Clustering - III

Prof. Dan A. Simovici

UMB
1 Hierarchical Clustering
Hierarchical clustering is a recursive process that begins with a metric space of objects \((S, d)\) and results in a chain of partitions of the set of objects.

In each of the partitions, similar objects belong to the same block and objects that belong to distinct blocks tend to be dissimilar.
In *agglomerative hierarchical clustering*, the construction of this chain begins with the unit partition \( \pi^1 = \alpha_S \). If the partition constructed at step \( k \) is

\[
\pi^k = \{ \pi_1^k, \ldots, \pi_{m_k}^k \},
\]

then two distinct blocks \( \pi_p^k \) and \( \pi_q^k \) of this partition are selected using a *selection criterion*. These blocks are fused and a new partition

\[
\pi^{k+1} = \{ \pi_1^k, \ldots, \pi_{p-1}^k, \pi_{p+1}^k, \ldots, \pi_{q-1}^k, \pi_{q+1}^k, \ldots, \pi_p^k \cup \pi_q^k \}
\]

is formed.

We have \( \pi^k \prec \pi^{k+1} \). The process must end because the poset \( (\text{PART}(S), \preceq) \) is of finite height. The algorithm halts when the one-block partition \( \omega_S \) is reached.
The chain of partitions $\pi^1, \pi^2, \ldots$ generates a hierarchy on the set $S$. Therefore, all tools developed for hierarchies, including the notion of a dendrogram, can be used for hierarchical algorithms.
When data to be clustered are numerical (that is, when $S \subseteq \mathbb{R}^n$), we can define the centroid of a nonempty subset $U$ of $S$ as:

$$c_U = \frac{1}{|U|} \sum \{o|o \in U\}.$$ 

If $\pi = \{U_1, \ldots, U_m\}$ is a partition of $S$, then the sum of the squared errors of $\pi$ is the number

$$sse(\pi) = \sum_{i=1}^{m} \sum \{d^2(o, c_{U_i})|o \in U_i\}, \quad (1)$$

where $d$ is the Euclidean distance in $\mathbb{R}^n$. 
If two blocks $U$ and $V$ of a partition $\pi$ are fused into a new block $W$ to yield a new partition $\pi'$ that covers $\pi$, then the variation of the sum of squared errors is given by

$$sse(\pi') - sse(\pi) = \sum \{d^2(o, c_W) | o \in U \cap V\}$$
$$- \sum \{d^2(o, c_U) | o \in U\} - \sum \{d^2(o, c_V) | o \in V\}.$$ 

The centroid of the new cluster $W$ is given by

$$c_W = \frac{1}{|W|} \sum \{o | o \in W\} = \frac{|U|}{|W|} c_U + \frac{|V|}{|W|} c_V.$$
This allows us to evaluate the increase in the sum of squared errors:

\[
\text{sse}(\pi') - \text{sse}(\pi) = \sum \{d^2(o, c_W) \mid o \in U \cup V\}
\]

\[
= - \sum \{d^2(o, c_U) \mid o \in U\} - \sum \{d^2(o, c_V) \mid o \in V\}
\]

\[
= \sum \{d^2(o, c_W) - d^2(o, c_U) \mid o \in U\} + \sum \{d^2(o, c_W) - d^2(o, c_V) \mid o \in V\}.
\]

Observe that:

\[
\sum \{d^2(o, c_W) - d^2(o, c_U) \mid o \in U\}
\]

\[
= \sum_{o \in U} ((o - c_W)'(o - c_W) - (o - c_U)'(o - c_U))
\]

\[
= |U|(c_W^2 - c_U^2) + 2(c_U - c_W) \sum_{o \in U} o
\]

\[
= |U|(c_W^2 - c_U^2) + 2|U|(c_U - c_W)c_U
\]

\[
= (c_W - c_U)(|U|(c_W + c_U) - 2|U|c_U)
\]

