Machine Learning

Assignment 10.3

Submitted By: Ranji Raj

March 7, 2021

Hierarchical agglomerative clustering

points	A	B	C	D	E	F
x	8	9	14	2.5	6	7
y	9	8	9	6	4	5

Table 1: Data

	A	B	C	D	E	F
A	0.000000	1.414214	6.000000	6.264982	5.385165	4.123106
B	1.414214	0.000000	5.099020	6.800735	5.000000	3.605551
C	6.000000	5.099020	0.000000	11.884864	9.433981	8.062258
D	6.264982	6.800735	11.884864	0.000000	4.031129	4.609772
E	5.385165	5.000000	9.433981	4.031129	0.000000	1.414214
F	4.123106	3.605551	8.062258	4.609772	1.414214	0.000000

Table 2: Euclidean distance matrix

a) By using Single (MIN/SLINK) linkage

- From distance matrix we observe (A,B) and (E,F) has the minimum distance (1.414) so group them.

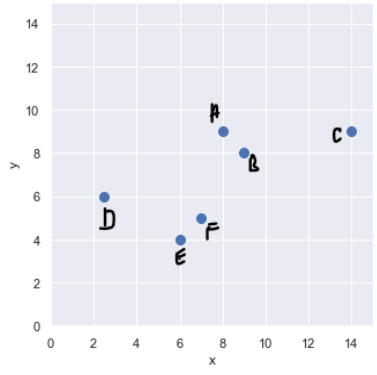

Figure 1: Data in 2D

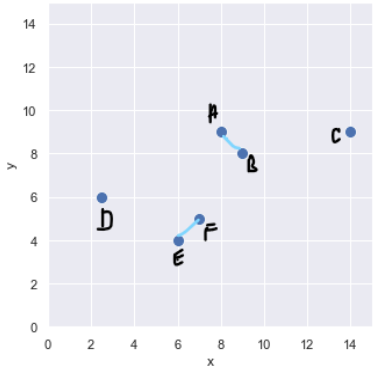

Figure 2: Initial groups

Calculate,

- $d(C,(A+B)) = min(d(C,A), d(C,B)) = min(6, 5.099) = 5.099$
- Likewise, $d(D,(A+B)) = 6.26$, $d(C,(E+F)) = 8.06$, $d(D,(E+F)) = 4.03$, $d((A+B),(E+F)) = \mathbf{3.6} \Rightarrow Merge$.

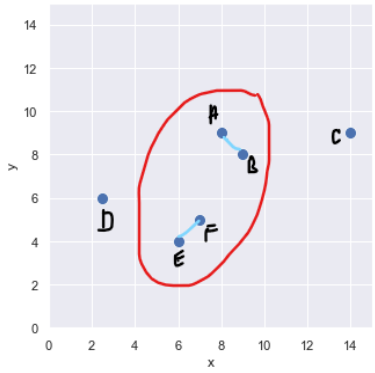

Figure 3: Merge-1

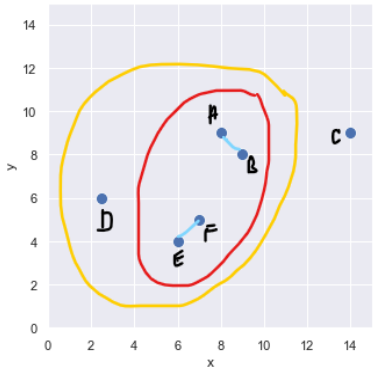

Figure 4: Merge-2

- We then calculate the distances of (A+B+E+F) to D and C.
- We observe $d((A+B+E+F), C) = 5.099$, $d((A+B+E+F), D) = \mathbf{4.03} \Rightarrow Merge$
- Finally merge (A+B+E+F+D) with C.

