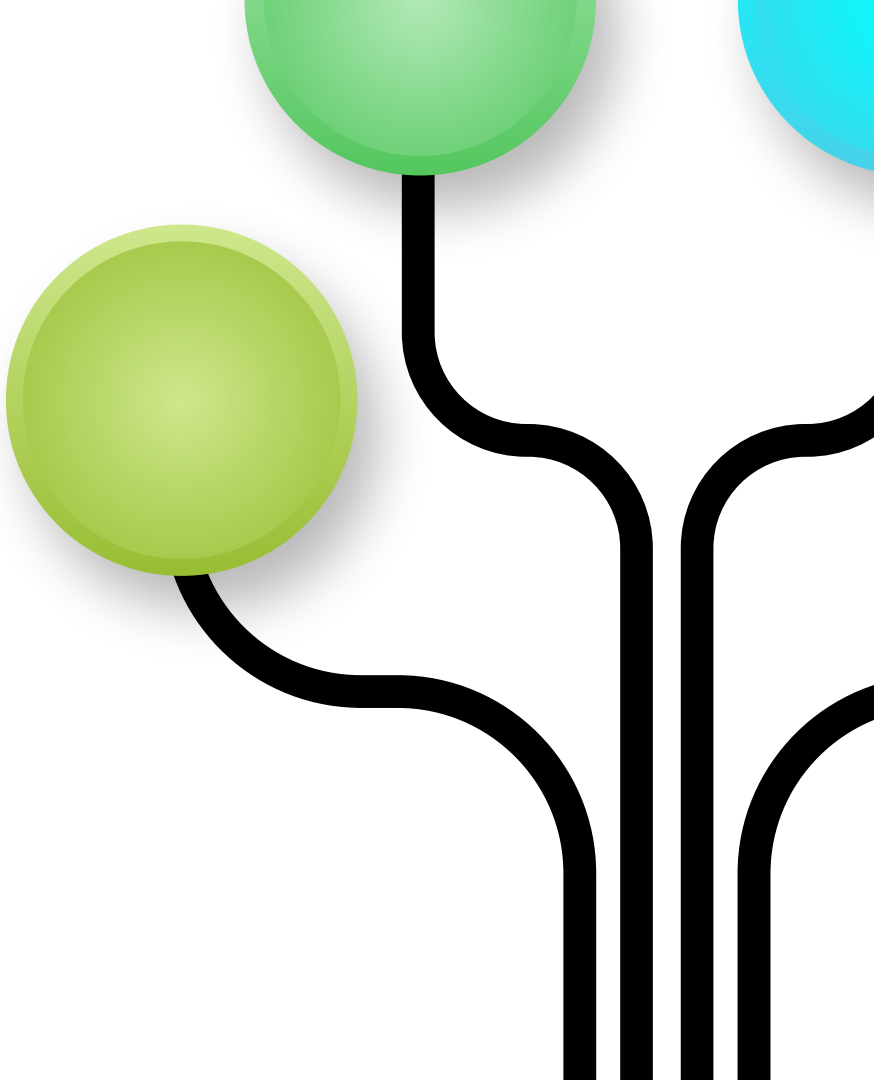
Hierarchical Clustering

By Pritam Channawar
ADS Spring2023



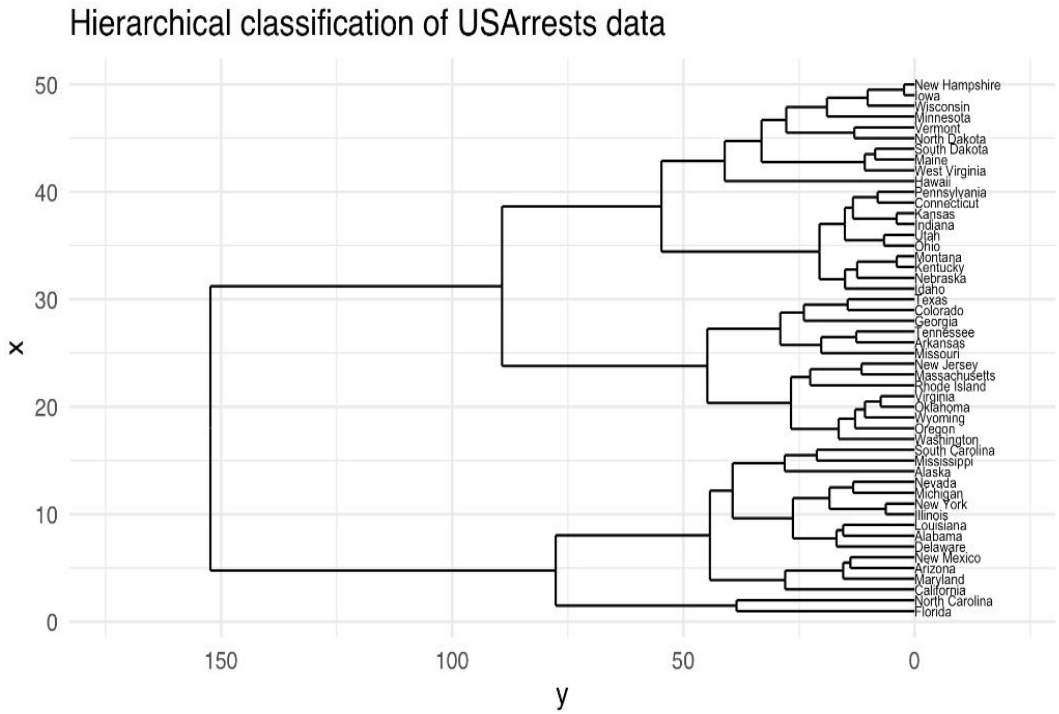
> **Hierarchical clustering is a method of cluster analysis which seeks to build a hierarchy of clusters**

Hierchical clustering and dendrograms

The result of hierarchical clustering is a *tree* where *leafs* are labelled by sample points and internal nodes correspond to merging operations

The tree conveys more information: if the tree is properly decorated, it is possible to reconstruct the different merging steps and to know which rule was applied when some merging operation was performed

The tree is called a *dendrogram*



Questions

- How to build the dendrogram?
- How to choose the cut?

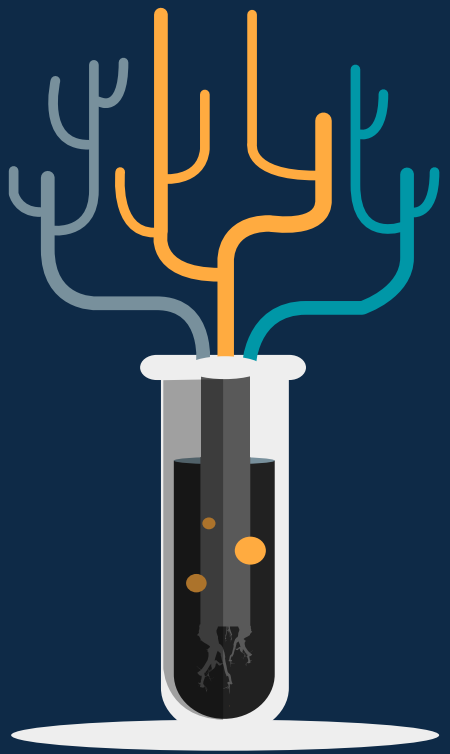
Bird-eye view at hierarchical agglomerative clustering methods

All hierarchical agglomerative clustering methods (HACMs) can be described by the following general algorithm.

- At each stage distances between clusters are recomputed by the Lance-Williams dissimilarity update formula according to the particular clustering method being used.

- Identify the 2 closest points and combine them into a cluster (treating existing clusters as points too)

- If more than one cluster remains, return to step 1.



Lance-Williams update formula

- Suppose that clusters Ci and Cj were next to be merged At this point, all of the current pairwise cluster distances are known
- The recursive update formula gives the updated cluster distances following the pending merge of clusters Ci and Cj
- Let dij,dik, and djk be shortands for the pairwise distances between clusters Ci,Cj and Ck d(ij) k be the short and for the distance between the new cluster Ci ∪ Cj and Ck(k∉{i,j})

Lance-Williams update formula

An algorithm belongs to the *Lance-Williams family* if the updated cluster *distance* $d_{(ij)k}$ can be computed recursively by

$$d_{(ij)k} = \alpha_i d_{ik} + \alpha_j d_{jk} + \beta d_{ij} + \gamma |d_{ik} - d_{jk}|$$

where $\alpha_i, \alpha_j, \beta$, and γ are parameters, which may depend on cluster sizes, that together with the cluster *distance* function d_{ij} determine the clustering algorithm.

