Crisp cluster validity indices:
Dunn, D. Bouldin and Silhouette

Joaquim Viegas

December 2014

1 Notation

Adapted from Arbelaitz et al. (2013) [1].

Dataset X of N objects represented as objects in a p-dimensional space: $X = \{x_1, x_2, ..., x_N\} \subseteq \Re^p$

A partition or clustering in X is a set of disjoint clusters that partition X into K groups: $C = \{c_1, c_2, ..., c_k\}$ where $\cup_{c_k \in C} c_k = X, c_k \cap c_l = \emptyset \forall k \neq l$.

The centroid of a cluster c_k is its mean vector, $\overline{c_k} = \frac{1}{|c_k|} \sum_{x_i \in c_k} x_i$.

Dataset mean vector: $\overline{X} = \frac{1}{N} \sum_{x_i \in X} x_i$.

Euclidean distance: $d_e(x_i, x_k) = \sqrt{\sum_{j=1}^{p} (x_{ij} - x_{kj})^2}$.

2 Dunn Index (D↑) [3]

It is a ratio-type index where the cohesion is estimated by the nearest neighbor distance and the separation by the maximum cluster diameter. The original index is defined as

$$D(C) = \frac{min_{c_k \in C}\{min_{c_l \in C \setminus c_k}\{\delta(c_k, c_l)\}\}}{max_{c_k \in C}\{\Delta(c_k)\}} \tag{1}$$

where

$$\delta(c_k, c_l) = min_{x_i \in c_k} min_{x_j \in c_l}\{d_e(x_i, x_j)\} \tag{2}$$

$$\Delta(c_k) = max_{x_i, x_j \in c_k}\{d_e(x_i, x_j)\} \tag{3}$$

3 Davis-Bouldin Index (DB↓) [2]

It estimates the cohesion based on the distance from the points in a cluster to its centroid and the separation based on the distance between centroids. It is defined as

$$DB(C) = \frac{1}{K} \sum_{c_k \in C} max_{c_l \in C \setminus c_k} \left\{ \frac{S(c_k) + S(c_l)}{d_e(\overline{c_k}, \overline{c_l})} \right\} \tag{4}$$

where

$$S(c_k) = \frac{1}{|c_k|} \sum_{x_i \in c_k} d_e(x_i, \overline{c_k}) \tag{5}$$

4 Silhouette Index (Sil↑) [4]

This index is a normalized summation-type index. The cohesion is measured based on the distance between all the points in the same cluster and the separation is based on the nearest neighbor distance. It is defined as

$$Sil(C) = \frac{1}{N} \sum_{c_k \in C} \sum_{x_i \in c_k} \frac{b(x_i, c_k) - a(x_i, c_k)}{max\{a(x_i, c_k), b(x_i, c_k)\}} \tag{6}$$

where

$$a(x_i, c_k) = \frac{1}{|c_k|} \sum_{x_j \in c_k} d_e(x_i, x_j) \tag{7}$$

$$min_{c_l \in C \setminus c_k} \left\{ \frac{1}{|c_l|} \sum_{x_j \in c_l} d_e(x_i, x_j) \right\} \tag{8}$$

References

[1] Olatz Arbelaitz, Ibai Gurrutxaga, Javier Muguerza, Jesús M. Pérez, and Iñigo Perona. An extensive comparative study of cluster validity indices. *Pattern Recognition*, 46(1):243–256, January 2013. cvi multiple.

[2] David L. Davies and Donald W. Bouldin. A cluster separation measure. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, (2):224–227, 1979. bouldin cvi.

[3] C Dunn. A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters. *Journal of Cybernetics*, 3(3):32–57, 1973. dunn cvi.

[4] Peter Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65, November 1987. sil cvi.