TDT4300 — Assignment 3
CLUSTERING
GROUP 98

2 Hierarchical Agglomerative Clustering (HAC)

a) The most popular sort of hierarchical clustering is agglomerative clustering, which is used to organize items into clusters based on their similarity. The procedure begins by treating each item as if it were a singleton cluster. Following that, pairs of clusters are combined one by one until all clusters have been merged into one large cluster holding all items. The ultimate result is a dendrogram, which it a tree-based representation of the items.

Single linkage, MIN-link, yields the shortest distance between two locations a and b such that a belongs to A and b belongs to B for two clusters A and B.

The complete linkage, MAX-link, yields the longest distance between two locations a and b such that a belongs to A and b belongs to B for two clusters A and B.

b) MIN-LINK:

First step: we consider each point as a single cluster.
Second step: we merge cluster A and cluster D together and we name the resulting cluster C1.
Third step: we merge one of the points that consists cluster C1 with the closest point. For this purpose, we will calculate the distance that separate each point and choose to merge cluster C1 with the points which coordinates help obtain the lowest Euclidean value.
we remind the Euclidean formula: **d = √[(x2 – x1)2 + (y2 – y1)2]**
We find that the lowest value of d is given with the point's B coordinates with point's D coordinates such that $\mathbf{d = \sqrt{[(4 - 5)^2 + (3 - 6)^2]} = 3.16}$

Fourth step: we merge cluster C1 with the point B, and we name it cluster C2 then we will redo step 3 in order to look for the closest point to one of the points that consists cluster C2.
We find that the point C is the closest to point A such as $\mathbf{d = \sqrt{[(5 - 9)^2 + (7 - 8)^2]} = 4.12}$

Fifth step: we merge cluster C2 with the point C and name it C3 then begin to search again for the closest point in the cluster to point E.
Sixth step: we find that the point B in cluster C3 is the closest to point E such as

$$\mathbf{d = \sqrt{[(11 - 4)^2 + (3 - 3)^2]} = 2.6}$$

We finally have our last cluster.

MAX-LINK:

First step: we consider each point as a single cluster.
Second step: we merge the shortest links first, in our case cluster A and cluster D, and we name the resulting cluster C1.

Third step: we merge cluster C1 with point B given that distance between point A and point B is the longest such as

dist({A, D}, {B}) = max(dist(A, B), dist(D, B))= 4.12

We choose to merge cluster C1 with B instead of C because:
dist({A, D}, {C}) = max(dist(A, C), dist(D, C))= 4.47 > 4.12 ➔ we choose the lowest value which is 4.12.

Fourth step: we now have to choose whether to merge cluster C1{A, D, B} with E, C or merge E with C. We have the following results:

dist({A, D, B}, {C}) = max(dist(A, C), dist(D, C), dist(B, C))= 7.07
dist({A, D, B}, {E}) = max(dist(A, E), dist(D, E), dist(B, E))= 7.2
dist({E}, {C}) = max(dist(E, C))= 5.3

We will choose to merge cluster E with cluster C because the distance between them has the lowest value.

