Hierarchical clustering

Based in part on slides from textbook, slides of Susan Holmes

©Jonathan Taylor

December 2, 2012
Hierarchical clustering

Description

- Produces a set of nested clusters organized as a hierarchical tree.
- Can be visualized as a *dendrogram*: A tree like diagram that records the sequences of merges or splits.
Hierarchical clustering

- Produces a set of nested clusters organized as a hierarchical tree.
- Can be visualized as a dendrogram - a tree-like diagram that records the sequences of merges or splits.

A clustering and its dendrogram.
Hierarchical clustering

**Strengths**

- Do not have to assume any particular number of clusters. Each horizontal cut of the tree yields a clustering.
- The tree may correspond to a meaningful taxonomy: (e.g., animal kingdom, phylogeny reconstruction, ...)
- Need only a similarity or distance matrix for implementation.
Hierarchical clustering

**Agglomerative**

- Start with the points as individual clusters.
- At each step, merge the closest pair of clusters until only one cluster (or some fixed number $k$ clusters) remain.
Hierarchical clustering

**Divisive**

- Start with one, all-inclusive cluster.
- At each step, split a cluster until each cluster contains a point (or there are $k$ clusters).
Hierarchical clustering

Agglomerative Clustering Algorithm

1. Compute the proximity matrix.
2. Let each data point be a cluster.
3. **While** there is more than one cluster:
   1. Merge the two closest clusters.
   2. Update the proximity matrix.

The major difference is the computation of proximity of two clusters.
Hierarchical clustering

Starting point for agglomerative clustering.
Hierarchical clustering

Intermediate point, with 5 clusters.
Hierarchical clustering

We will merge $C_2$ and $C_5$. 
Hierarchical clustering

How do we update the proximity matrix?

Proximity Matrix

<table>
<thead>
<tr>
<th></th>
<th>C1</th>
<th>C5</th>
<th>C3</th>
<th>C4</th>
</tr>
</thead>
<tbody>
<tr>
<td>C1</td>
<td>?</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C2 U C5</td>
<td>?</td>
<td>?</td>
<td>?</td>
<td>?</td>
</tr>
<tr>
<td>C3</td>
<td>?</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C4</td>
<td>?</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Hierarchical clustering

We need a notion of similarity between clusters.
Hierarchical clustering

Single linkage uses the minimum distance.
Hierarchical clustering

Complete linkage uses the maximum distance.
Hierarchical clustering

Group average linkage uses the average distance between groups.
Hierarchical clustering

Centroid uses the distance between the centroids of the clusters (presumes one can compute centroids...)}
Cluster Similarity: MIN or Single Link

- Similarity of two clusters is based on the two most similar (closest) points in the different clusters
- Determined by one pair of points, i.e., by one link in the proximity graph.

Proximity matrix and dendrogram of single linkage.
Hierarchical clustering

Distance matrix for nested clusterings

\[
\begin{pmatrix}
1 & 0.24 & 0.22 & 0.37 & 0.34 & 0.23 \\
2 & 0.15 & 0.20 & 0.14 & 0.25 \\
3 & 0.15 & 0.28 & 0.11 \\
4 & 0.29 & 0.22 \\
5 & 0.39 \\
6
\end{pmatrix}
\]
Hierarchical clustering

Nested cluster representation and dendrogram of single linkage.
Hierarchical clustering

Single linkage

- Can handle irregularly shaped regions fairly naturally.
- Sensitive to noise and outliers in the form of “chaining”.
The Iris data (single linkage)
The Iris data (single linkage)
Hierarchical clustering

Cluster Similarity: MAX or Complete Linkage

- Similarity of two clusters is based on the two least similar (most distant) points in the different clusters
- Determined by all pairs of points in the two clusters

Proximity matrix and dendrogram of complete linkage.
Hierarchical clustering

Nested cluster and dendrogram of complete linkage.
Hierarchical clustering

Complete linkage

- Less sensitive to noise and outliers than single linkage.
- Regions are generally compact, but may violate “closeness”. That is, points may much closer to some points in neighbouring cluster than its own cluster.
- This manifests itself as breaking large clusters.
- Clusters are biased to be globular.
The Iris data (complete linkage)
The Iris data (complete linkage)
The Iris data (complete linkage)
Hierarchical clustering

Nested cluster and dendrogram of group average linkage.
Hierarchical clustering

**Average linkage**

- Given two elements of the partition $C_r, C_s$, we might consider

$$d_{GA}(C_r, C_s) = \frac{1}{|C_r||C_s|} \sum_{x \in C_r, y \in C_s} d(x, y)$$

- A compromise between single and complete linkage.
- Shares globular clusters of complete, less sensitive than single.
The Iris data (average linkage)
The Iris data (average linkage)
Hierarchical clustering

Ward’s linkage

- Similarity of two clusters is based on the increase in squared error when two clusters are merged.
- Similar to average if dissimilarity between points is distance squared. Hence, it shares many properties of average linkage.
- A hierarchical analogue of $K$-means.
- Sometimes used to initialize $K$-means.
The Iris data (Ward’s linkage)

cdis(iris[, -5])
hclust(*, "ward")
The Iris data (Ward’s linkage)
NCI data (complete linkage)
NCI data (single linkage)
NCI data (average linkage)
NCI data (Ward’s linkage)
Hierarchical clustering

Computational issues

- $O(n^2)$ space since it uses the proximity matrix.
- $O(n^3)$ time in many cases as there are $N$ steps, and at each step a matrix of size $N^2$ must be updated and/or searched.
Hierarchical clustering

**Statistical issues**

- Once a decision is made to combine two clusters, it cannot be undone.
- No objective function is directly minimized.
- Different schemes have problems with one or more of the following:
  - Sensitivity to noise and outliers.
  - Difficulty handling different sized clusters and convex shapes.
  - Breaking large cluster.