CS724: Topics in Algorithms
Evaluation of Clustering Quality
Internal Measures

Prof. Dan A. Simovici

The Davies-Bouldin Index

The Davies-Bouldin index is designed for evaluating the quality of non-overlapping clusterings.

Definition

Let (S, d) be a metric space. A *dispersion measure* on (S, d) is a function $s : \mathcal{P}(S) \longrightarrow \mathbb{R}_{\geqslant 0}$ such that $s(C) = 0$ if and only if $|C| = 1$.

Example

The function *sse* is a dispersion measure.

Example

The function $\delta : \mathcal{P}(S) \longrightarrow \mathbb{R}_{\geqslant 0}$ defined by

$$\delta(C) = \frac{\sum\{d(x,y) \mid x,y \in C, x \neq y\}}{|C|(|C|-1)}$$

yields the mean distance between all pairs of objects in C. It is immediate to see that $\delta(C) = 0$ if and only if $|C| = 1$, so δ is a dispersion measure.

Example

The diameter $diam : \mathcal{P}(S) \longrightarrow \mathbb{R}_{\geqslant 0}$ is a dispersion function for obvious reasons.

Definition

Let $\kappa = \{C_1, \ldots, C_k\}$ be a clustering in a metric space (S, d), s_i be the dispersion of C_i, and r_{ij} be the distance between the representatives c_i and c_j of the clusters C_i and C_j (usually chosen as the centroids of the clusters C_i and C_j) for $1 \leqslant i, j \leqslant k$.

A *cluster similarity measure* is a function $r : \mathbb{R}^3_{\geqslant 0} \longrightarrow \hat{\mathbb{R}}$ that satisfies the following conditions:

- $r(s_i, s_j, m_{ij}) \geqslant 0$;
- $r(s_i, s_j, m_{ij}) = r(s_j, s_i, m_{ij})$;
- $r(s_i, s_j, m_{ij}) = 0$ if and only if $s_i = s_j$;
- if $s_j = s_k$ and $m_{ij} < m_{ik}$, then $r(s_i, s_j, m_{ij}) > r(s_i, s_k, m_{ik})$;
- if $m_{ik} = m_{ij}$ and $s_j > s_k$, then $r(s_i, s_j, m_{ij}) > r(s_i, s_k, m_{ik})$.

UMASS BOSTON

- When the distance between cluster centers increases while their dispersions remain constant, the similarity of the clusters decreases.
- If the distances between cluster centroids remains constant while the dispersion increase, the similarity increases.

Example

Consider the function r given by

$$r(s, s', m) = \frac{s + s'}{m}$$

for $s, s', m \in \mathbb{R}_{\geqslant 0}$. It is immediate that r satifies the conditions imposed on similarity measures.

Definition

Let $\kappa = \{C_1, \ldots, C_k\}$ be a clustering in a metric space (S, d). The *Davies-Bouldin index* of κ is the clustering average similarity measure r_κ given by

$$r_\kappa = \frac{1}{k}\sum_{i=1}^{k} \max\{r_{ij} \mid 1 \leqslant j \leqslant k\}.$$

The "best" clustering is the one that minimizes the average similarity measure.

Example

Consider a data set in $\mathbb{R}^2$ shown next:

$$\mathbf{v}_1 = \begin{pmatrix} 2 \\ 1 \end{pmatrix}, \mathbf{v}_2 = \begin{pmatrix} 2 \\ 3 \end{pmatrix}, \mathbf{v}_3 = \begin{pmatrix} 8 \\ 1 \end{pmatrix}, \mathbf{v}_4 = \begin{pmatrix} 8 \\ 3 \end{pmatrix}$$

grouped into two clusterings:

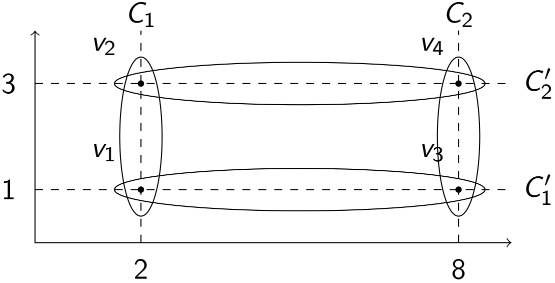
$$\kappa = \{C_1, C_2\}, \kappa' = \{C_1', C_2\},$$

where

$$C_1 = \{\mathbf{v}_1, \mathbf{v}_2\}, C_2 = \{\mathbf{v}_3, \mathbf{v}_4\},$$

and

$$C_1' = \{\{\mathbf{v}_1, \mathbf{v}_3\}, C_2' = \{\mathbf{v}_2, \mathbf{v}_4\}.$$