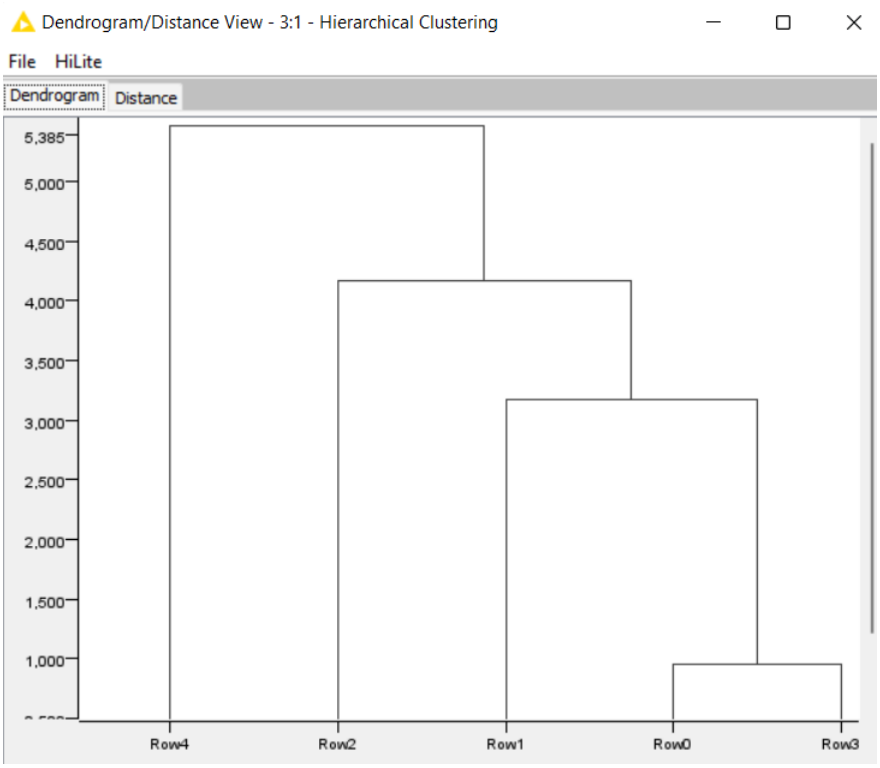
Fifth step: Now we have 2 clusters {A, D, B} and {E, C}. We will merge them together and have our final cluster.

c) The workflow:

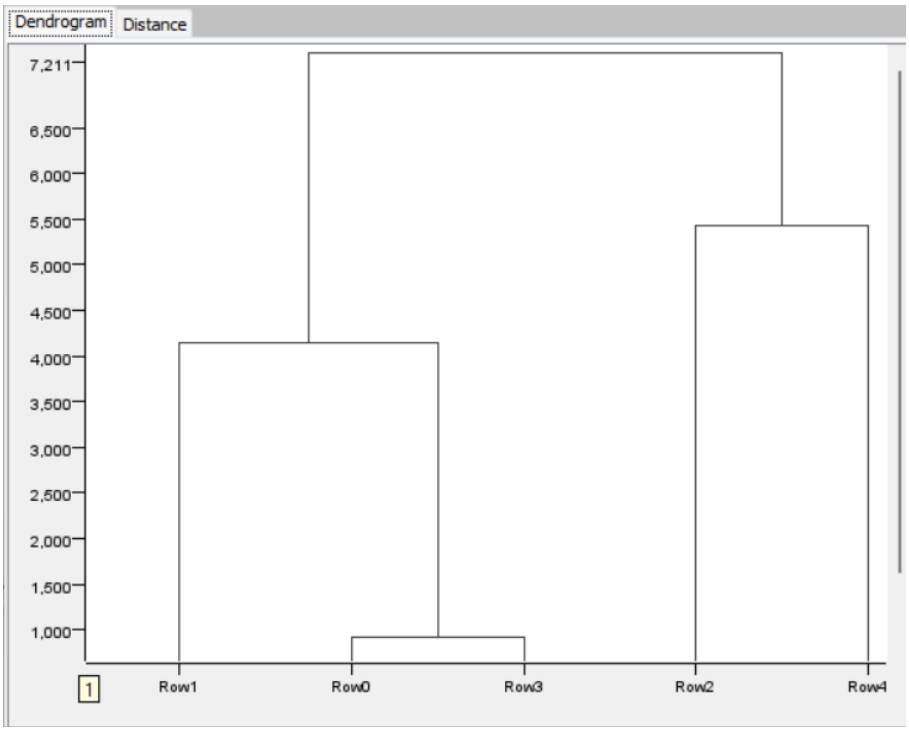
We load the **points hac.csv** into the CSV reader and then configure for the first time the Hierarchical clustering node with SINGLE then with COMPLETE linkage methods. Then we execute.

We notice that the results are identical to the ones obtained above.