Lance-Williams update formula

Method	α_i $(\alpha_{i'})$	β	γ
Single Linkage	$\frac{1}{2}$	0	$-\frac{1}{2}$
Complete Linkage	$\frac{1}{2}$	0	$+\frac{1}{2}$
Unweighted Average	$\frac{\lvert X_i\rvert}{\lvert X_i\rvert+\lvert X_{i'}\rvert}$	0	0
Weighted Average	$\frac{1}{2}$	0	0
Unweighted Centroid	$\frac{\lvert X_i\rvert}{\lvert X_i\rvert+\lvert X_{i'}\rvert}$	$-\frac{\lvert X_i\rvert\lvert X_{i'}\rvert}{(\lvert X_i\rvert+\lvert X_{i'}\rvert)^2}$	0
Weighted Centroid	$\frac{1}{2}$	$-\frac{1}{4}$	0
Ward	$\frac{\lvert X_i\rvert+\lvert X_j\rvert}{\lvert X_i\rvert+\lvert X_{i'}\rvert+\lvert X_j\rvert}$	$-\frac{\lvert X_j\rvert}{\lvert X_i\rvert+\lvert X_{i'}\rvert+\lvert X_j\rvert}$	0

Packages

- Numpy
- Scipy
- Sklearn

Distance of Measure

```
def euclidean_distance(u, v):
    """Return the euclidean distance between two vectors."""
    diff = u - v
    return sqrt(dot(diff, diff))

def manhattan_distance(u, v):
    """Return the Manhattan/City Block distance between two vectors."""
    return abs(u-v).sum()
```

Linkage Function

```
def UPGMA_link(clusters, i, j, dendrogram):
    n_i, n_j = len(dendrogram[i]), len(dendrogram[j])
    a_i = n_i / (n_i + n_j)
    a_j = n_j / (n_i + n_j)
    update_fn = lambda d_ik,d_jk: a_i*d_ik + a_j*d_jk
    return _general_link(clusters, i, j, update_fn)
```

$$d_{(ij)k} = \alpha_i d_{ik} + \alpha_j d_{jk} + \beta d_{ij} + \gamma |d_{ik} - d_{jk}|$$

Hierarchical Agglomerative Clustering using group average linkage also known as UPGMA(nweighted pair group method with arithmetic mean). Cluster j is clustered with cluster i when the pairwise average of values between the clusters is the smallest in the vector space. Lance-Williams parameters: M{S{alpha}(i) = |i|/(|i|+|j|); S{beta} = 0; S{gamma} = 0}

Method	α_i $(\alpha_{i'})$	β	γ
Single Linkage	$\frac{1}{2}$	0	$-\frac{1}{2}$
Complete Linkage	$\frac{1}{2}$	0	$+\frac{1}{2}$
Unweighted Average	$\frac{\lvert X_i\rvert}{\lvert X_i\rvert+\lvert X_{i'}\rvert}$	0	0
Weighted Average	$\frac{1}{2}$	0	0
Unweighted Centroid	$\frac{\lvert X_i\rvert}{\lvert X_i\rvert+\lvert X_{i'}\rvert}$	$-\frac{\lvert X_i\rvert\lvert X_{i'}\rvert}{(\lvert X_i\rvert+\lvert X_{i'}\rvert)^2}$	0
Weighted Centroid	$\frac{1}{2}$	$-\frac{1}{4}$	0
Ward	$\frac{\lvert X_i\rvert+\lvert X_j\rvert}{\lvert X_i\rvert+\lvert X_{i'}\rvert+\lvert X_j\rvert}$	$-\frac{\lvert X_j\rvert}{\lvert X_i\rvert+\lvert X_{i'}\rvert+\lvert X_j\rvert}$	0

Dataset

[[4,4], [8,4], [15,8], [24,4], [24,12]]

Linkage types Implemented

Implemented linkage clustering algorithms with L1 and L2 distances.

