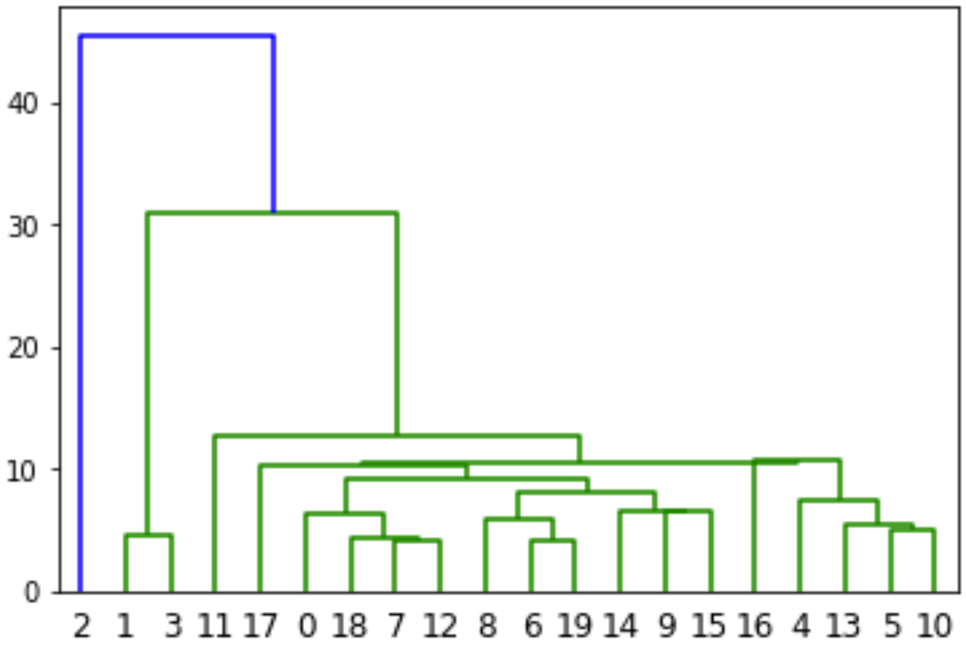
Hierarchical Clustering

Part 3



Distance between two clusters

Cluster P and cluster Q are merged and formed a new cluster denoted by $P+Q$.
Cluster P contains n_P objects and cluster Q contains n_Q objects. Another Cluster R contains n_R objects.
The distance between cluster R and cluster $P+Q$ is given below for few linkage methods.

Distance between two clusters

Cluster P and cluster Q are merged and formed a new cluster denoted by $P+Q$.
Cluster P contains n_P objects and cluster Q contains n_Q objects. Another Cluster R contains n_R objects.
The distance between cluster R and cluster $P+Q$ is given below for few linkage methods.

$$d(R,P+Q) = w_1 d(R, P) + w_2 d(R,Q) + w_3 d(P,Q) + w_4 |d(R, P) - d(R,Q)|$$

where the weights w_1, w_2, w_3, w_4 are method specific, provided by the table below:

Name	w_1	w_2	w_3	w_4
Single	1/2	1/2	0	-1/2
Complete	1/2	1/2	0	1/2
Average	$n_P/(n_P+n_Q)$	$n_Q/(n_P+n_Q)$	0	0
Weighted	1/2	1/2	0	0
Centroid	$n_P/(n_P+n_Q)$	$n_Q/(n_P+n_Q)$	$-(n_P n_Q)/(n_P+n_Q)^2$	0
Median	1/2	1/2	-1/4	0
Ward	$(n_R+n_P)/(n_R+n_P+n_Q)$	$(n_R+n_Q)/(n_P+n_P+n_Q)$	$n_R/(n_R+n_P+n_Q)$	0
Felxibeta	$(1-\beta)/2$	$(1-\beta)/2$	β	0

