Representative-based clustering

→ K means

→ EM

Hierarchical

Density-based

→ Kernel density estimation

graph clustering

→ Spectral

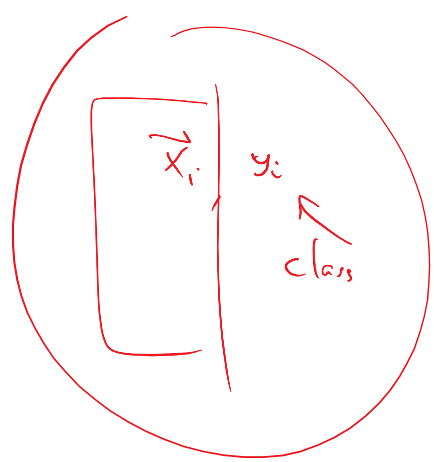
→ Markov chain

Validation / Evaluation of clustering results ?

① External Measure

where we have ground-truth.

(any classification dataset)

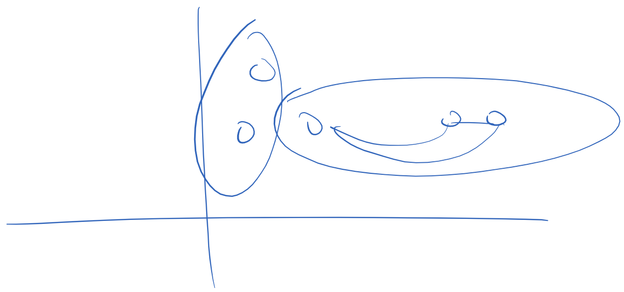


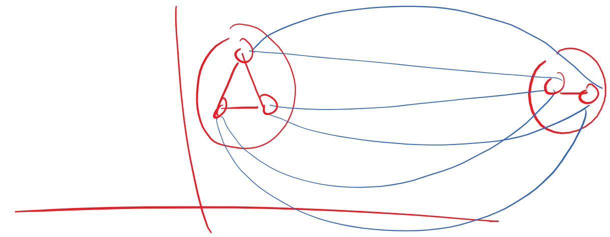
② Internal Measures

→ pair-wise distances

+

cluster output





③ Relative Measures

Selecting $\underline{\underline{k}}$ # of clusters

Cluster <u>stability</u>

→ how stable/robust are the clusters under small perturbations

Clustering tendency

→ Are there clusters?

D versus Random Points

External

$$T = \{T_1, T_2, \ldots, T_k\} \longleftarrow \text{ground truth}$$

True cluster 1

from some algo $= \underbrace{\{C_1, C_2, \ldots, C_k\}}$

k is given

algo $\underbrace{\qquad\qquad}_{\text{extracted clusters}}$

Contingency table / Confusion matrix

	T_1	T_2	$\dots$	T_k	
C_1	n_{11}	n_{12}	$\dots$		n_1
C_2					n_2
$\vdots$					$\vdots$
C_k					n_k
	m_1	m_2		m_k	

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

$$n_i = |C_i|$$

$$m_j = |T_j|$$

① **Purity** for each cluster C_i

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

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

weighted sum

Point

$C_1 \rightarrow T_2$

$C_2 \rightarrow T_3$

$\boxed{C_3 \rightarrow T_{\dots}}$

$$\text{Prec}_{C_1} = \frac{90}{94}$$

$$\text{recall}_1 = \frac{90}{200}$$

	T_1	T_2	T_3	
C_1	2	90	2	94
C_2	1	5	100	106

lunts

$c_2 \rightarrow T_3$

$\boxed{c_3 \rightarrow T_2}$

$\text{recall}_1 = \frac{T_2}{200}$

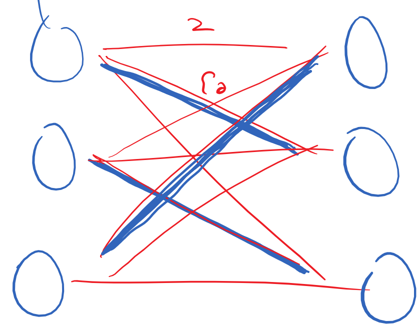
c_2	1	5	100	106
c_3	99	100	1	200
	100	200	100	400
	m_1	m_2	m_3	

