Lecture 6 : Introduction to Clustering

04/11/2024

Clustering ≠ Classification

↳ No-labels ↳ We have labels

Feature Selection (Supervised Learning)

① Variable Ranking

ⓐ Compare each feature with the Ground Truth classification & quantify the level of correlation

ⓑ Sort according with the correlation & choose those more correlated

- linear correlation

$$R(i) = \frac{\sum_k (x_i^k - \bar{x}_i)(y^k - \bar{y})}{\sqrt{\sigma_{x_i}^2 \, \sigma_y^2}}$$

$R_1 \ldots R_m$

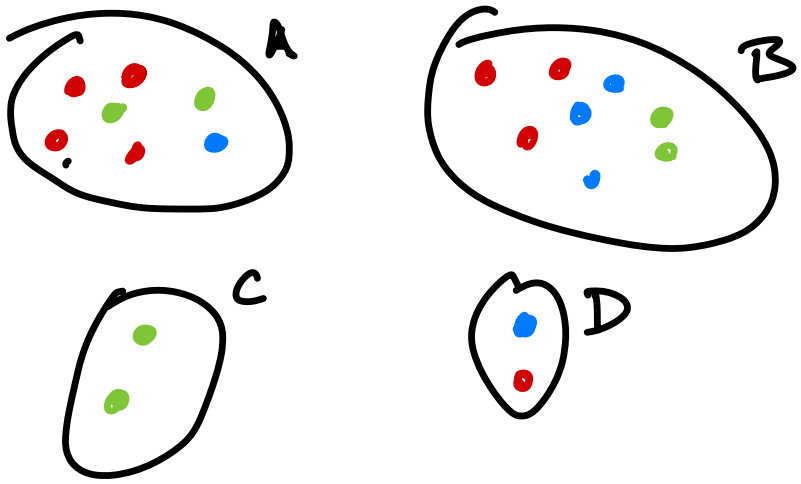
• Mutual Information

$$MI(i) = \int\int p(x_i, y) \log \frac{p(x_i, y)}{p(x_i)\,p(y)}$$

$$= \sum_{\forall y} \sum_{\forall x_i} p(x_i, y) \log \frac{p(x_i, y)}{p(x_i)\,p(y)}$$

