Statistical Machine Learning — 25 — H Clustering — 12/01/2015

Lecturer: Prof. Homrighausen — Scribe: Lyuou Zhang

1 Hierarchical Clustering

1.1 Linkages

Notation: Define $X_1, \ldots, X_n$ to be the data

Let the dissimiliarities be d_{ij} between each pair X_i, X_j

At any level, clustering assignments can be expressed by sets $G = \{i_1, i_2, \ldots, i_r\}$. given the indices of points in this group. Define $|G|$ to be the size of G.

Linkage: The function $d(G, H)$ that takes two groups G, H and returns the linkage distance between them.

Agglomerative clustering, given the linkage:

- Start with each point in its own group
- Until there is only one cluster, repeatedly merge the two groups G, H that minimize $d(G, H)$.

1.2 Single linkage

In single linkage (a.k.a nearest-neighbor linkage), the linkage distance between G, H is the smallest dissimilarity between two points in different groups:

$$d_{\text{single}}(G, H) = \min_{i \in G,\, j \in H} d_{ij}$$

1.3 Complete linkage

In complete linkage (i.e. farthest-neighbor linkage), linkage distance between G, H is the largest dissimilarity between two points in different clusters:

$$d_{\text{complete}}(G, H) = \max_{i \in G,\, j \in H} d_{ij}.$$

1.4 Average linkage

In average linkage, the linkage distance between G, H is the average dissimilarity over all points in different clusters:

$$d_{\text{average}}(G, H) = \frac{1}{|G| \cdot |H|} \sum_{i \in G,\, j \in H} d_{ij}.$$

1.5 Common properties

Single, complete, and average linkage share the following:

- They all operate on the dissimilarities d_{ij}. This means that the points we are clustering can be quite general (number of mutations on a genome, polygons, faces, whatever).
- Running agglomerative clustering with any of these linkages produces a dendrogram with no inversions.

No inversions means that the linkage distance between merged clusters only increases as we run the algorithm.

1.6 Shortcomings of single and complete linkage

Single linkage:	Often suffers from chaining, that is, we only need a single pair of points to be close to merge two clusters. Therefore, clusters can be too spread out and not compact enough.
Complete linkage:	Often suffers from crowding, that is, a point can be closer to points in other clusters than to points in its own cluster. Therefore, the clusters are compact, but not far enough apart.

Average linkage tries to strike a balance between these two.

1.7 Shortcomings of average linkage

- It isn't clear what properties the resulting clusters have when we cut an average linkage tree.
- Results of average linkage clustering can change with a monotone increasing transformation of the dissimilarities (that is, if we changed the distance, but maintained the ranking of the distances, the cluster solution could change).

Neither of these problems afflict single or complete linkage.

1.8 Hierarchical agglomerative clustering in R

The function `hclust` in base **R** performs the necessary computations. E.g.

```
Delta = dist(x)
out.average = hclust(Delta,method='average')
plot(out.average)
```

1.9 Centroid linkage

Centroid linkage is a commonly used and relatively new approach. Assume

- $X_i \in \mathbb{R}^p$
- $d_{ij} = ||X_i - X_j||_2^2$

Let $\overline{X}_G$ and $\overline{X}_H$ denote group averages for G, H. Then

$$d_{\text{centroid}} = ||\overline{X}_G - \overline{X}_H||_2^2$$

Centroid linkage is

- ... quite intuitive
- ... widely used
- ... nicely analogous to K-means.
- ... very related to average linkage (and much, much faster)

However, it has a very unsavory feature: inversions.

1.10 Linkages summary

	No inversions?	Unchanged w/ monotone transformation?	Cut interpretation?	Notes
Single	✓	✓	✓	chaining
Complete	✓	✓	✓	crowding
Average	✓	X	X	
Centroid	X	X	X	inversions

Cut interpretation Suppose we cut the tree at height $h = 1$.

Single	For each point X_i, there is another point X_j in the same cluster with $d_{ij} \leq 1$ (assuming more than 1 point in cluster). Also, no points in different clusters are closer than 1.
Complete	For each point X_i, every other point X_j in the same cluster has $d_{ij} \leq 1$.