**Hierarchical clustering**

Advantages:

1. It allows us to plot dendrograms (visualizations of a binary hierarchical clustering), which can help with the interpretation of the results by creating meaningful taxonomies.
2. We do not need to specify the number of clusters upfront

Two main approaches:

1. Agglomerative

Take the opposite approach to divisive. Starting with each sample as an individual cluster and merge the closet pairs of clusters until only one cluster remains.

1. Divisive:

Start with one cluster that encompasses all samples, and iteratively split the cluster into smaller clusters until each cluster only contains one sample.

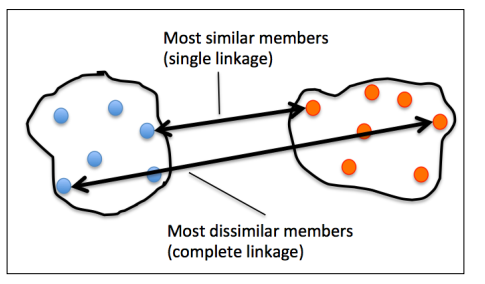
Two standard algorithms for Agglomerative:

1. Single linkage

Compute the distances between the most similar members for each pair of clusters and merge the two clusters for which the distance between the most similar members in the smallest.

1. Complete linkage

Similar to single linkage but, instead of comparing the most similar members in each pair of clusters, we compare the most dissimilar members to perform the merge.



1. Average linkage

Merge the cluster pairs based on the minimum average distances between all group members in two clusters.

1. Ward’s linkage

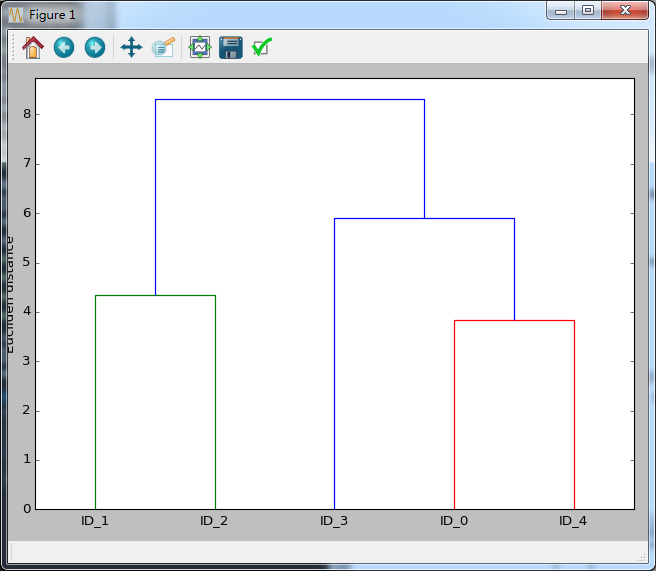
Those two clusters that lead to the minimum increase of the total within-cluster SSE are merged

Steps of agglomerative clustering using the complete linkage approach:

1. Compute the distance matrix of all samples
2. Represent each data point as a singleton cluster
3. Merge the two closet clusters based on the distance of the most dissimilar (distant) members
4. Update the similarity matrix
5. Repeat steps 2 to 4 until one single cluster remains

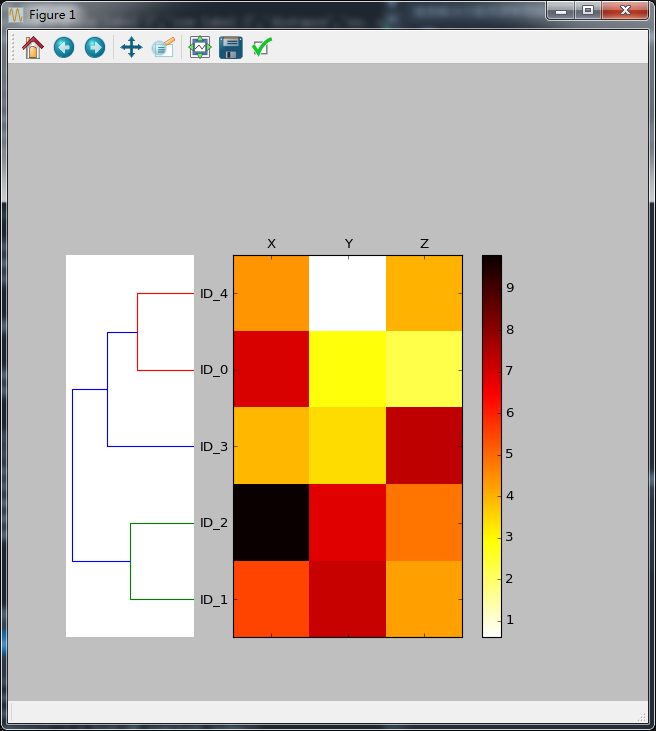
Python code:

1. Distance Matrix: (pdist function from Scipy’s spatial.distance)
2. Complete linkage agglomeration: linkage function from scipy’s cluster.hierarchy submodule ,which returns a so-called linkage matrix
3. visulaize the results in the form of a dendrogram



1. attaching dendrograms to a heat map

code:



. In addition to a simple dendrogram, the color-coded values of each sample and feature in the heat map provide us with a nice summary of the dataset.