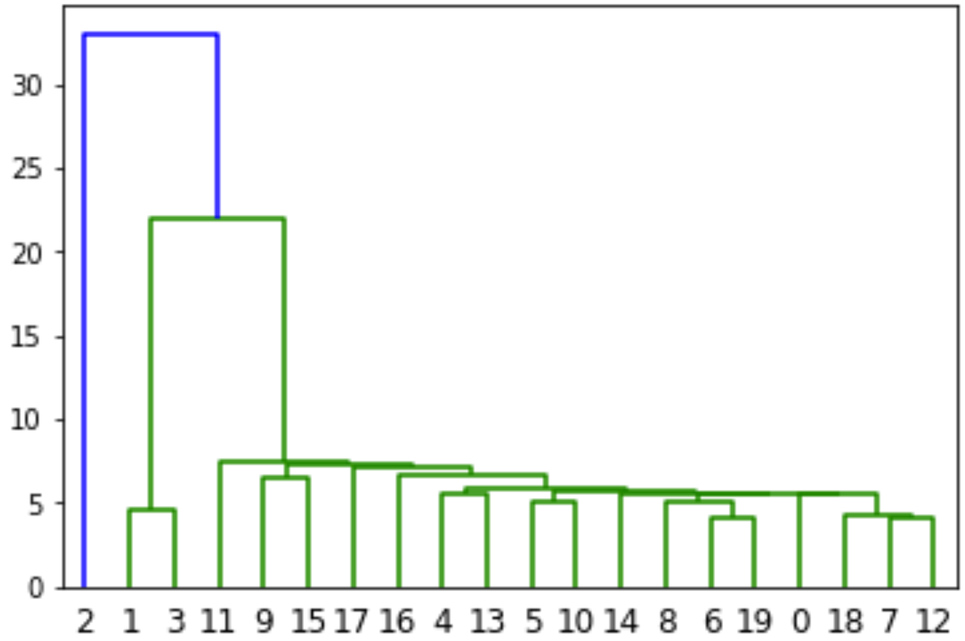
Linkage method: Single

Distance between clusters C_i and C_j is the minimum distance between any object in C_i and any object in C_j

Can handle non-elliptical shapes

Sensitive to noise and outliers

It produces stretched out clusters



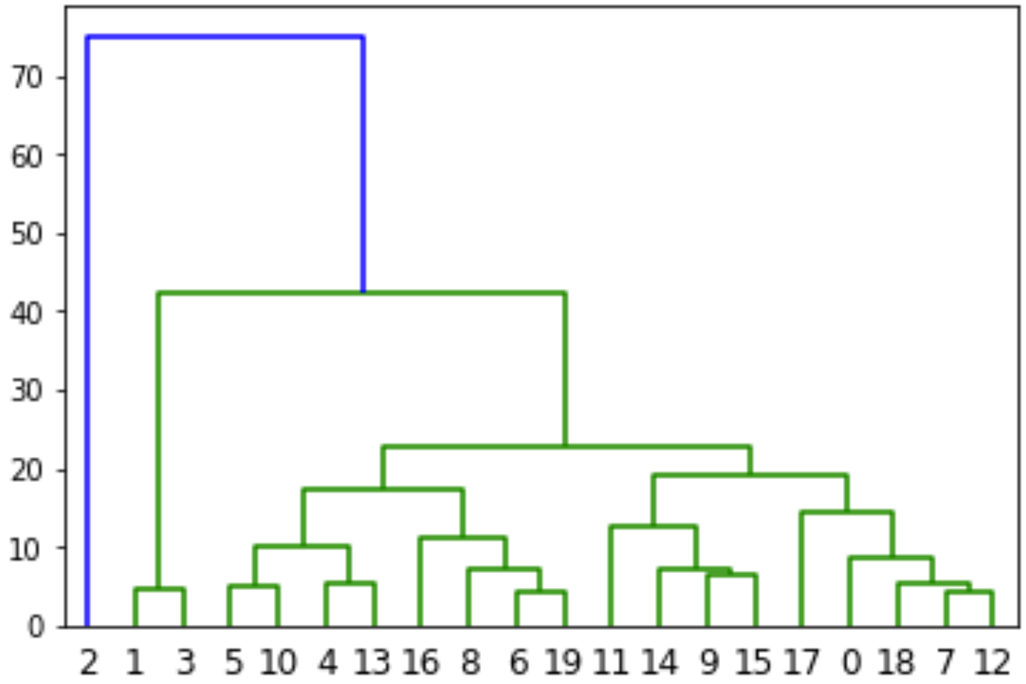
Hierarchical clustering with single linkage method

Linkage method: Complete

Distance between clusters C_i and C_j is the maximum distance between any object in C_i and any object in C_j

