Lecture 24

Thursday, November 30, 2023 10:24 AM

Clustering evaluation

→ External measures

$\mathcal{C} = \{C_1, C_2, \ldots, C_r\}$ r = # of clusters chosen by the user

↳ Clustering using algo A

$|C_i| = n_i$

$T = \{T_1, T_2, \ldots, T_k\}$ k = # of 'true' clusters

ground truth partitioning

$|T_i| = m_i$

$|C_i \cap T_j| = n_{ij}$

	T_1	T_2	. T_j	T_k
C_1	n_{11}	n_{12}		
C_2				
C_i			n_{ij}	
C_r				

r k

$T = \{$ Iris data
$T_1 =$ setosa 50
$T_2 =$ virginica 50
$T_3 =$ versicolor 50 $\}$

$C = \{ C_1 = 75,\ C_2 = 30,\ C_3 = 45 \}$

$r = 3$

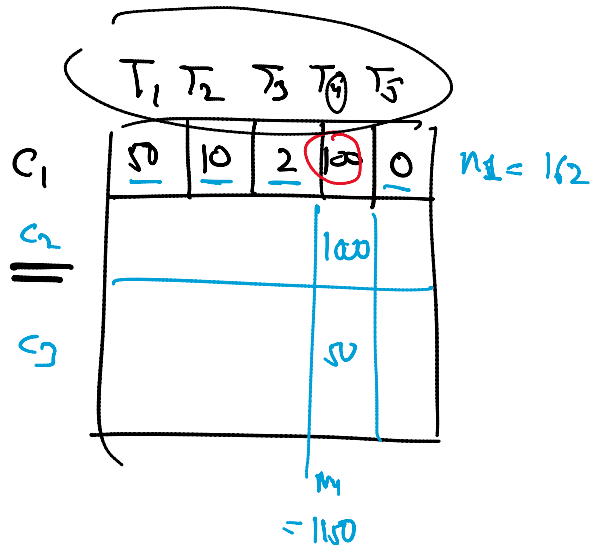
Purity

Purity

for each cluster C_i → find the max over the true partitions

$$\text{Purity}_i = \frac{1}{(n_i)} \max_{j=1}^{k} \{ n_{ij} \}$$

(C_i)



Precision$_i$ (accuracy C_i)

$$\text{Purity} = \sum_{i=1}^{r} \frac{n_i}{n} \text{Purity}_i$$

$$\text{Purity} = \frac{1}{n} \sum_{i=1}^{r} \max_{j=1}^{k} \{ n_{ij} \}$$

$$\text{Prec}(C_i) = \frac{100}{162}$$

$$\text{recall}(C_i) = \frac{100}{1150}$$

F-score : balance precision & recall

$$\underset{(C_i)}{\text{Prec}_i} = \text{Purity}_i = \frac{1}{(n_i)} \max_{j=1}^{k} \{ n_{ij} \}$$

$$\underset{(C_i)}{\text{recall}_i} = \frac{1}{(n_{j^*})} \max_{j=1}^{k} \{ n_{ij} \}$$

$$j^* = \arg\max_{j=1}^{k} \{ n_{ij} \}$$

F_i = harmonic mean (Prec_i, recall_i)

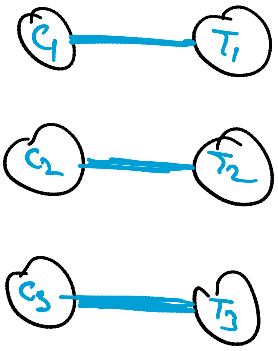
$$= \frac{2 \times \text{Prec}_i \times \text{recall}_i}{\quad}$$

$Prec_i + recall_i$

$$F = \frac{1}{r}\sum_{i=1}^{r} F_i \qquad (macro)$$

Maximum matching

	T_1	T_2	T_3	
C_1	50	100	0	$150 = n_1$
C_2	10	1000	50	$1060 = n_2$
C_3	100	90	150	$340 = n_3$
	160	1190	200	1550
	m_1	m_2	m_3	n



$$Purity_1 = \frac{100}{150}$$

$$Purity_2 = \frac{1000}{1060}$$

$$Purity_3 = \frac{150}{340}$$

weighted avg = purity

one to many

or many to one

$$\frac{50}{150}$$

$$\frac{1000}{1060}$$

$$\frac{150}{340}$$

weighted avg = maximum matching

one-to-one

Information Theoretic

Entropy

$$\frac{n_i}{n} = P_{C_i}$$

$$H(\mathcal{C}) = -\sum_{i=1} P(C_i) \log P(C_i)$$

$$H(T) = -\sum_{j=1}^{k} P(T_j) \log P(T_j) \qquad P_{T_j} = \frac{n_j}{n}$$

Mutual information

$$I(\mathcal{C}, T) = \sum_{i=1}^{r} \sum_{j=1}^{k} p_{ij} \log \left(\frac{p_{ij}}{P(C_i) \cdot P(T_j)} \right)$$