\[
= |U|(c_W - c_U)^2.
\]
Using the equality
\[
\mathbf{c}_W - \mathbf{c}_U = \frac{|U|}{|W|} \mathbf{c}_U + \frac{|V|}{|W|} \mathbf{c}_V - \mathbf{c}_U = \frac{|V|}{|W|} (\mathbf{c}_V - \mathbf{c}_U),
\]
we obtain
\[
\sum \{ d^2(o, \mathbf{c}_W) - d^2(o, \mathbf{c}_U) \mid o \in U \} = \frac{|U||V|^2}{|W|^2} (\mathbf{c}_V - \mathbf{c}_U)^2.
\]
Similarly, we have
\[
\sum \{ d^2(o, \mathbf{c}_W) - d^2(o, \mathbf{c}_V) \mid o \in V \} = \frac{|U|^2|V|}{|W|^2} (\mathbf{c}_V - \mathbf{c}_U)^2,
\]
so,
\[
sse(\pi') - sse(\pi) = \frac{|U||V|}{|W|} (\mathbf{c}_V - \mathbf{c}_U)^2. \quad (2)
\]
The dissimilarity between two clusters $U$ and $V$ can be defined using one of the following real-valued, two-argument functions defined on the set of subsets of $S$:

\[
sl(U, V) = \min\{d(u, v) | u \in U, v \in V\}; \\
cl(U, V) = \max\{d(u, v) | u \in U, v \in V\}; \\
gav(U, V) = \frac{\sum\{d(u, v) | u \in U, v \in V\}}{|U| \cdot |V|}; \\
cen(U, V) = \|\mathbf{c}_U - \mathbf{c}_V\|^2; \\
ward(U, V) = \frac{|U||V|}{|U| + |V|} \left(\mathbf{c}_V - \mathbf{c}_U\right)^2.
\]

The names of the functions $sl$, $cl$, $gav$, and $cen$ defined above are acronyms of the terms “single link”, “complete link”, “group average”, and “centroid”, respectively. They are linked to variants of the hierarchical clustering algorithms.

Note that in the case of the $ward$ function the value equals the increase in the sum of the square errors when the clusters $U$, $V$ are replaced with their union.
Hierarchical Clustering

The computation of the dissimilarity between a new cluster and existing clusters is described next.

**Theorem**

Let $U$ and $V$ be two clusters of the clustering $\pi$ that are joined into a new cluster $W$. Then, if $Q \in \pi - \{U, V\}$, we have

\[
\begin{align*}
sl(W, Q) &= \frac{1}{2}sl(U, Q) + \frac{1}{2}sl(V, Q) - \frac{1}{2}\left|sl(U, Q) - sl(V, Q)\right|; \\
cl(W, Q) &= \frac{1}{2}cl(U, Q) + \frac{1}{2}cl(V, Q) + \frac{1}{2}\left|cl(U, Q) - cl(V, Q)\right|; \\
gav(W, Q) &= \frac{|U|}{|U| + |V|}gav(U, Q) + \frac{|V|}{|U| + |V|}gav(V, Q); \\
cen(W, Q) &= \frac{|U|}{|U| + |V|}cen(U, Q) + \frac{|V|}{|U| + |V|}cen(V, Q) \\
&\quad - \frac{|U||V|}{(|U| + |V|)^2}cen(U, V); \\
ward(W, Q) &= \frac{|U| + |Q|}{|U| + |V| + |Q|}ward(U, Q) + \frac{|V| + |Q|}{|U| + |V| + |Q|}ward(V, Q).
\end{align*}
\]
Proof

The first two equalities follow from the fact that

\[
\min\{a, b\} = \frac{1}{2} (a + b) - \frac{1}{2} |a - b|,
\]

\[
\max\{a, b\} = \frac{1}{2} (a + b) + \frac{1}{2} |a - b|,
\]

for every \(a, b \in \mathbb{R}\).

For the third equality, we have

\[
gav(W, Q) = \frac{\sum\{d(w, q) | w \in W, q \in Q\}}{|W| \cdot |Q|}
\]

\[
= \frac{\sum\{d(u, q) | u \in U, q \in Q\}}{|W| \cdot |Q|} + \frac{\sum\{d(v, q) | v \in V, q \in Q\}}{|W| \cdot |Q|}
\]

\[
= \frac{|U|}{|W|} \frac{\sum\{d(u, q) | u \in U, q \in Q\}}{|U| \cdot |Q|} + \frac{|V|}{|W|} \frac{\sum\{d(v, q) | v \in V, q \in Q\}}{|V| \cdot |Q|}
\]

\[
= \frac{|U|}{|U| + |V|} gav(U, Q) + \frac{|V|}{|U| + |V|} gav(V, Q).
\]
Proof (cont’d)

