EVALUATING CLUSTERING METHODS

(1) Internal clustering validation:

- No ground-truth available.
- How good is the cluster structure?

(2) External cluster evaluation:

- Take a benchmark with known labels
- How "close" are the clusters to the true labels?

INTERNAL EVALUATION

$C_1, C_2, \_\_\_, C_k$ = output clusters

DUNN INDEX

(1) Inter-cluster distance

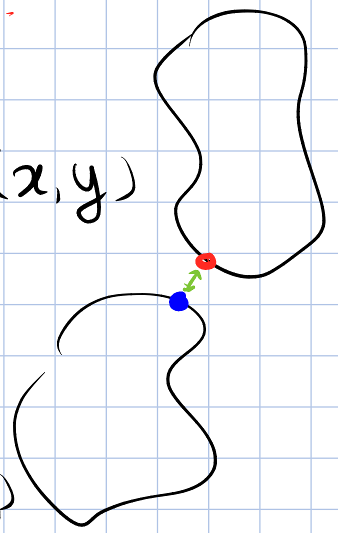
= how well-separated the clusters are.

Examples:

distance

- $\delta(C_i, C_j) = \min_{\substack{x \in C_i \\ y \in C_j}} d(x,y)$ ← single-linkage distance

- average linkage distance → $\delta(C_i, C_j) = \frac{1}{|C_i||C_j|} \sum_{\substack{x \in C_i \\ y \in C_j}} d(x,y)$

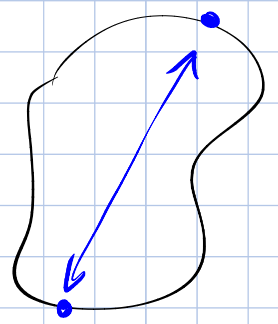


(2) Intra-cluster distance = how "concentrated"

Example:

diameter

- $\Delta(C_i) = \max_{x, x' \in C_i} d(x, x')$



- $\Delta(C_i) = \frac{1}{|C_i|(|C_i|-1)} \sum_{x, x' \in C_i} d(x, x')$ → average distance.

The Dunn Index:

$$DI = \frac{\min_{1 \le i < j \le k} \delta(C_i, C_j)}{\max_{1 \le l \le k} \Delta(C_l)}$$

worst case (numerator) ; worst case (denominator)

high = good separation (numerator)

→ as high as possible

low = good concentration (denominator)

PROPERTIES:

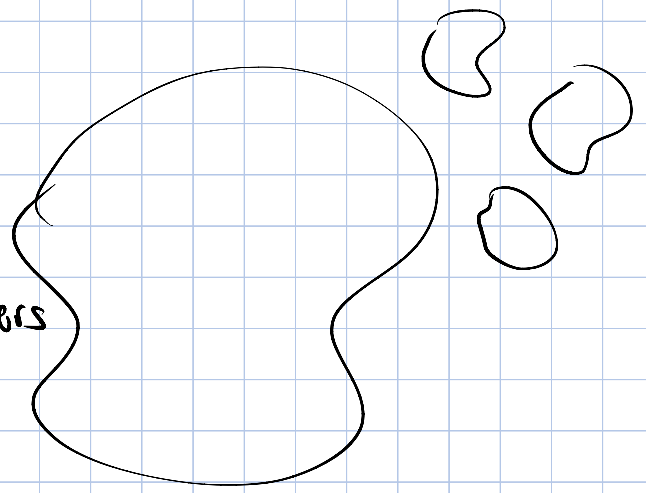
- $DI \in [0, \infty)$
- High DI ⇒ better clustering
- Issue:

When we have a mix of small & big clusters

↓

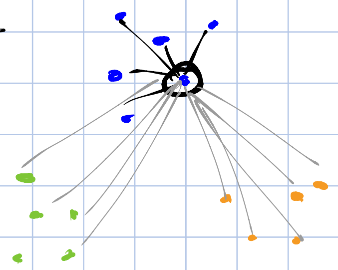
DI → low

(even if method works well)



Silhouette Analysis

- Clusters: $C_1, C_2, \ldots, C_k$.
- Pick $x \in C_i$:



$$a(x) = \frac{1}{|C_i|-1} \sum_{x' \in C_i} d(x, x') \rightarrow \text{within-cluster distance.}$$

$$b(x) = \min_{j \neq i} \frac{1}{|C_j|} \sum_{y \in C_j} d(x, y) \rightarrow \text{between-cluster distance.}$$

Silhouette Coefficient:

$$s(x) = \frac{b(x) - a(x)}{\max(a(x), b(x))}$$

If $x \in C_i$ and $|C_i| = 1$

↓

$s(x) = 0$.