Type of clustering algorithms:

- Single
- Complete
- WPGMA(weighted pair group method with arithmetic mean)
- UPGMA(unweighted pair group method with arithmetic mean)
- Ward
- PDLM(prototype distance linkage - mean)

Single Linkage with with L2

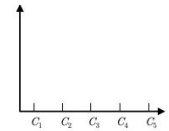
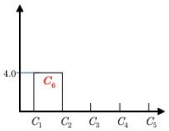
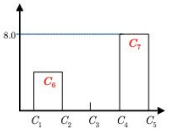
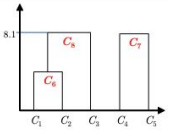
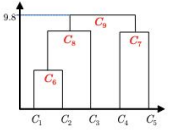
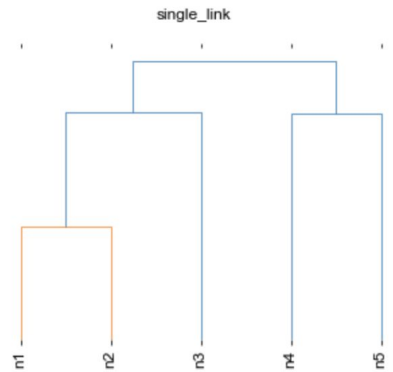
step	Clusters	Distance Matrix	Dendrogram
1	$C_1 = \{x_1\}$ $C_2 = \{x_2\}$ $C_3 = \{x_3\}$ $C_4 = \{x_4\}$ $C_5 = \{x_5\}$	$\begin{array}{c} \\ C_1 \\ C_2 \\ C_3 \\ C_4 \end{array}\begin{array}{c} \begin{array}{cccc} C_2 & C_3 & C_4 & C_5 \end{array} \\ \begin{pmatrix} 4.0 & 11.7 & 20.0 & 21.5 \\ & 8.1 & 16.0 & 17.9 \\ & & 9.8 & 9.8 \\ & & & 8.0 \end{pmatrix} \end{array}$	
2	$C_3 = \{x_3\}$ $C_4 = \{x_4\}$ $C_5 = \{x_5\}$ $C_6 = C_1 \cup C_2$ $= \{x_1, x_2\}$	$\begin{array}{c} \\ C_3 \\ C_4 \\ C_5 \end{array}\begin{array}{c} \begin{array}{ccc} C_4 & C_5 & C_6 \end{array} \\ \begin{pmatrix} 9.8 & 9.8 & 8.1 \\ & 8.0 & 16.0 \\ & & 17.9 \end{pmatrix} \end{array}$	
3	$C_3 = \{x_3\}$ $C_6 = C_1 \cup C_2$ $= \{x_1, x_2\}$ $C_7 = C_4 \cup C_5$ $= \{x_4, x_5\}$	$\begin{array}{c} \\ C_3 \\ C_6 \end{array}\begin{array}{c} \begin{array}{cc} C_6 & C_7 \end{array} \\ \begin{pmatrix} 8.1 & 9.8 \\ & 16 \end{pmatrix} \end{array}$	
4	$C_7 = C_4 \cup C_5$ $= \{x_4, x_5\}$ $C_8 = C_3 \cup C_6$ $= \{x_1, x_2,$ $x_3\}$	$\begin{array}{c} C_8 \\ C_7\,(9.8) \end{array}$	
5	$C_9 = C_7 \cup C_8$ $= \{x_1, x_2,$ $x_3, x_4, x_5\}$	The End	

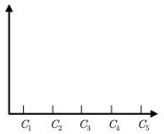
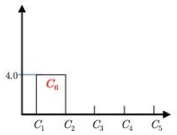
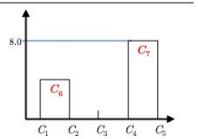
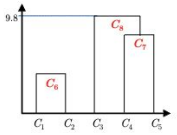
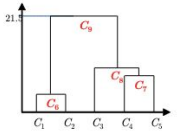
Figure 23.3: Agglomerative single linkage algorithm illustration

```
k=5
[[        inf  4.         11.70469991 20.         21.54065923]
 [ 4.                inf  8.06225775 16.         17.88854382]
 [11.70469991  8.06225775         inf  9.8488578   9.8488578 ]
 [20.         16.          9.8488578          inf  8.        ]
 [21.54065923 17.88854382  9.8488578   8.                 inf]]
k=4
[[        inf  8.06225775 16.         17.88854382]
 [ 8.06225775         inf  9.8488578   9.8488578 ]
 [16.          9.8488578          inf  8.        ]
 [17.88854382  9.8488578   8.                 inf]]
k=3
[[        inf  8.06225775 16.        ]
 [ 8.06225775         inf  9.8488578 ]
 [16.          9.8488578          inf]]
k=2
[[      inf 9.8488578]
 [9.8488578       inf]]
```

```
: labels = ['n'+str(i+1) for i in range(len(vectors))]
clusterer.dendrogram.draw(title=clusterer.linkage.__name__, labels=labels)
```