For the third equality, we have

$$gav(W, Q) = \frac{\sum\{d(w, q)|w \in W, q \in Q\}}{|W| \cdot |Q|}$$

$$= \frac{\sum\{d(u, q)|u \in U, q \in Q\}}{|W| \cdot |Q|} + \frac{\sum\{d(v, q)|v \in V, q \in Q\}}{|W| \cdot |Q|}$$

$$= \frac{|U|}{|W|} \frac{\sum\{d(u, q)|u \in U, q \in Q\}}{|U| \cdot |Q|} + \frac{|V|}{|W|} \frac{\sum\{d(v, q)|v \in V, q \in Q\}}{|V| \cdot |Q|}$$

$$= \frac{|U|}{|U| + |V|} gav(U, Q) + \frac{|V|}{|U| + |V|} gav(V, Q).$$
Hierarchical Clustering

Proof (cont’d)

The equality involving the function $cen$ is immediate. The last equality can be easily translated into

\[
\frac{|Q||W|}{|Q| + |W|} (c_Q - c_W)^2
= \frac{|U| + |Q|}{|U| + |V| + |Q|} \frac{|U||Q|}{|U| + |Q|} (c_Q - c_U)^2 \\
+ \frac{|V| + |Q|}{|U| + |V| + |Q|} \frac{|V||Q|}{|V| + |Q|} (c_Q - c_V)^2 \\
- \frac{|Q|}{|U| + |V| + |Q|} \frac{|U||V|}{|U| + |V|} (c_Q - c_U)^2,
\]

which can be verified replacing $|W| = |U| + |V|$ and $c_W = \frac{|U|}{|W|} c_U + \frac{|V|}{|W|} c_V$. 
These equalities can be presented as a single equality involving several coefficients. (the Lance-Williams Formula): Let $U$ and $V$ be two clusters of the clustering $\pi$ that are joined into a new cluster $W$. Then, if $Q \in \pi - \{U, V\}$, the dissimilarity between $W$ and $Q$ can be expressed as

$$d(W, Q) = a_U d(U, Q) + a_V d(V, Q) + bd(U, V) + c|d(U, Q) - d(V, Q)|,$$

where the coefficients $a_U, a_V, b, c$ are given by the following table:

<table>
<thead>
<tr>
<th>Function</th>
<th>$a_U$</th>
<th>$a_V$</th>
<th>$b$</th>
<th>$c$</th>
</tr>
</thead>
<tbody>
<tr>
<td>sl</td>
<td>$\frac{1}{2}$</td>
<td>$\frac{1}{2}$</td>
<td>0</td>
<td>$-\frac{1}{2}$</td>
</tr>
<tr>
<td>cl</td>
<td>$\frac{1}{2}$</td>
<td>$\frac{1}{2}$</td>
<td>0</td>
<td>$\frac{1}{2}$</td>
</tr>
<tr>
<td>gav</td>
<td>$</td>
<td>U</td>
<td>$</td>
<td>$</td>
</tr>
<tr>
<td>cen</td>
<td>$</td>
<td>U</td>
<td>+</td>
<td>V</td>
</tr>
<tr>
<td>ward</td>
<td>$</td>
<td>U</td>
<td>+</td>
<td>V</td>
</tr>
</tbody>
</table>
Hierarchical Clustering

The variant of the algorithm that makes use of the function sl is known as the *single-link* clustering. It tends to favor elongated clusters. We use single-link clustering for the metric space \((S, d_1)\), where \(S \subseteq \mathbb{R}^2\) consists of seven objects, \(S = \{o_1, \ldots, o_7\}\) (see Figure ??).
The distances $d_1(o_i, o_j)$ for $1 \leq i, j \leq 7$ between the objects of $S$ are specified by the $7 \times 7$ matrix

$$D^1 = \begin{pmatrix}
0 & 1 & 3 & 6 & 8 & 11 & 10 \\
1 & 0 & 2 & 5 & 7 & 10 & 9 \\
3 & 2 & 0 & 3 & 5 & 8 & 7 \\
6 & 5 & 3 & 0 & 2 & 5 & 4 \\
8 & 7 & 5 & 2 & 0 & 3 & 4 \\
11 & 10 & 8 & 5 & 3 & 0 & 3 \\
10 & 9 & 7 & 4 & 4 & 3 & 0
\end{pmatrix}.$$
Hierarchical Clustering