C_1
C_2
v_2
v_4
3
C'_2
v_1
v_3
1
C'_1
2
8

The centroids of the clusters are:

cluster	C_1	C_2	C_1'	C_2'
centroid	$\begin{pmatrix}2\\2\end{pmatrix}$	$\begin{pmatrix}8\\2\end{pmatrix}$	$\begin{pmatrix}5\\1\end{pmatrix}$	$\begin{pmatrix}5\\3\end{pmatrix}$

We choose the dispersion measure as the sum of the square errors. Its values for the clusters are:

$$sse(C_1) = 2, sse(C_2) = 2, sse(C_1') = 18, sse(C_2') = 18.$$

Thus, $r_{12} = 0.8$ and $r'_{12} = 18$, hence $r_\kappa = 0.8$ and $r_{\kappa'} = 18$, giving the edge to κ.

The Dunn Quality Indices

A related family of cluster quality indices is known as *Dunn quality indices*. For a clustering $\kappa = \{C_1, \ldots, C_k\}$ a Dunn index is a function

$$\Delta(\kappa) = \frac{\min_{1 \leqslant i < j \leqslant k} D(C_i, C_j)}{\max_{1 \leqslant j \leqslant k} s(C_j)},$$

where s is a dispersion measure, and $D(C_i, C_j)$ is an intercluster dissimilarity (which can be the least distance between two points in different clusters, the maximum distance between two such points, or the distance between the centroids of the clusters, etc.). Note that if a cluster has a high value of the dispersion this impacts negatively the value of the index due to the presence of max in the denominator.

The Silhouette Coefficient

Let $\kappa = \{C_1, \ldots, C_k\}$ be a clustering on a dissimilarity space (S, d), where $k > 1$. The *silhouette coefficient* of an object compares the similarity between an object and other objects located in the same cluster, and the similarity of the same object to objects located in other clusters.
Suppose that $x \in S$ is assigned to the cluster C_p and $\{x\} \subset C_p$. Define

$$a(x) = \frac{1}{|C_p|} \sum \{d(x, u) \mid u \in C_p - \{x\}\}.$$

For $r \neq p$ define $d(x, C_r) = \frac{1}{|C_r|} \sum \{d(x, y) \mid y \in C_r\}$ and

$$b(x) = \min\{d(x, C_r) \mid 1 \leqslant r \leqslant k \text{ and } r \neq p\}.$$

The cluster C_r that defines $b(x)$, that is, $b(x) = d(x, C_r)$ is the *neighbour* of x and represents the second-best choice for object x.

Definition

The *silhouette* of x is the number $s(x)$ defined as

$$s(x) = \frac{b(x) - a(x)}{\max\{a(x), b(x)\}} = \begin{cases} 1 - \frac{a(x)}{b(x)} & \text{if } a(x) < b(x), \\ 0 & \text{if } a(x) = b(x), \\ \frac{b(x)}{a(x)} - 1 & \text{if } a(x) > b(x). \end{cases}$$

If $C_p = \{x\}$ we define $s(x) = 0$.

Note that $-1 \leqslant s(x) \leqslant 1$. When $s(x)$ is close to 1, the within dissimilarity $a(x)$ is much smaller than the smallest between dissimilarity $b(x)$. Therefore, x is well-classified; the second best-choice of a cluster for x is not nearly as closes as the actual choice.
When $a(x)$ is close to 0, then $a(x)$ and $b(x)$ are about the same, hence it not clear whether x has been correctly assigned to C_p.
When $a(x)$ is close to -1, then $a(x)$ is larger than $b(x)$, so x is closer to some cluster other than C_p; we say that x has been missassigned.