19 (Green, Blue, Red) y

(A, B, C, D) Feature x_i



$$p(\text{Red}) = \frac{8}{19}$$

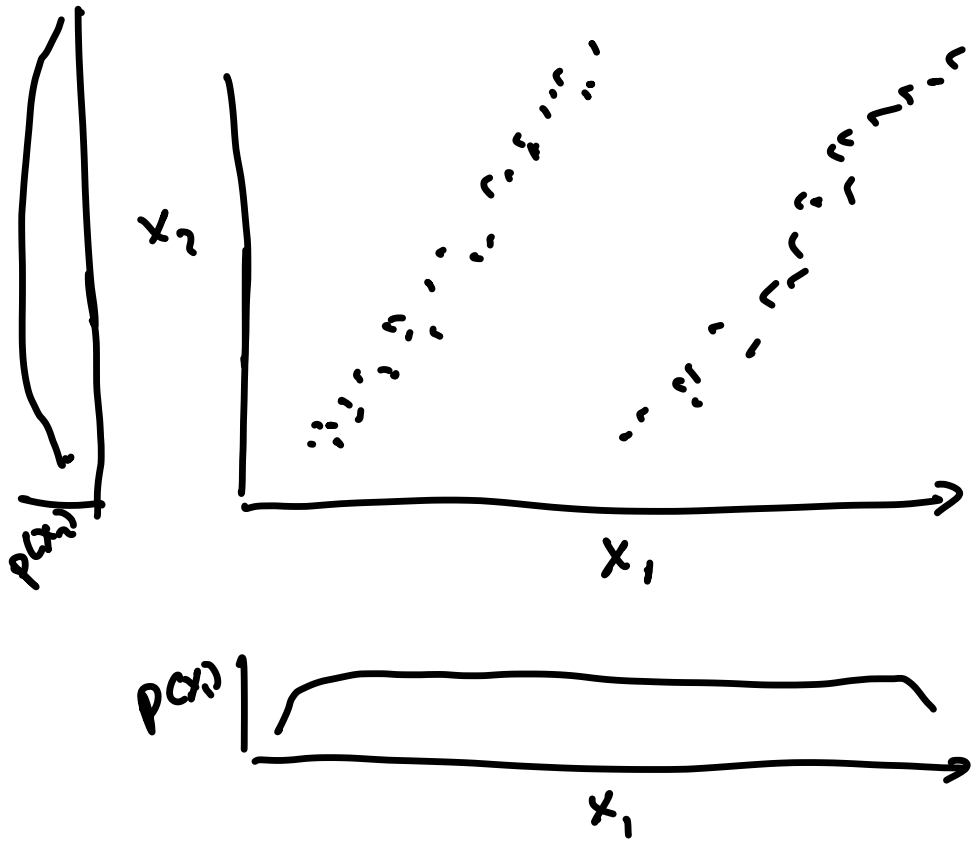
$$p(A) = \frac{7}{19}$$

$$p(\text{Red}, A) = \frac{4}{19}$$

$$\Sigma \; p\,\frac{4}{19} \cdot \log \frac{4/19}{7/19 \cdot 8/19} + \dots$$

Problems of variable ranking

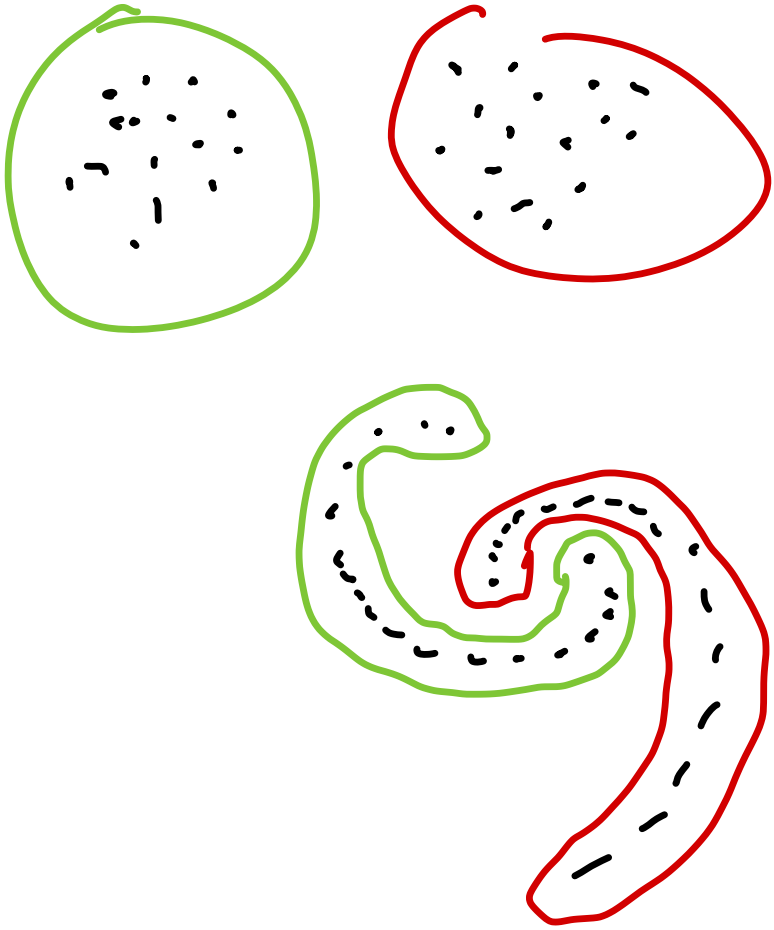
(A) Redundant variables



(2) Subset selection

- (a) Explore subsets of features
- (b) Rank

CLUSTERING



CLUSTERING METHODS CLASSIFIED ACCORDING TO THEIR OUTPUT

FLAT CLUSTERING : $Cl(i) = z^i$

↑ integer

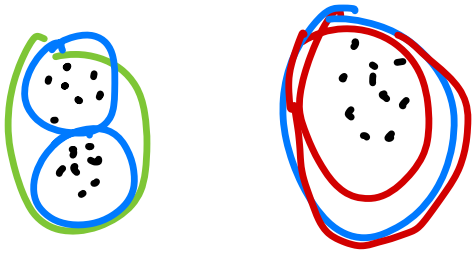
FUZZY CLUSTERING : $Cl(i) = \vec{\mu}^i$