More equal sized or balanced clusters

Less vulnerable to noise



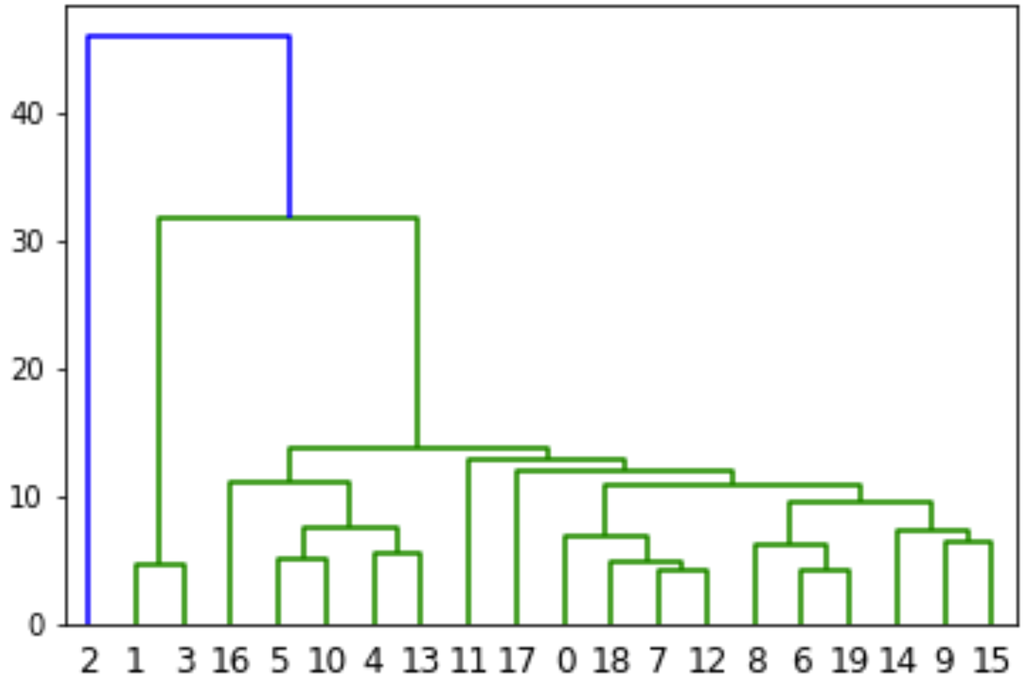
Hierarchical clustering with complete linkage method

Linkage method: Average

Distance between clusters C_i and C_j is the average distance between any object in C_i and any object in C_j

In-between single linkage and complete linkage

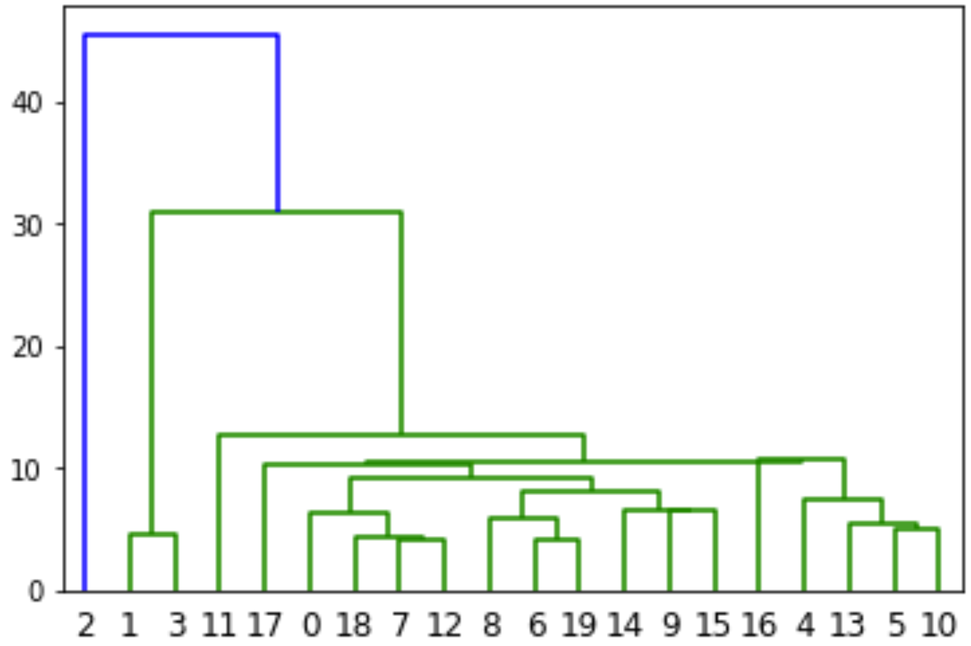
Less vulnerable to noise



Hierarchical clustering with average linkage method

Linkage method: Centroid

Distance between clusters C_i and C_j is the distance between centre of C_i and centre of C_j



