |C|} d_{BC}$$

$$- \frac{|C|}{|A| + |B| + |C|} d_{AB}$$

Adv: · Not sensitive to noise
· Tends to generate balanced clusters

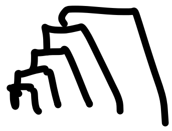
Disadvantage: · Biased towards convex clusters

It can be thought as hierarchical version of K-means

Made with Goodnotes

Noise ?

Yes → Bibliography

No → SL 

Bibliography

Yes → Use the tested method

No → Ward's method

DIANA

$2^{N-1} - 1$ possible partitions

~~Distance~~ Matrix

Dissimilarities

	a	b	c	d	e
a	$\emptyset$	2	6	10	9
b		$\emptyset$	5	9	8
c			$\emptyset$	4	5
d				$\emptyset$	3
e					$\emptyset$

1 Compute the average distance from each element to the others

a	6.75
b	6.0
c	5.0
d	6.5
e	6.25

Made with Goodnotes

a	6.75
b	6.0
c	5.0
d	6.5
e	6.25

2 Choose the element with highest average distance to nucleate a new cluster

3

	AN	AP	$\Delta = AP - AN$
b	2	7.33	5.33
c	6	4.66	-1.33
d	10	5.33	-4.67
e	9	5.33	-3.67

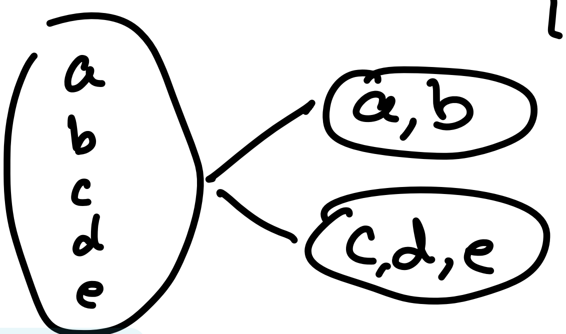
4 Consider the elements with positive Δ as elements of the new cluster

$\{a, b\}$ $\{c, d, e\}$

5 Repeat steps 3 and 4 until no positive Δ is found

	AN	AP	Δ
c	5.5	4.5	-1
d	9.5	3.5	-6
e	8.5	4.0	-4.5

$\{a, b\}$ $\{c, d, e\}$



Made with Goodnotes

6. Consider the cluster with highest pair dissimilarity for the next partition (Repeat points 2, 3, 4 & 5 in this cluster)

7. Stop when each data point belongs to a different cluster

Made with Goodnotes

Clustering → k-means

→ Fuzzy c-means

→ Hierarchical clustering

→ Agglomerative (SL, Ward's)

→ Divisive (DIANA)

Made with Goodnotes