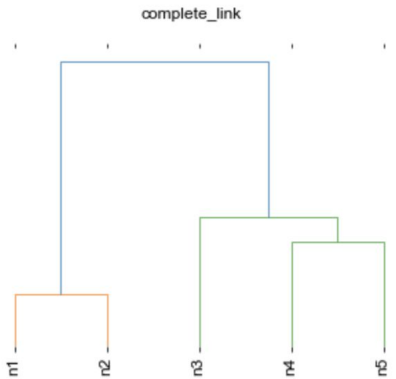


Complete Linkage with L2

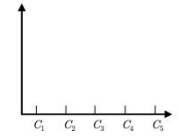
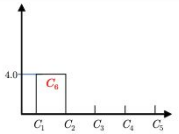
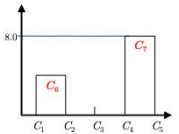
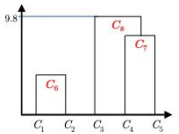
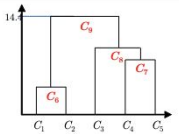
step	Clusters	Distance Matrix	Dendrogram
1	$C_1 = \{x_1\}$ $C_2 = \{x_2\}$ $C_3 = \{x_3\}$ $C_4 = \{x_4\}$ $C_5 = \{x_5\}$	$\begin{array}{c} \\ C_1 \\ C_2 \\ C_3 \\ C_4 \end{array}\begin{array}{c} \begin{array}{cccc} C_2 & C_3 & C_4 & C_5 \end{array} \\ \begin{pmatrix} 4.0 & 11.7 & 20.0 & 21.5 \\ & 8.1 & 16.0 & 17.9 \\ & & 9.8 & 9.8 \\ & & & 8.0 \end{pmatrix} \end{array}$	
2	$C_3 = \{x_3\}$ $C_4 = \{x_4\}$ $C_5 = \{x_5\}$ $C_6 = C_1 \cup C_2$ $= \{x_1, x_2\}$	$\begin{array}{c} \\ C_3 \\ C_4 \\ C_5 \end{array}\begin{array}{c} \begin{array}{ccc} C_4 & C_5 & C_6 \end{array} \\ \begin{pmatrix} 9.8 & 9.8 & 11.7 \\ & 8.0 & 20.0 \\ & & 21.5 \end{pmatrix} \end{array}$	
3	$C_3 = \{x_3\}$ $C_6 = C_1 \cup C_2$ $= \{x_1, x_2\}$ $C_7 = C_4 \cup C_5$ $= \{x_4, x_5\}$	$\begin{array}{c} \\ C_3 \\ C_6 \end{array}\begin{array}{c} \begin{array}{cc} C_6 & C_7 \end{array} \\ \begin{pmatrix} 11.7 & 9.8 \\ & 21.5 \end{pmatrix} \end{array}$	
4	$C_7 = C_4 \cup C_5$ $= \{x_4, x_5\}$ $C_8 = C_3 \cup C_6$ $= \{x_1, x_2,$ $x_3\}$	$\begin{array}{c} C_8 \\ C_7\,(21.5) \end{array}$	
5	$C_9 = C_7 \cup C_8$ $= \{x_1, x_2,$ $x_3, x_4, x_5\}$	The End	

```
k=5
[[        inf  4.         11.70469991 20.         21.54065923]
 [ 4.         inf  8.06225775 16.         17.88854382]
 [11.70469991  8.06225775        inf  9.8488578   9.8488578 ]
 [20.         16.          9.8488578         inf  8.        ]
 [21.54065923 17.88854382  9.8488578   8.                inf]]
k=4
[[        inf 11.70469991 20.         21.54065923]
 [11.70469991         inf  9.8488578   9.8488578 ]
 [20.          9.8488578         inf  8.        ]
 [21.54065923  9.8488578   8.                inf]]
k=3
[[        inf 11.70469991 21.54065923]
 [11.70469991         inf  9.8488578 ]
 [21.54065923  9.8488578         inf]]
k=2
[[        inf 21.54065923]
 [21.54065923         inf]]
```

```
labels = ['n'+str(i+1) for i in range(len(vectors))]
clusterer.dendrogram.draw(title=clusterer.linkage.__name__, labels=labels)
```