Hierarchical clustering with centroid linkage method

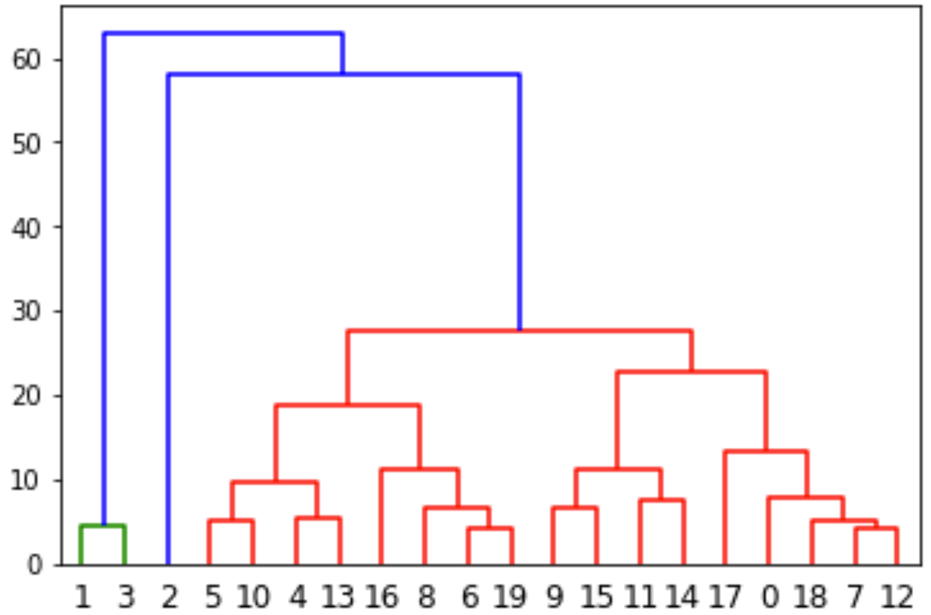
Linkage method: Ward

Distance between clusters C_i and C_j is the difference between the total within cluster sum of squares for the two clusters separately, and the within cluster sum of squares resulting from merging the two clusters in cluster C_{ij}

Behave like k-means clustering in hierarchical framework

Suitable for initializing seed in k-means clustering

Less vulnerable to noise



Hierarchical clustering with ward linkage method

Distance between two clusters

Cluster P and cluster Q are merged and formed a new cluster denoted by $P+Q$.
Cluster P contains n_P objects and cluster Q contains n_Q objects. Another Cluster R contains n_R objects.
The distance between cluster R and cluster $P+Q$ is given below for few linkage methods.

$$d(R,P+Q) = w_1 d(R, P) + w_2 d(R,Q) + w_3 d(P,Q) + w_4 |d(R, P) - d(R,Q)|$$

where the weights w_1, w_2, w_3, w_4 are method specific, provided by the table below:

Name	w_1	w_2	w_3	w_4
Single	1/2	1/2	0	-1/2
Complete	1/2	1/2	0	1/2
Average	$n_P/(n_P+n_Q)$	$n_Q/(n_P+n_Q)$	0	0
Weighted	1/2	1/2	0	0
Centroid	$n_P/(n_P+n_Q)$	$n_Q/(n_P+n_Q)$	$-(n_P n_Q)/(n_P+n_Q)^2$	0
Median	1/2	1/2	-1/4	0
Ward	$(n_R+n_P)/(n_R+n_P+n_Q)$	$(n_R+n_Q)/(n_P+n_P+n_Q)$	$n_R/(n_R+n_P+n_Q)$	0
Felxibeta	$(1-\beta)/2$	$(1-\beta)/2$	β	0

That’s all for today

#HappyLearning