Dendrogram of the SINGLE linkage:



Dendrogram for the COMPLETE linkage:



3 DBSCAN Clustering

a) We pick the first point P12 = (4,5) and we apply the Euclidean formula on its neighbour points. We find the following result:
P14 = (2,3) : d = $\sqrt{[(2-4)^2 + (3-5)^2]}$ = 2.82 < 4
P9 = (5,1) : d = $\sqrt{[(5-4)^2 + (1-5)^2]}$ = 4.12 > 4
P13 = (8,4) : d = $\sqrt{[(8-4)^2 + (4-5)^2]}$ = 4.12 > 4
P12 circle contains only 2 points (including itself) which is less than the minPoints(=3)
We conclude that **P12 is a border point.**

We conduct the same analysis for the point P13 = (8,4) and find that its circle contains only 2 points, thus **P13 is a border point**.

We pick next the point P3 = (5,1) and we apply the Euclidean formula on its neighbour points. We find the following result:
P14 = (2,3) : d = $\sqrt{[(2-5)^2 + (3-1)^2]}$ = 3.6 < 4
P8 = (8,3) : d = $\sqrt{[(8-5)^2 + (3-1)^2]}$ = 3.6 < 4
P3 circle contains 3 points (including itself) which is equal to minPoints.
We conclude that **P3 is a core point**.

P14(2,3) is also a core point because its circle contains itself and point P3 and P12.
P13 = (8,4) and P8 = (8,3) are border points because their circle only contains 2 points (including themselves).

We pick next the point P4 = (13,11) and compare it to its neighbour points. We find the following result:
P10 = (14,12) : d = $\sqrt{[(14-13)^2 + (12-11)^2]}$ = 1.4 < 4
P11 = (12,9) : d = $\sqrt{[(12-13)^2 + (9-11)^2]}$ = 2.2 < 4
P4 circle contains 3 points (including itself) which is equal to minPoints.
We conclude that **P4 is a core point**.

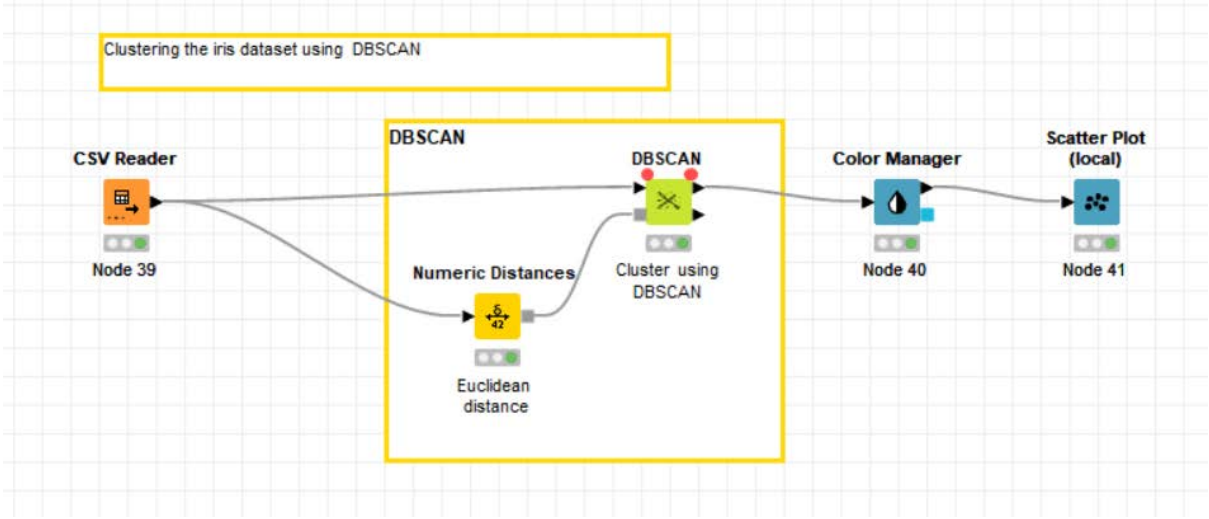
We pick next the point P11 = (12,9) and compare it to its neighbour points. We find the following result:
P5 = (12,6) : d = $\sqrt{[(12-12)^2 + (6-9)^2]}$ = 3 < 4
P10 = (14,12) : d = $\sqrt{[(14-12)^2 + (12-9)^2]}$ = 3.4 < 4
P5 circle contains 3 points (including itself) which is equal to minPoints.
We conclude that **P5 is a core point**.