↑ vector with k components

μ^i_j → defines the degree of membership of point i to cluster j

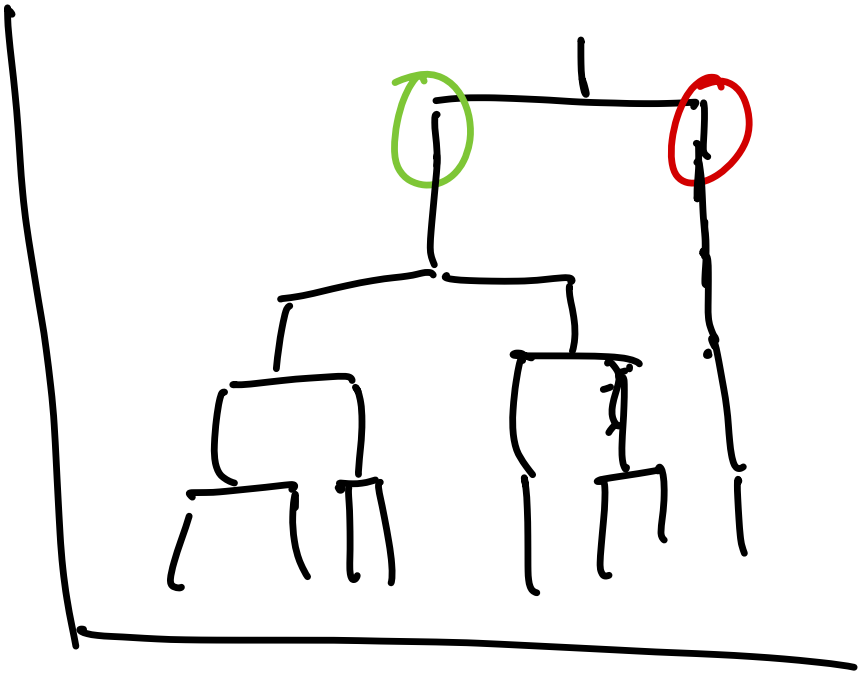
$$\sum_j^K \mu^i_j = 1$$

Hierarchical clustering



• Dendrogram

Binary tree



K-means clustering

< maximize inter cluster dissimilarity
minimizing intra cluster dissimilarity

$$c^l = \frac{\sum^N \delta(z^i, l)\, x^i}{\sum^N \delta(z^i, l)}$$

Average coordinates of points belonging to cluster l

$z^i : Cl(i)$; $\delta(z^i, l) \begin{cases} 1 & \text{if } i \in l \\ \emptyset & \text{otherwise} \end{cases}$

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

Loss of k-means

Target: find z that minimizes $L(z)$

(NP-complete problem)

k-means algorithm

① Pick k-centers (among the data points)

② Iterate

ⓐ Assign each data point to its nearest center

minimize intra cluster dis.

ⓑ Recompute the centers

$$c^l = \frac{\sum \delta(z^i, l) \, x^i}{\sum \delta(z^i, l)}$$

maximize the distance between centers

Kernel K-means

① Find distances in the feature space

$$d_{il}^2 = \|\vec{\phi}^i - \vec{\phi}^l\|^2 = (\vec{\phi}^i - \vec{\phi}^l)\cdot(\vec{\phi}^i - \vec{\phi}^l) =$$

$$\vec{\phi}^i\cdot\vec{\phi}^i + \vec{\phi}^l\cdot\vec{\phi}^l - 2\,\vec{\phi}^i\vec{\phi}^l = K^{ii} + K^{ll} - 2K^{il}$$

$$d_{il}^2 = \|\vec{\phi}^i - \vec{v}^l\|^2 = (\vec{\phi}^i - \vec{v}^l)\cdot(\vec{\phi}^i - \vec{v}^l)$$

l ↑ centroid

$$= \vec{\phi}^i\cdot\vec{\phi}^i - 2\,\vec{\phi}^i v^l + v^l v^l =$$

$$K^{ii} - 2\,\vec{\phi}^i\cdot\frac{\sum\delta(z^j,l)\,\vec{\phi}^j}{\sum\delta(z^j,l)} + \vec{v}^l\frac{\sum_m\delta(z^m,l)\vec{\phi}^m}{\sum\delta(z^m,l)}$$

$$d_{il}^2 =$$

$$K^{ii} - 2\frac{\sum\delta(z^j,l)\,K^{ij}}{\sum\delta(z^j,l)} + \frac{\sum_{n,m}\delta(z^m,l)\,\delta(z^n,l)\,K_{nm}}{\left(\sum_n\delta(z^n,l)\right)^2}$$

① Compute the Kernel (Gram) matrix

② Choose randomly K centers

③ Iterate $z^i = \underset{l\in\text{cluster}}{\operatorname{argmin}}\,(d_{il}^2)$