C_i T_j

Observed joint prob

joint prob if C_i is independent of T_j

$$I(\mathcal{C}, T) \in [0, \infty]$$

$$p_{ij} = \frac{n_{ij}}{n} = \frac{|C_i \cap T_j|}{n}$$

$$P(C_i) = \frac{n_i}{n} \qquad P(T_j) = \frac{n_j}{n}$$

normalized mutual information : $\in [0, 1]$

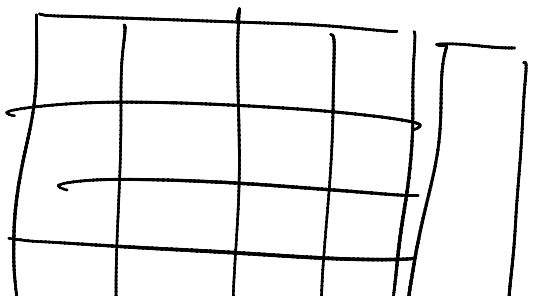
geometric mean of two ratios

$$\frac{I(\mathcal{C}, T)}{H(\mathcal{C})} \quad \text{and} \quad \frac{I(C, T)}{H(T)}$$

$$NMI = \sqrt{\frac{I(\mathcal{C}, T)}{H(\mathcal{C})} \cdot \frac{I(\mathcal{C}, T)}{H(T)}} = \frac{I(\mathcal{C}, T)}{\sqrt{H(\mathcal{C}) \cdot H(T)}}$$

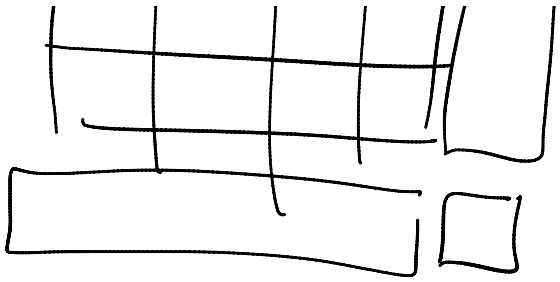
Confusion matrix

Contingency table



Confusion matrix

Contingency table



Pairwise labels

$$T = \{ \underset{y_1}{T_1}\ \underset{y_2}{T_2} \ldots \underset{y_k}{T_k} \} \quad \Big| \quad \mathcal{C} = \{ \underset{\hat{y}_1}{C_1}\ \underset{\hat{y}_2}{C_2} \ldots \underset{\hat{y}_r}{C_r} \}$$

$\vec{x_i}$ and $\vec{x_j}$	True	Predicted
1) True Positives (TP) :	$y_i = y_j$	and $\hat{y}_i = \hat{y}_j$
2) True negatives :	$y_i \neq y_j$	and $\hat{y}_i \neq \hat{y}_j$
3) False Positive :	$y_i \neq y_j$	and $\hat{y}_i = \hat{y}_j$
4) False negative :	$y_i = y_j$	and $\hat{y}_i \neq \hat{y}_j$