vs

Maximum matching

$c_1 \rightarrow T_2$
$c_2 \rightarrow T_3$
$c_3 \rightarrow T_1$

Measure



1-1 matching
with the
highest sum
of edge weights

F-measure:

$$F_i = \frac{2 \cdot \text{Prec}_i \, \text{Recall}_i}{\text{Prec}_i + \text{Recall}_i}$$

$$\text{Prec}_i = \text{Purity}_i = \frac{1}{n_i} \max_j \{n_{ij}\}$$

F-measure $\mu_F = \frac{1}{k} \sum F_i$

$$j_i^* = \arg\max_j \{n_{ij}\}$$

$$\text{Recall}_i = \frac{n_{ij_i^*}}{m_{j_i^*}}$$

---

r: # of clusters found

	T_1	T_2	T_3
c_1			

T_1	T_2
0	50

r : # of clusters found

$r < k$

C_1			
C_2			

$r < k$

$r = 4$

T_1	T_2
0	50
50	0
10	40
30	20

$r > k$

T_2

C_1	C_3

C_2	C_4

Information Theory

$$\boxed{H = -\sum_{i=1}^{k} p_i \log p_i}$$

Conditional Entropy

$H(T|C_i)$ (row-wise entropy) $\rightarrow$

T_1	T_2	T_3	
0 (n_{11})	100 (n_{12})	0 (n_{13})	n_1
0	0	100	n_2
100	0	0	n_3
100	200	100	

$\boxed{0 \quad 1 \quad 0}$

$$H(T|C_i) = -\sum^{k} \left(\frac{n_{ij}}{n_i}\right) \log\left(\frac{n_{ij}}{n_i}\right)$$

$$H(T|C) = \underbrace{\frac{n_i}{n} H(T|C_i)}_{\text{weighted sum}}$$

C_1	50	100	50
C_2	40	-	-
C_3	10		

$\frac{1}{4}, \frac{1}{2}, \frac{1}{4}$

Perfect clustering $H(T|C) = 0$

$$\boxed{H(T|C) = \underbrace{H(T,C)}_{\text{joint entropy}} - \underbrace{H(C)}_{\text{entropy of clustering}}}$$

joint entropy

entropy of clustering

$$-\sum_{i}\sum_{j} p_{ij} \log p_{ij}$$

$$-\sum_{i} p_{c_i} \log p_{c_i}$$

Pairwise

TP: how many points x_i & x_j belong to the same cluster C_k & they also belong to the same true cluster T_a

$$^{n}C_2 = \binom{n}{2} = \frac{n.(n-1)}{2}$$

all distinct pair of points

$T_1 = \{1, 2, 3\}$ $T_2 = \{4, 5, 6\}$

$C_1 = \{1, 4\}$, $C_2 = \{2, 3, 5, 6\}$

TP: $1 + 1 = 2$

TN: # of pairs of points that belong to diff clusters & also diff true clusters

$C_1' = \{1, 2, 3\}$

$C_2' = \{4, 5, 6\}$

	T	$\bar{T}$
C	TP	FP
$\bar{C}$	FN	TN

Jaccard: $\frac{TP}{TP + FN + FP}$ | Ignores TN

FM: Folkes Mallow

geometric mean of Prec & Recall

TP

recall = $\frac{TP}{TP+FN}$

FM: $\sqrt{Prec. \ Recall}$

$$= \sqrt{\frac{TP^2}{(TP+FP)(TP+FN)}}$$

No ground truth!!!

D $\rightarrow$ $C = \{c_1, c_2, \ldots, c_k\}$

$W = \{ w_{ij} \}$

Pair-wise distance matrix

$w_{ij} = \|\vec{x}_i - \vec{x}_j\|^2$

for graph data

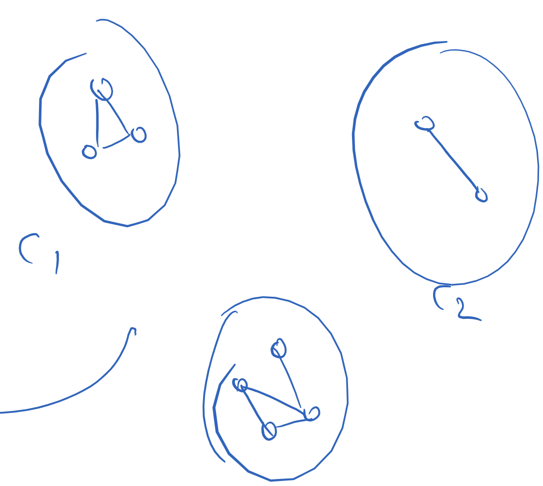
w_{ij} = # of hops in the shortest path

$$W(S,T) = \sum_{x_a \in S} \sum_{x_b \in T} w_{ab}$$

weight of the cut

$$W_{in} = \frac{1}{2} \sum^{k} w(C_i, C_i)$$

Internal weights



N:

N_{in} = how many pairs of internal points

$\sum_{i=1}$

$n_i = |C_i|$

$\underline{W_{out}}: \frac{1}{2} \sum_{i=1}^{k} W(C_i, \overline{C_i})$

external weight

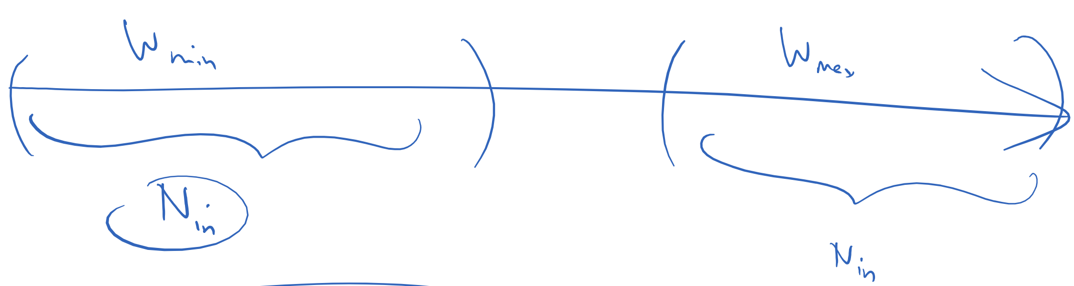
$\underline{N_{out}}: n - N_{in}$

external pairs of points

① BetaCV : $\dfrac{W_{in}/N_{in}}{W_{out}/N_{out}}$ smaller is better!

② C-index N_{in} : # of internal pairs

$W_{min}(N_{in})$ = N_{in} smallest weight in the W matrix



$C_{index} = \dfrac{W_{in} - W_{min}(N_{in})}{W_{max}(N_{in}) - W_{min}(N_{in})}$ Smaller the better

Varye

Silhouette Coefficient

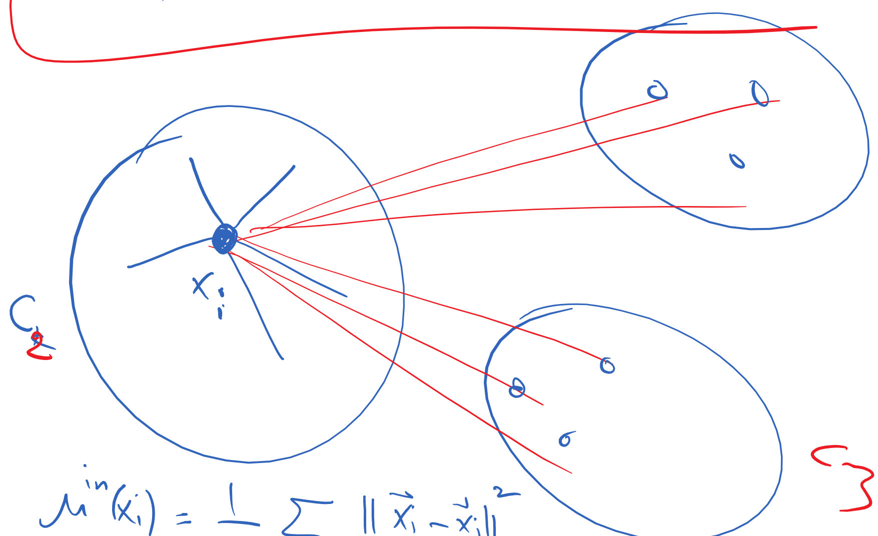
$$\boxed{\underset{\text{Per point } \vec{x}_i}{S_i} = \frac{\mu^{out}(x_i) - \mu^{in}(x_i)}{\max\{\mu^{out}, \mu^{in}\}}}$$

S_i close to 1 is best!

