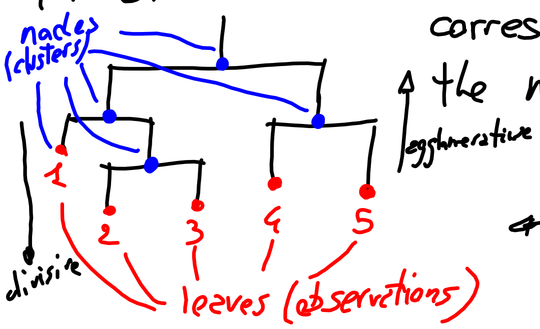
Hierarchical clustering

We saw that the biggest limitation of k-means is the choice of the number of clusters. Alternatives, called hierarchical clustering methods, organize the observations in groups in a hierarchical fashion, so the creation of a new cluster always results in splitting in two an existing cluster (or the other way around)

The structure is that of a binary tree, where the leaves correspond to the observations (in red) and the nodes to clusters (in blue)



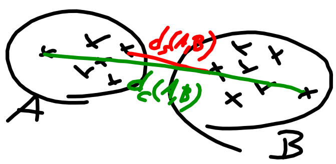
← the plot is called dendrogram

To create the dendrogram (i.e., perform the clustering), we can proceed in two ways

- agglomerative: starting from the leaves (the situation in which $K=n$, each observation is in its own cluster), we proceed by consecutive aggregation of the clusters with smallest dissimilarity, until we reach the situation of $K=1$ (one single cluster)
- divisive: starting from the root (all the observations in the same cluster) we proceed by consecutive splitting of the groups by separating the observations with largest dissimilarity, until $K=n$

To perform hierarchical clustering is necessary to define the dissimilarity between two clusters. When $K=n$, obviously it is just $d(i,i')$, but in a later stage, when we have more observations in the same group, we need to define $d(G,G')$, where G and G' are two groups. There are several options, the traditional ones are:

- single link: $d_S(G,G') = \min_{i \in G, i' \in G'} d(i,i')$
- complete link: $d_C(G,G') = \max_{i \in G, i' \in G'} d(i,i')$
- average link: $d_A(G,G') = \frac{1}{n_G n_{G'}} \sum_{i \in G} \sum_{i' \in G'} d(i,i')$



Grouping based on different measures results in different cluster structures
- single link tends to work better in recovering filiform types of structures
- complete link " " " " " " spheroidal " " "

Choice of k

If we need find the best k we can use the same principle used in K-means ("useful" splits result in larger decreases in D_{within}, "useless" splits in small decreases in D_{within}).
We construct the vertical lines of the dendrogram proportional to the decrease in D_{within} $\longrightarrow$ cut where the lines are longer.