	True: P	True: N
Predicted: P	TP	FP
Predicted: N	FN	TN
		$\binom{n}{2}$

$\binom{n}{2}$ pairs $= \frac{n(n-1)}{2}$

Jaccard : $\frac{TP}{TP + FN + FP}$

FM : Fowlkes - Mallows

geometric mean of prec & recall

$$\text{prec} = \frac{TP}{TP+FP} \qquad \text{recall} = \frac{TP}{TP+FN}$$

$$\sqrt{\frac{TP}{(TP+FP)} \cdot \frac{TP}{(TP+FN)}}$$

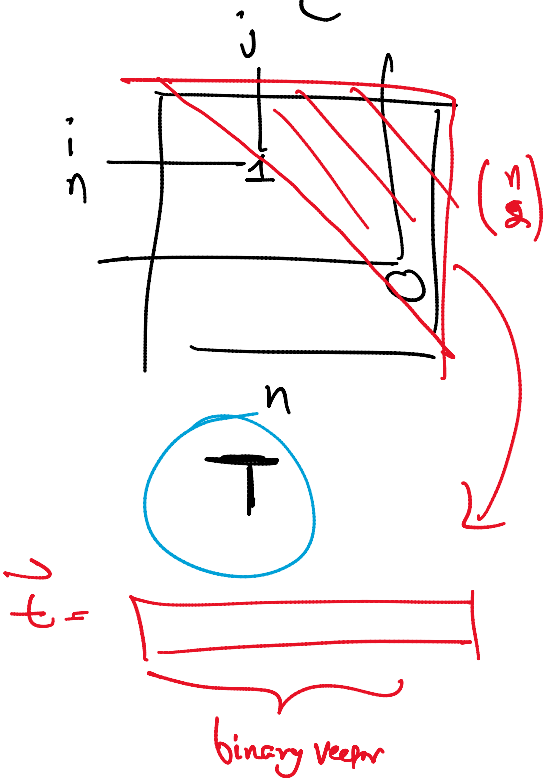
F-score : harmonic mean

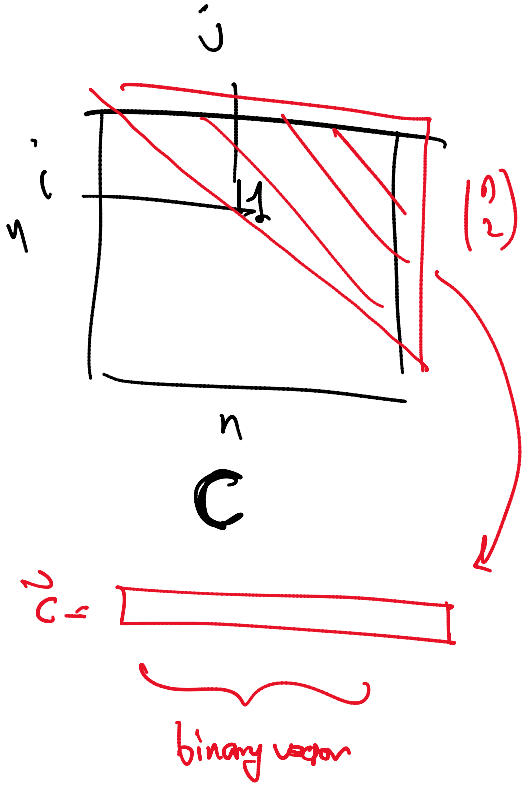
Correlation measure

define two matrices

$$T(i,j) = \begin{cases} 1 & \text{if } y_i = y_j \\ 0 & \text{otherwise} \end{cases}$$

$$C(i,j) = \begin{cases} 1 & \text{if } \hat{y}_i = \hat{y}_j \\ 0 & \text{otherwise} \end{cases}$$





Correlation?