Single-link Clustering

\[ \pi^1 = \{\{o_i\} \mid 1 \leq i \leq 7\}\{\{o_1\}, \{o_2\}, \{o_3\}, \{o_4\}, \{o_5\}, \{o_6\}, \{o_7\}\}. \]

Two of the closest clusters are \{o_1\}, \{o_2\}; these clusters are fused into the cluster \{o_1, o_2\}, the new partition is

\[ \pi^2 = \{\{o_1, o_2\}, \ldots, \{o_7\}\}, \]

and the matrix of dissimilarities becomes the \(6 \times 6\)-matrix

\[ D^2 = \begin{pmatrix}
0 & 2 & 5 & 7 & 10 & 9 \\
2 & 0 & 3 & 5 & 8 & 7 \\
5 & 3 & 0 & 2 & 5 & 4 \\
7 & 5 & 2 & 0 & 3 & 4 \\
10 & 8 & 5 & 3 & 0 & 3 \\
9 & 7 & 4 & 4 & 3 & 0 \\
\end{pmatrix}. \]
The next pair of closest clusters is \( \{o_1, o_2\} \) and \( \{o_3\} \). These clusters are fused into the cluster \( \{o_1, o_2, o_3\} \), and the new \( 5 \times 5 \)-matrix is:

\[
D^3 = \begin{pmatrix}
0 & 3 & 5 & 8 & 7 \\
3 & 0 & 2 & 5 & 4 \\
5 & 2 & 0 & 3 & 4 \\
8 & 5 & 3 & 0 & 3 \\
7 & 4 & 4 & 3 & 0 \\
\end{pmatrix},
\]

which corresponds to the partition

\[
\pi^3 = \{\{o_1, o_2, o_3\}, \{o_4\}, \ldots, \{o_7\}\}.
\]
Hierarchical Clustering

Next, the closest clusters are \( \{ \mathbf{o}_4 \} \) and \( \{ \mathbf{o}_5 \} \). Fusing these yields the partition

\[
\pi^4 = \{ \{ \mathbf{o}_1, \mathbf{o}_2, \mathbf{o}_3 \}, \{ \mathbf{o}_4, \mathbf{o}_5 \}, \{ \mathbf{o}_6 \}, \{ \mathbf{o}_7 \} \}
\]

and the \( 4 \times 4 \)-matrix

\[
D^4 = \begin{pmatrix}
0 & 3 & 8 & 7 \\
3 & 0 & 3 & 4 \\
8 & 3 & 0 & 3 \\
7 & 4 & 3 & 0
\end{pmatrix}
\]
We have three choices now since there are three pairs of clusters at distance 3 of each other: 

\[
\{\mathbf{o}_1, \mathbf{o}_2, \mathbf{o}_3\}, \{\mathbf{o}_4, \mathbf{o}_5\}, \{\mathbf{o}_6\}, \{\mathbf{o}_7\},
\]

and

\[
\{\mathbf{o}_4, \mathbf{o}_5\}, \{\mathbf{o}_6\}, \{\mathbf{o}_7\}.
\]

By choosing to fuse the first pair, we obtain the partition

\[
\pi^5 = \left\{ \{\mathbf{o}_1, \mathbf{o}_2, \mathbf{o}_3, \mathbf{o}_4, \mathbf{o}_5\}, \{\mathbf{o}_6\}, \{\mathbf{o}_7\} \right\},
\]

which corresponds to the $3 \times 3$-matrix

\[
D^5 = \begin{pmatrix}
0 & 3 & 4 \\
3 & 0 & 3 \\
4 & 3 & 0
\end{pmatrix}.
\]

Observe that the large cluster $\{\mathbf{o}_1, \mathbf{o}_2, \mathbf{o}_3, \mathbf{o}_4, \mathbf{o}_5\}$ formed so far has an elongated shape, which is typical for single-link variant of the algorithm.
Next, we coalesce \( \{o_1, o_2, o_3, o_4, o_5\} \) and \( \{o_6\} \), which yields

\[
\pi^6 = \{\{o_1, o_2, o_3, o_4, o_5, o_6\}, \{o_7\}\}
\]

and

\[
D^6 = \begin{pmatrix} 0 & 3 \\ 3 & 0 \end{pmatrix}.
\]

Finally, we join the last two clusters, and the clustering is completed.
Hierarchical Clustering

The dendrogram

Dendrogram of single-link clustering.