Example

Starting from the iris dataset we remove the species attribute by

```
ir <- iris[,1:4]
```

and apply the pam algorithm of the package clust:

```
pamc <- pam(ir,3)
```

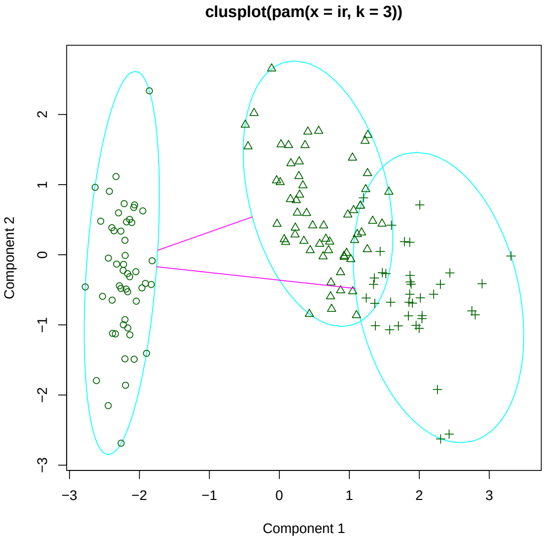
The plot of the pamc object contains two subplots: the clusplot, which we discussed previously and the sihouette plot. These plots can be obtained by writing

```
> pdf(''pamc-clusplot.pdf'')
> plot(pamc,which.plots=1)
> dev.off()
```

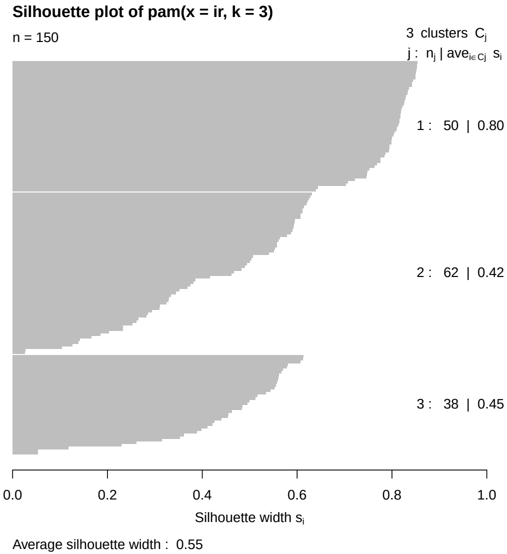
and

```
> pdf(''pamc-silh.pdf'')
> plot(pamc,which.plots=2)
> dev.off()
```

The plot which is generated is determined by the parameter `which.plots` (1 for `clusplot` and 2 for the `silhouette` plot.



These two components explain 95.81 % of the point variability.


Silhouette plot of pam(x = ir, k = 3)
n = 150
3 clusters C_j
j : n_j | $\text{ave}_{i \in C_j}$ s_i
1 : 50 | 0.80
2 : 62 | 0.42
3 : 38 | 0.45
0.0
0.2
0.4
0.6
0.8
1.0
Silhouette width s_i
Average silhouette width : 0.55

The silhouette function can be used to determine the best number of clusters. Consider, for example a uni-dimensional set of objects defined as

```
x <- c(rnorm(50),rnorm(50,mean=5),rnorm(50,mean=15))
```

and define an array w as

```
w <- numeric(20)
```

Then write

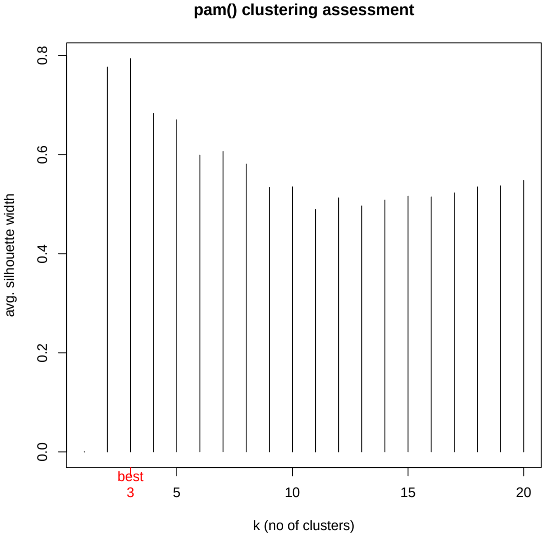
```
x <- c(rnorm(50),rnorm(50,mean=5),rnorm(30,mean=15))
w <- numeric(20)
for(k in 2:20)
   w[k] <- pam(x,k)$silinfo$avg.width
k.best <- which.max(w)

cat(’’silhouette-optimal number of clusters is: ’’,k.best,’’\n’’)

plot(1:20,w,type=’’h’’,main=’’pam() clustering assessment’’,
    xlab=’’k (no of clusters)’’,ylab=’’avg. silhouette width’’)

axis(1,k.best,paste(’’best’’,k.best,sep=’’\n’’),col=’’red’’,col.axi
```

The best value is $k = 3$, as it also follows from the following graph:


pam() clustering assessment
avg. silhouette width
best
k (no of clusters)