$$\cos\Theta = \frac{\vec{t}}{\|\vec{t}\|} \cdot \frac{\vec{c}}{\|\vec{c}\|}$$

Hubert statistic

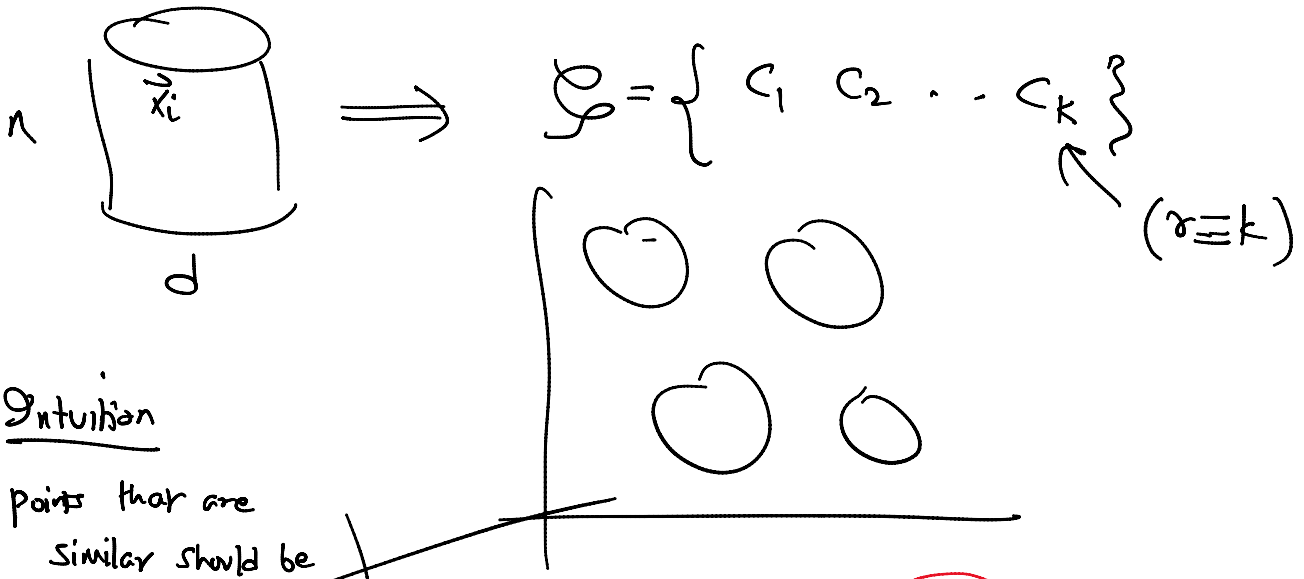
↳ Computed using 2×2 confusion matrix

(TP FP TN FN)

- Computed using 2x2 Confusion matrix (TP, FP, TN, FN)

Internal / Intrinsic

No T : no ground truth labels / partitions



Intuition

points that are

1) similar should be in the same cluster → intra-cluster distance should be small // similarity // high

2) diff clusters should be far apart → inter-cluster distance should be high / large

$W \equiv$ proximity matrix $\equiv$ distance matrix

$n \times n$ of distance

$w_{ij} = \|\vec{x}_i - \vec{x}_j\|$ ← L_2 norm for points (shortest path length in a graph)

$W = \{ w_{ij} \}_{i,j=1 \ldots n}$

$$W = \{ w_{ij} \}_{i,j=1 \cdots n}$$

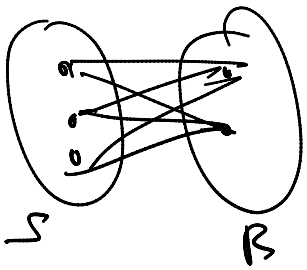
symmetric matrix $\quad w_{ij} \geq 0$

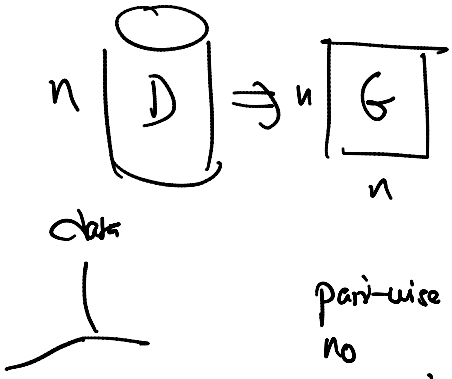
---

clustering $\equiv$ partition $\equiv$ k-way graph cut

$$W(S,R) = \sum_{i \in S} \sum_{j \in R} w_{ij}$$

$S, R \subseteq D$





pair-wise
no
Coordinate!

$$D \Longrightarrow \mathcal{C} = \{ C_1, C_2, \cdots, C_k \}$$

$$W(C_i, C_i) = \sum_{j \in C_i} \sum_{j \in C_i} w_{ij} \equiv \text{intra-cluster distance for } C_i$$

$$W_{in} = \frac{1}{2} \sum_{i=1}^{k} W(C_i, C_i)$$

over all clusters / intra-cluster

$$W_{out} = \frac{1}{2} \sum_{i=1}^{k} W(C_i, \bar{C}_i) \qquad \text{inter-cluster}$$

Normalized cut

$$\max \quad Nc = \sum^{k} \frac{W(C_i, \bar{C}_i)}{} = \sum^{k} \frac{W(C_i, \bar{C}_i)}{}$$

$$\max_{\{C_1, C_2 \cdots C_k\}} \quad Nc = \sum_{i=1}^{k} \frac{W(C_i, \bar{C}_i)}{vol(C_i)} = \sum_{i=1}^{k} \frac{w(C_i, \bar{C}_i)}{W(C_i, V)}$$

↳ W is a distance matrix

min if W were a kernel matrix