Comments:

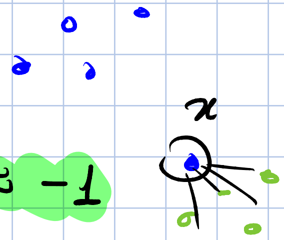
- $a(x) < b(x) \Rightarrow s(x) = 1 - \frac{a(x)}{b(x)} \in (0, 1]$
- $a(x) > b(x) \Rightarrow s(x) = \frac{b(x)}{a(x)} - 1 \in [-1, 0)$
- $a(x) = b(x) \Rightarrow s(x) = 0$.

$\Rightarrow \quad -1 \leq s(x) \leq 1$

Good clustering:

$a(x) << b(x) \rightarrow s(x) \approx +1$.

Bad Clustering:

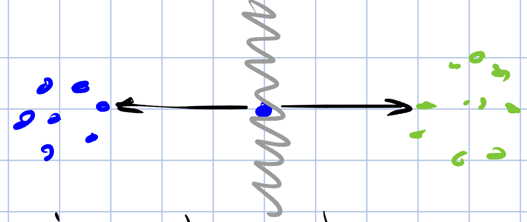


$b(x) << a(x) \rightarrow s(x) \approx -1$

⊗ $s(x) = 0$

↓

x is on the decision boundary.



MEAN SILHOUETTE COEFFICIENT:

$$SI = \frac{1}{n} \sum_{x} s(x)$$

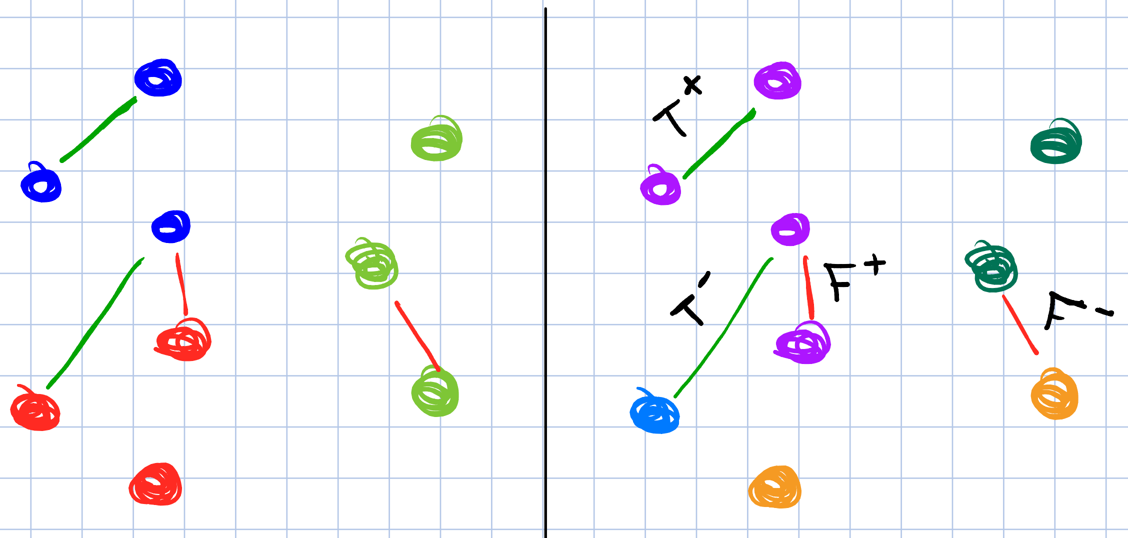
→ overall quality of algorithm.

EXTERNAL EVALUATION

Compare results to some benchmark.

- labelled data.
- data with concensus on "the truth"

RAND INDEX



"truth" | our algorithm

Data set: $\{x_1, x_2, \_\_, x_n\}$

Ground truth: $C^* = \{C_1^*, C_2^*, \_\_, C_m^*\}$

Our algorithm: $C = \{C_1, C_2, \_\_, C_k\}$

Denote:

- $x_i \sim x_j$ if x_i, x_j in the same cluster in C
- $x_i \overset{*}{\sim} x_j$ if x_i, x_j in the same cluster in C^*

Define:

$$(x_i, x_j) \longrightarrow \begin{cases} T^+ & \text{if } x_i \sim x_j \text{ and } x_i \overset{*}{\sim} x_j \\ T^- & \text{if } x_i \not\sim x_j \text{ and } x_i \overset{*}{\not\sim} x_j \\ F^+ & \text{if } x_i \sim x_j \text{ and } x_i \overset{*}{\not\sim} x_j \\ F^- & \text{if } x_i \not\sim x_j \text{ and } x_i \overset{*}{\sim} x_j \end{cases}$$

TP = total number of T^+'s

TN = " " " T^-'s

FP = " " " F^+'s

FN = " " " F^-'s

RAND INDEX:

$$RI = \frac{TP + TN}{\underbrace{TP + TN + FP + FN}_{\binom{n}{2} = \frac{n \cdot (n-1)}{2}}} \in [0, 1]$$

$RI \approx 1$ - good clustering