Clustering

Hierarchical Clustering

One potential disadvantage of k-means clustering is that it requires us to pre-specify the number of clusters k. Hierarchical clustering is an alternative approach that does not require that we commit to a particular choice of k. Hierarchical clustering has an added advantage over k-means clustering in that it yields an attractive tree-based representation of the observations, known as a *dendrogram*. In the next section, we describe bottom-up or agglomerative clustering. This is the most common type of hierarchical clustering, referring to the fact that a dendrogram (generally depicted as an upside-down tree) is used.

Agglomerative Clustering Algorithm

The hierarchical clustering dendrogram is generated using a straightforward algorithm. We begin by defining some sort of *similarity* measure between each pair of observations. Most often, Euclidean distance is used; we will discuss the choice of dissimilarity measure later in this chapter. The algorithm proceeds iteratively. Starting out at the bottom of the dendrogram, each of the n observations is treated as its own cluster. The two clusters that are most similar to each other are then *fused* so that there are now $n-1$ clusters. Next, the two clusters that are most similar to each other are *fused* again, so that there are now $n-2$ clusters. The algorithm proceeds in this fashion until all of the observations belong to one single cluster, and the dendrogram is complete.

Agglomerative Clustering Algorithm

1. Begin with n observations and a measure (such as Euclidean distance) of all the $\binom{n}{2} = n(n-1)/2$ pairwise dissimilarities. Treat each observation as its own cluster.
2. For $i = n, n-1, \ldots, 2$:
   (a) Examine all pairwise inter-cluster dissimilarities among the i clusters and identify the pair of least dissimilar clusters (that is, most similar). Fuse these two clusters. The similarity between these two clusters indicates the height between these two clusters indicate the height in the dendrogram at which the fusion at where the fusion should be placed.
   (b) Compute the new pairwise inter-cluster similarities among the $i-1$ remaining clusters.

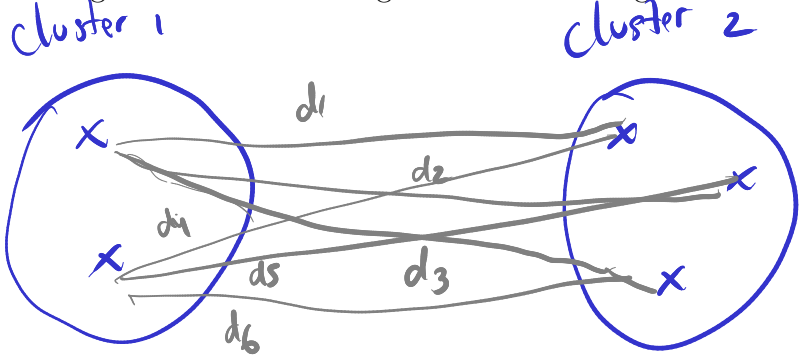
The concept of similarity between a pair of observations needs to be extended to a pair of groups of observations. This extension is achieved by developing the notion of linkage, which defines the

dissimilarity between two groups of observations. The four most common types of linkage are: *complete*, *average*, *single*, and *centroid*. They are briefly described in the following table.

Linkage Description

Linkage	Description
Complete	Maximal inter-cluster similarity. Compute all pairwise similarities between the observations in cluster A and the observations in cluster B, and record the *largest* of these dissimilarities.
Single	Minimal inter-cluster similarity. Compute all pairwise similarities between the observations in cluster A and the observations in cluster B, and record the *smallest* of these similarities. Single linkage can result in extended, trailing clusters in which single observations are fused one at a time.
Average	Mean inter-cluster similarity. Compute all pairwise similarities between the observations in cluster A and the observations in cluster B, and record the *average* of these similarities.
Centroid	Similarity between the centroid for cluster A (a mean vector of length p) and the centroid for cluster B. Centroid linkage can result in undesirable *inversions*.

Average and complete linkage are most popular among statisticians because they tend to yield balanced dendrogram. Centroid linkage is often used in genomics.



Complete linkage = $\max(d_1, d_2, d_3, d_4, d_5, d_6)$

single linkage = $\min(d_1, d_2, d_3, d_4, d_5, d_6)$

Average linkage = $\frac{d_1+d_2+d_3+d_4+d_5+d_6}{6}$