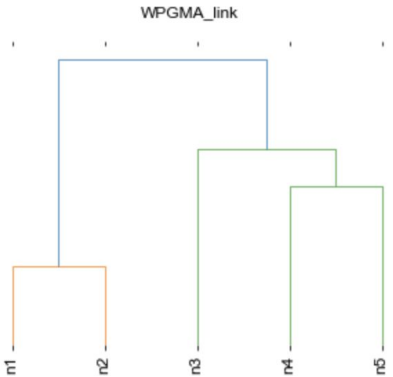


WPGMA Linkage with L2

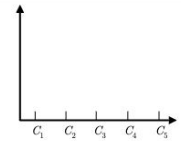
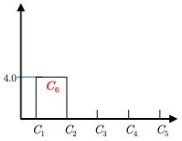
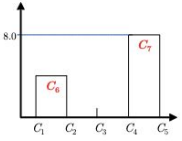
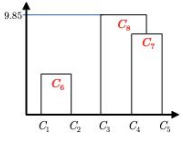
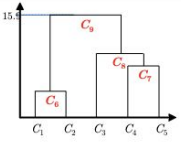
step	Clusters	Distance Matrix	Dendrogram
1	$C_1 = \{x_1\}$ $C_2 = \{x_2\}$ $C_3 = \{x_3\}$ $C_4 = \{x_4\}$ $C_5 = \{x_5\}$	$\begin{array}{c} \\ C_1 \\ C_2 \\ C_3 \\ C_4 \end{array}\begin{array}{c} \begin{array}{cccc} C_2 & C_3 & C_4 & C_5 \end{array} \\ \begin{pmatrix} 4.0 & 11.7 & 20.0 & 21.5 \\ & 8.1 & 16.0 & 17.9 \\ & & 9.8 & 9.8 \\ & & & 8.0 \end{pmatrix} \end{array}$	
2	$C_3 = \{x_3\}$ $C_4 = \{x_4\}$ $C_5 = \{x_5\}$ $C_6 = C_1 \cup C_2$ $= \{x_1, x_2\}$	$\begin{array}{c} \\ C_3 \\ C_4 \\ C_5 \end{array}\begin{array}{c} \begin{array}{ccc} C_4 & C_5 & C_6 \end{array} \\ \begin{pmatrix} 9.8 & 9.8 & 9.9 \\ & 8.0 & 18.0 \\ & & 19.7 \end{pmatrix} \end{array}$	
3	$C_3 = \{x_3\}$ $C_6 = C_1 \cup C_2$ $= \{x_1, x_2\}$ $C_7 = C_4 \cup C_5$ $= \{x_4, x_5\}$	$\begin{array}{c} \\ C_3 \\ C_6 \end{array}\begin{array}{c} \begin{array}{cc} C_6 & C_7 \end{array} \\ \begin{pmatrix} 9.9 & 9.8 \\ & 18.85 \end{pmatrix} \end{array}$	
4	$C_6 = C_1 \cup C_2$ $= \{x_1, x_2\}$ $C_8 = C_3 \cup C_7$ $= \{x_3, x_4,$ $x_5\}$	C_8 $C_6\,(14.375)$	
5	$C_9 = C_7 \cup C_8$ $= \{x_1, x_2,$ $x_3, x_4, x_5\}$	The End	

```
k=5
[[        inf  4.         11.70469991 20.         21.54065923]
 [ 4.         inf  8.06225775 16.         17.88854382]
 [11.70469991  8.06225775        inf  9.8488578   9.8488578 ]
 [20.         16.         9.8488578         inf  8.        ]
 [21.54065923 17.88854382  9.8488578   8.                inf]]
k=4
[[        inf  9.88347883 18.         19.71460152]
 [ 9.88347883         inf  9.8488578   9.8488578 ]
 [18.         9.8488578         inf  8.        ]
 [19.71460152  9.8488578   8.                inf]]
k=3
[[        inf  9.88347883 18.85730076]
 [ 9.88347883         inf  9.8488578 ]
 [18.85730076  9.8488578         inf]]
k=2
[[       inf 14.3703898]
 [14.3703898        inf]]
```

```
labels = ['n'+str(i+1) for i in range(len(vectors))]
clusterer.dendrogram.draw(title=clusterer.linkage.__name__, labels=labels)
```