Eventually, it's safe to say that **P10 = (14,12) is also a core point** because its circle contains 3 points (including itself) which is equal to minPoints.

Point **P5 = (12,6) on the other hand is a border point** because it only contains 2 points inside its circle including itself.

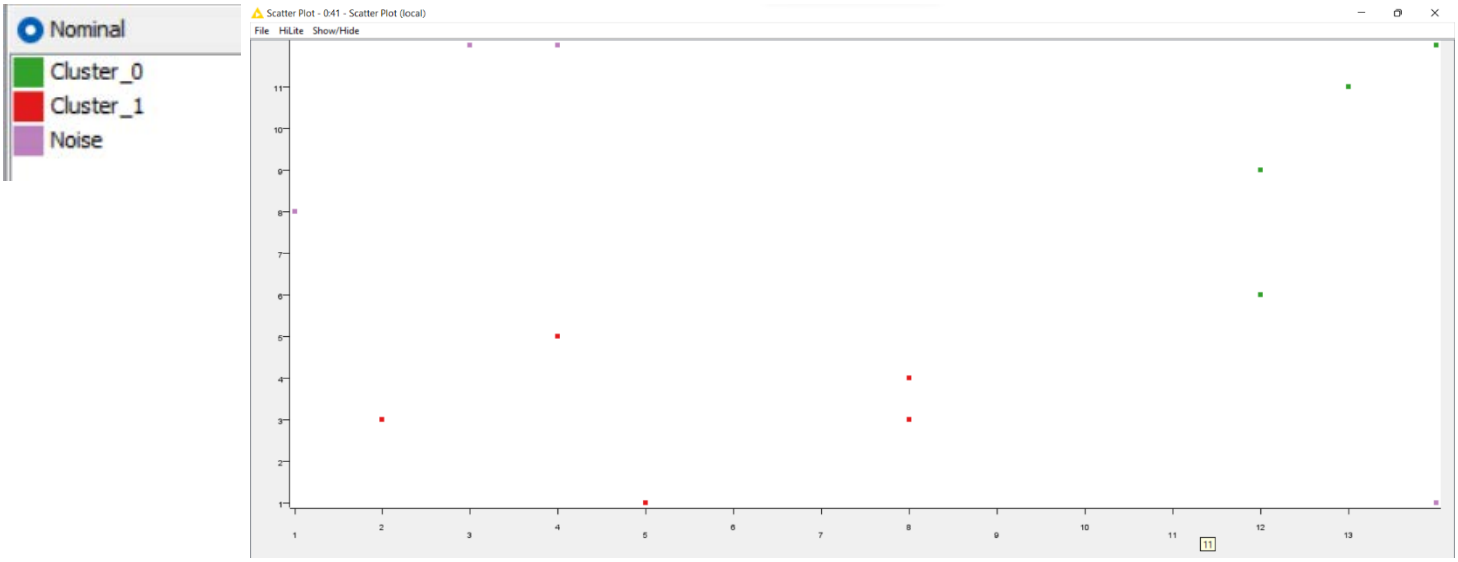
P0 = (14,1), P1 = (1,8), P2 = (3,12), P6 = (4,12) and P7 = (1,8) are considered noise points because their circle only contains themselves.

b) The workflow:



We load the CSV reader with the **points dbscan.csv file**, then we configure Eps = 4 and MinPts = 3 in the DBSCAN node after we had precised the Euclidean configuration in the Numeric Distance node.

We notice that we obtain the same results as the ones obtained in the analyses above. The scatter plot:



Row ID	Count
Noise	5
Cluster_0	4
Cluster_1	6

Table "default" - Rows: 15 | Spec - Columns: 3 | Properties | Flow Variables

Row ID	x	y	Cluster
0	14	1	Noise
1	1	8	Noise
2	3	12	Noise
3	5	1	Cluster_1
4	13	11	Cluster_0
5	12	6	Cluster_0
6	4	12	Noise
7	1	8	Noise
8	8	3	Cluster_1
9	5	1	Cluster_1
10	14	12	Cluster_0
11	12	9	Cluster_0
12	4	5	Cluster_1
13	8	4	Cluster_1
14	2	3	Cluster_1