$S_i \in [-1, 1]$

$S_i = -1$

$\Rightarrow$ mis-clustered point



$$\mu^{in}(x_i) = \frac{1}{n_k} \sum_{x_j \in C_k} \underbrace{\|\vec{x}_i - \vec{x}_j\|^2}$$

$$\mu^{out}(x_i) = \min_{C_k} \left\{ \text{avg distance to cluster } C_k \right\}$$

$$\boxed{SC: \frac{1}{n} \sum s_i}$$

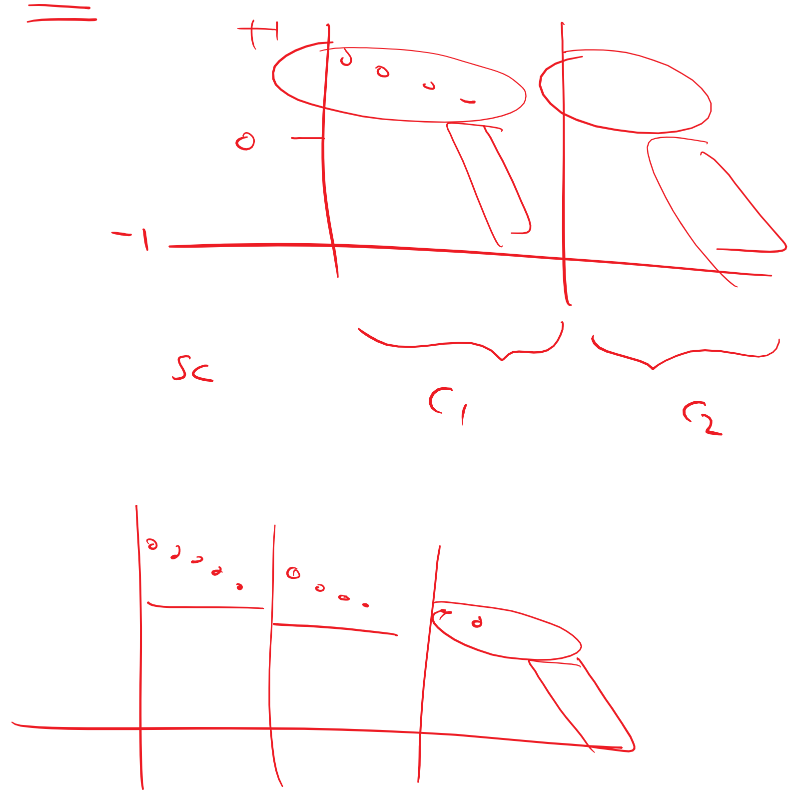
avg over all points

Relative Measure to select k

D — k = 2

k = 3

fix algo



gap statistic to select k

$C = \{c_1, c_2 \dots c_k\}$

D $\rightarrow$ W_{in} : total internal weight

is this good or bad?

generate random samples in the same data space as D

q random

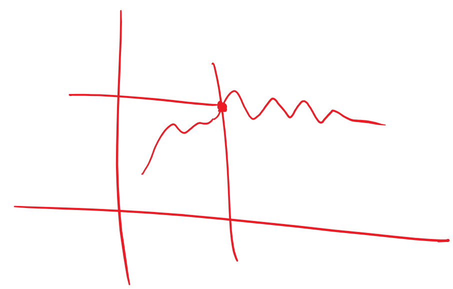
$$\left\{ \begin{array}{l} R_1 \rightarrow W_{in}^{k}(R_1) \\ R_2 \rightarrow W_{in}^{k}(R_2) \\ \vdots \end{array} \right\}$$

$\hat{\mu}^k \leftarrow$ mean internal

random samples

$$\left\{\begin{array}{l} R_2 \longrightarrow W_{in}(R_2) \\ \vdots \\ R_q \longrightarrow W_{in}^{k}(R_q) \end{array}\right\} \quad \begin{array}{l} \mu \longleftarrow \text{mean internal} \\ \sigma^{k} \longleftarrow \text{Stdev.} \end{array}$$

$$gap(k) = \underbrace{\mu^{k}}_{\text{from } q \text{ random samples}} - \underbrace{W_{in}^{k}}_{\text{from } D}$$



① Pick the max k

② Choose smallest k such that

$$\left[\, gap(k) \geq \underbrace{gap(k+1) - \sigma^{k+1}}_{\text{within a deviation of the next value of } k} \,\right]$$