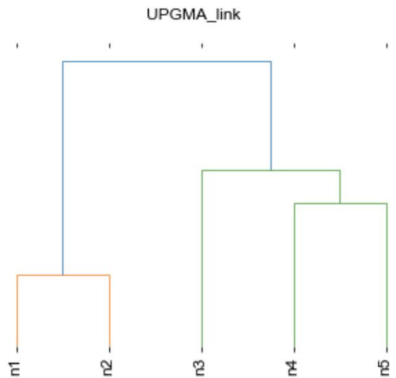


UPGMA Linkage with L2

step	Clusters	Distance Matrix	Dendrogram
1	$C_1 = \{x_1\}$ $C_2 = \{x_2\}$ $C_3 = \{x_3\}$ $C_4 = \{x_4\}$ $C_5 = \{x_5\}$	$\begin{array}{c} \\ C_1 \\ C_2 \\ C_3 \\ C_4 \end{array} \begin{array}{c} \begin{array}{cccc} C_2 & C_3 & C_4 & C_5 \end{array} \\ \begin{pmatrix} 4.0 & 11.7 & 20.0 & 21.5 \\ & 8.1 & 16.0 & 17.9 \\ & & 9.8 & 9.8 \\ & & & 8.0 \end{pmatrix} \end{array}$	
2	$C_3 = \{x_3\}$ $C_4 = \{x_4\}$ $C_5 = \{x_5\}$ $C_6 = C_1 \cup C_2$ $= \{x_1, x_2\}$	$\begin{array}{c} \\ C_3 \\ C_4 \\ C_5 \end{array} \begin{array}{c} \begin{array}{ccc} C_4 & C_5 & C_6 \end{array} \\ \begin{pmatrix} 9.8 & 9.8 & 9.88 \\ & 8.0 & 18.00 \\ & & 19.72 \end{pmatrix} \end{array}$	
3	$C_3 = \{x_3\}$ $C_6 = C_1 \cup C_2$ $= \{x_1, x_2\}$ $C_7 = C_4 \cup C_5$ $= \{x_4, x_5\}$	$\begin{array}{c} \\ C_3 \\ C_6 \end{array} \begin{array}{c} \begin{array}{cc} C_6 & C_7 \end{array} \\ \begin{pmatrix} 9.9 & 9.85 \\ & 18.86 \end{pmatrix} \end{array}$	
4	$C_6 = C_1 \cup C_2$ $= \{x_1, x_2\}$ $C_8 = C_3 \cup C_7$ $= \{x_3, x_4,$ $x_5\}$	$\begin{array}{c} C_8 \\ C_6\,(15.866) \end{array}$	
5	$C_9 = C_7 \cup C_8$ $= \{x_1, x_2,$ $x_3, x_4, x_5\}$	The End	

```
k=5
[[        inf  4.         11.70469991 20.         21.54065923]
 [ 4.                inf  8.06225775 16.         17.88854382]
 [11.70469991  8.06225775         inf  9.8488578   9.8488578 ]
 [20.         16.          9.8488578          inf  8.        ]
 [21.54065923 17.88854382  9.8488578  8.                 inf]]
k=4
[[        inf  9.88347883 18.         19.71460152]
 [ 9.88347883         inf  9.8488578   9.8488578 ]
 [18.         9.8488578          inf  8.        ]
 [19.71460152  9.8488578  8.                 inf]]
k=3
[[        inf  9.88347883 18.85730076]
 [ 9.88347883         inf  9.8488578 ]
 [18.85730076  9.8488578          inf]]
k=2
[[        inf 15.86602678]
 [15.86602678         inf]]
```

