Clustering Calculations

Given. Two clusters in $\mathbb{R}^2$:

$$C_1 = \{(1,2),(0,-1)\}, \qquad C_2 = \{(0,0),(1,1)\}.$$

Euclidean distance is used for (a). Cosine similarity $\operatorname{cosSim}(x,y) = \dfrac{x^\top y}{\|x\|\,\|y\|}$ is used for (b)–(d).[1]

1. Linkage/similarity questions

(a) Weighted average intra-cluster distance (Euclidean). Each cluster has exactly one pair, so the weighted average of all within-cluster pairwise distances equals the simple average:

$$\bar{D}_{\text{intra}} = \frac{1}{2}\Big(\underbrace{\sqrt{(1-0)^2+(2-(-1))^2}}_{C_1} + \underbrace{\sqrt{(1-0)^2+(1-0)^2}}_{C_2}\Big) = \frac{\sqrt{10}+\sqrt{2}}{2} \approx 2.288245611.$$

(b) Single-link similarity (cosine). A single link with a similarity function is the maximum inter-cluster pairwise similarity. Compute the relevant cosines:

$$\begin{aligned}
\operatorname{cosSim}\big((1,2),(1,1)\big) &= \frac{1\cdot 1 + 2\cdot 1}{\sqrt{5}\,\sqrt{2}} = \frac{3}{\sqrt{10}} \approx 0.948683298,\\
\operatorname{cosSim}\big((0,-1),(1,1)\big) &= \frac{0\cdot 1 + (-1)\cdot 1}{1\cdot\sqrt{2}} = -\frac{1}{\sqrt{2}} \approx -0.707106781,\\
\operatorname{cosSim}\big((1,2),(0,0)\big) &= 0, \qquad \operatorname{cosSim}\big((0,-1),(0,0)\big) = 0.
\end{aligned}$$

Hence

$$\boxed{\text{single-link similarity} = \max = \frac{3}{\sqrt{10}} \approx 0.948683298}.$$

(c) Complete-link similarity (cosine). Complete link with a similarity function is the minimum inter-cluster pairwise similarity:

$$\boxed{\text{complete-link similarity} = \min = -\frac{1}{\sqrt{2}} \approx -0.707106781}.$$

(d) Average-link similarity (cosine). Average link takes the mean of all inter-cluster pairwise similarities. With the convention that any pair involving $(0,0)$ contributes 0:

$$\text{average-link similarity} = \frac{0+0+\frac{3}{\sqrt{10}}-\frac{1}{\sqrt{2}}}{4} = \frac{\frac{3}{\sqrt{10}}-\frac{1}{\sqrt{2}}}{4} \approx 0.060394129.$$

Remark. If one instead excludes undefined pairs with the zero vector, the average over the two defined pairs is $\big(\frac{3}{\sqrt{10}}-\frac{1}{\sqrt{2}}\big)/2 \approx 0.120788258$.

[1] Cosine similarity is undefined when either vector is the zero vector. In this solution, any pair involving $(0,0)$ is taken to have cosine similarity 0 by convention; if such pairs are excluded instead, the single/complete link values remain unchanged, while the average link doubles (see Remark at the end of item (d)).

2. On W_j'''' for average intra-cluster distance sequence W_j

Here j indexes discrete clustering levels; W_j is a sequence, not a continuous function, so the classical derivative W_j'''' is *not defined.* If "fourth derivative" is intended as the *fourth forward finite difference*, then

$$\Delta^4 W_j = W_{j+4} - 4W_{j+3} + 6W_{j+2} - 4W_{j+1} + W_j,$$

which cannot be numerically evaluated without the values $W_j, \dots, W_{j+4}$.

3. Weighted average purity

Predicted clusters: $C_1 = \{1, 2, 3, 4\}$, $C_2 = \{5, 6, 7, 8\}$.
Ground truth (hand labels): $G_1 = \{3, 4\}$, $G_2 = \{1, 2, 5, 6, 7, 8\}$.

$$\begin{aligned} |C_1 \cap G_1| = 2, \quad |C_1 \cap G_2| = 2 \;\Rightarrow\; \max = 2, \\ |C_2 \cap G_1| = 0, \quad |C_2 \cap G_2| = 4 \;\Rightarrow\; \max = 4. \end{aligned}$$

Weighted average purity (standard definition) is

$$\text{Purity} = \frac{\max_\ell |C_1 \cap G_\ell| + \max_\ell |C_2 \cap G_\ell|}{|C_1| + |C_2|} = \frac{2+4}{8} = \boxed{\tfrac{3}{4} = 0.75}.$$