Aglomerative Clustering and Clustering Evaluation

David Mareček

December 09, 2021
Hierarchical clustering

Each observation starts in its own cluster, and clusters are successively merged together.

The linkage criteria:

- **Minimum (Single linkage)** minimizes the minimum distance between pairs of clusters.
  \[ d_{\text{single}} = \min \{ d(x_i, x_j), \ i \in A, \ j \in B \} \]

- **Maximum (Complete linkage)** minimizes the maximum distance between pairs of clusters.
  \[ d_{\text{complete}} = \max \{ d(x_i, x_j), \ i \in A, \ j \in B \} \]

- **Average linkage** minimizes the average of the distances between all observations of pairs of clusters.
  \[ d_{\text{average}} = \frac{1}{|A| \cdot |B|} \sum_{i \in A} \sum_{j \in B} d(x_i, x_j) \]

- **Centroid linkage** minimizes the distance between centers of clusters.
  \[ d_{\text{centroid}} = d\left( \frac{1}{|A|} \sum_{i \in A} x_i, \frac{1}{|B|} \sum_{i \in B} x_i \right) \]
Hierarchical clustering

- **Ward linkage** is a variance minimizing approach. The distance between two clusters $A$ and $B$ is how much the sum of squares will increase when we merge them. It is similar to the k-means objective function but tackled with an agglomerative hierarchical approach.

\[
d_{\text{Ward}}(A, B) = \sum_{i \in A \cup B} ||x_i - m_{A \cup B}||^2 - \sum_{i \in A} ||x_i - m_A||^2 - \sum_{i \in B} ||x_i - m_B||^2,
\]

where $m_X$ is the mean (center) of cluster $X$. It also corresponds to the squared distance between the centers of the clusters

\[
d_{\text{Ward}}(A, B) = \frac{n_A n_B}{n_A + n_B} ||m_A - m_B||^2,
\]

where $n_A$ and $n_B$ are number of points in clusters A and B, respectively.
Hierarchical clustering

Hierarchical Clustering Dendrogram

Number of points in node (or index of point if no parenthesis).
Clustering Methods Comparison

<table>
<thead>
<tr>
<th>Method</th>
<th>Image 1</th>
<th>Image 2</th>
<th>Image 3</th>
<th>Image 4</th>
<th>Image 5</th>
<th>Image 6</th>
<th>Image 7</th>
<th>Image 8</th>
<th>Image 9</th>
<th>Image 10</th>
</tr>
</thead>
<tbody>
<tr>
<td>MiniBatchKMeans</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MeanShift</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SpectralClustering</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ward</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AgglomerativeClustering</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DBSCAN</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>OPTICS</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Birch</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GaussianMixture</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
If you have a testing data available with annotated gold label:

Rosenberg and Hirschberg (2007) define the following objectives for any cluster assignment:

- **Homogeneity** – each cluster contains only members of a single class
- **Completeness** – all members of a given class are assigned to the same cluster
- **V-measure** – their harmonic mean

If you do not have any labelled data:

- **Silhouette coefficient** – “unsupervised” consistency within clusters of data
Homogeneity

Homogeneity – To what extent each cluster contains only members of a single class?

\[ h = 1 - \frac{H(C|K)}{H(C)} \]

\( H(C|K) \) is the conditional entropy of the cluster assignments given the classes:

\[ H(C|K) = - \sum_{c=1}^{\left| C \right|} \sum_{k=1}^{\left| K \right|} \frac{n_{c,k}}{n} \log \frac{n_{c,k}}{n_k} \]

\( H(C) \) is the entropy of the classes:

\[ H(C) = - \sum_{c=1}^{\left| C \right|} \frac{n_c}{n} \log \frac{n_c}{n} \]
Completeness

Completeness – To what extent all members of a given class are assigned to the same cluster?

\[ c = 1 - \frac{H(K|C)}{H(K)} \]

\( H(K|C) \) is the conditional entropy of the classes given the cluster assignments:

\[ H(K|C) = -\sum_{c=1}^{|C|} \sum_{k=1}^{|K|} \frac{n_{c,k}}{n} \log \frac{n_{c,k}}{n_c} \]

\( H(K) \) is the entropy of the clusters:

\[ H(K) = -\sum_{k=1}^{|K|} \frac{n_k}{n} \log \frac{n_k}{n} \]
V-measure

V-measure – Harmonic mean of homogeneity and completeness:

\[ v = \frac{2 \cdot h \cdot c}{h + c} \]
Homogeneity and Completeness

(a) Homogeneity = 1
Completeness < 1

(b) Homogeneity = 0
Completeness = 1

(c) Completeness = 1
Homogeneity < 1

(d) Completeness = 0
Homogeneity = 1
Silhouette coefficient

How similar an object is to its own cluster (cohesion) compared to other clusters (separation)

Values between -1 and 1

The Silhouette Coefficient $s$ for a single sample is then given as:

$$s_i = \frac{b_i - a_i}{\max(a_i, b_i)}, \quad a_i = \frac{1}{C_I - 1} \sum_{j \in C_I, i \neq j} d(i, j), \quad b_i = \min_{J \neq I} \frac{1}{C_J} \sum_{j \in C_J} d(i, j)$$

- $a$ is the mean distance between a sample and all other points in the same cluster
- $b$ is the mean distance between a sample and all other points in the next nearest cluster

The mean over all points of a cluster is a measure of how tightly grouped all the points in the cluster are. Thus the mean over all data of the entire dataset is a measure of how appropriately the data have been clustered.