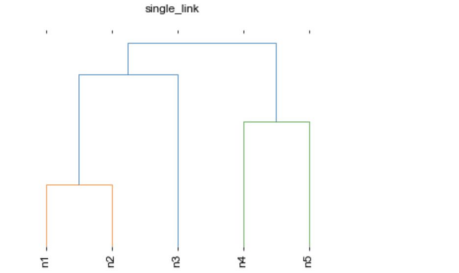
```
labels = ['n'+str(i+1) for i in range(len(vectors))]
clusterer.dendrogram.draw(title=clusterer.linkage.__name__, labels=labels)
```



Linkage with L1 distance

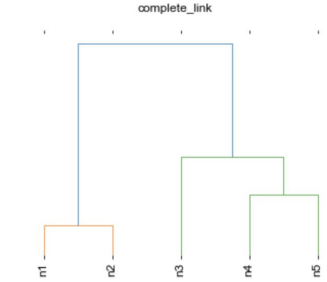
```
k=5
[[inf  4. 15. 20. 28.]
 [ 4. inf 11. 16. 24.]
 [15. 11. inf 13. 13.]
 [20. 16. 13. inf  8.]
 [28. 24. 13.  8. inf]]
k=4
[[inf 11. 16. 24.]
 [11. inf 13. 13.]
 [16. 13. inf  8.]
 [24. 13.  8. inf]]
k=3
[[inf 11. 16.]
 [11. inf 13.]
 [16. 13. inf]]
k=2
[[inf 13.]
 [13. inf]]
```

```
labels = ['n'+str(i+1) for i in range(len(vectors))]
clusterer.dendrogram.draw(title=clusterer.linkage.__name__
```



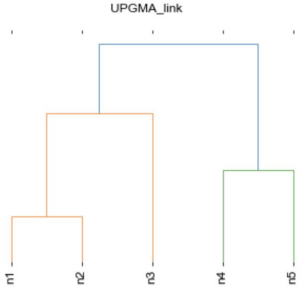

```
k=5
[[inf  4. 15. 20. 28.]
 [ 4. inf 11. 16. 24.]
 [15. 11. inf 13. 13.]
 [20. 16. 13. inf  8.]
 [28. 24. 13.  8. inf]]
k=4
[[inf 15. 20. 28.]
 [15. inf 13. 13.]
 [20. 13. inf  8.]
 [28. 13.  8. inf]]
k=3
[[inf 15. 28.]
 [15. inf 13.]
 [28. 13. inf]]
k=2
[[inf 28.]
 [28. inf]]
```

```
labels = ['n'+str(i+1) for i in range(len(vectors))]
clusterer.dendrogram.draw(title=clusterer.linkage.__name__,
```



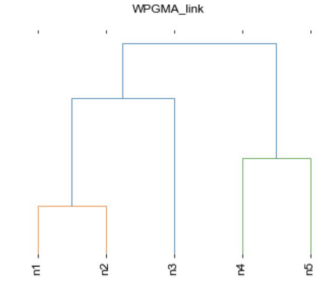

```
k=5
[[inf  4. 15. 20. 28.]
 [ 4. inf 11. 16. 24.]
 [15. 11. inf 13. 13.]
 [20. 16. 13. inf  8.]
 [28. 24. 13.  8. inf]]
k=4
[[inf 13. 18. 26.]
 [13. inf 13. 13.]
 [18. 13. inf  8.]
 [26. 13.  8. inf]]
k=3
[[inf 13. 22.]
 [13. inf 13.]
 [22. 13. inf]]
k=2
[[19.]
 [19.]]
```

```
labels = ['n'+str(i+1) for i in range(len(vectors))]
clusterer.dendrogram.draw(title=clusterer.linkage.__name__
```




```
k=5
[[inf  4. 15. 20. 28.]
 [ 4. inf 11. 16. 24.]
 [15. 11. inf 13. 13.]
 [20. 16. 13. inf  8.]
 [28. 24. 13.  8. inf]]
k=4
[[inf 13. 18. 26.]
 [13. inf 13. 13.]
 [18. 13. inf  8.]
 [26. 13.  8. inf]]
k=3
[[inf 13. 22.]
 [13. inf 13.]
 [22. 13. inf]]
k=2
[[ inf 17.5]
 [17.5  inf]]
```

```
labels = ['n'+str(i+1) for i in range(len(vectors))]
clusterer.dendrogram.draw(title=clusterer.linkage.__name__
```



Thank You!