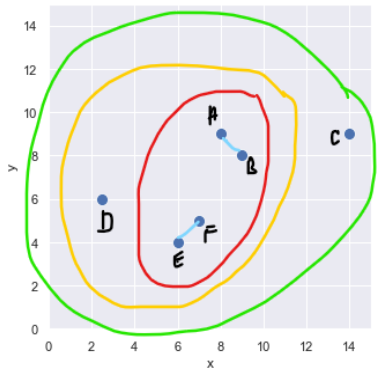

Figure 5: Merge-3

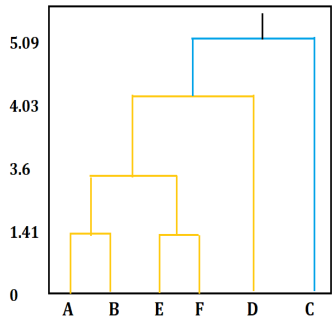

Figure 6: Dendogram-MIN link

b) By using Centroid (UPGMC) linkage

We calculate,

Centroid of A+B, $C^{AB} = (8.5, 8.5)$, Centroid of E+F, $C^{EF} = (6.5, 4.5)$
$d(C^{AB}, C) = 5.52$, $d(C^{AB}, D) = 6.5$, $d(C^{AB}, C^{EF}) = 4.47$, $d(C^{EF}, C) = 8.74$, $d(D, C) = 11.88$, $d(C^{EF}, D) = \mathbf{4.27} \Rightarrow Merge$

Centroid of D, E+F, $C^{DEF} = (\frac{6+7+2.5}{3}, \frac{6+4+5}{3}) = (5.16, 5)$

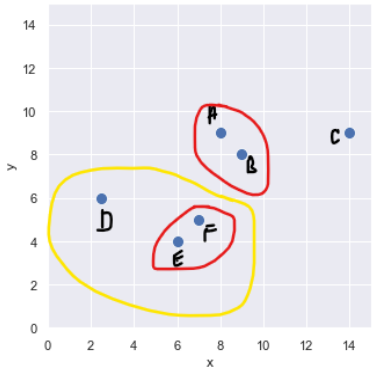

Figure 7: Merge-1

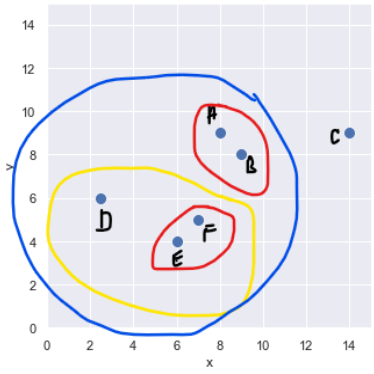

Figure 8: Merge-2

- Next calculate, $d(C^{AB}, C^{DEF}) = \mathbf{4.83} \Rightarrow Merge$, $d(C^{DEF}, C) = 9.7$
- Centroid of $C^{ABDEF} = (\frac{8+9+2.5+6+7}{5}, \frac{9+8+6+4+5}{5}) = (6.5, 6.4)$
- $d(C, C^{ABDEF}) = 7.93$, $C^{ABCDEF} = (\frac{8+9+14+2.5+6+7}{6}, \frac{9+8+9+6+4+5}{6}) = (7.75, 6.83)$

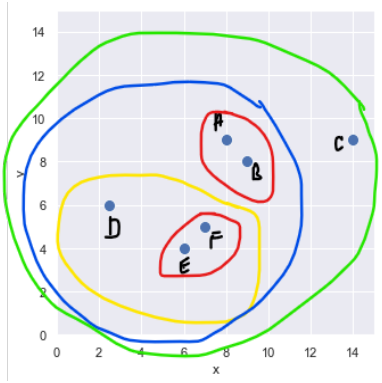

Figure 9: Merge-3

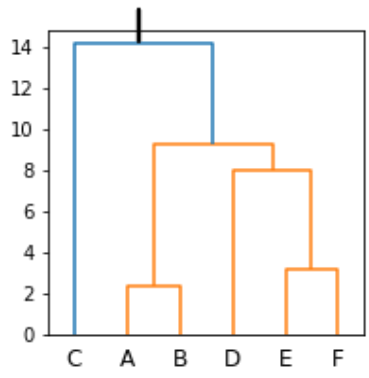

Figure 10: Dendogram-Centroid link

c) By using Complete (MAX/CLINK) linkage

We calculate,

- $d(C,(A+B)) = max(d(C,A), d(C,B)) = max(6, 5.099) = 6$
- $d(D,(A+B)) = max(d(D,A), d(D,B)) = max(6.26, 6.8) = 6.8$
- $d(C,(E+F)) = max(d(C,E), d(C,F)) = max(9.43, 8.06) = 9.43$
- $d(C,D) = 11.88$
- $d((A+B),(E+F)) = max(d(A,E), d(F,F), d(B,E), d(B,F)) = 5.38$
- $d(D,(E+F)) = max(d(D,E), d(D,F)) = max(4.03, 4.6) = \mathbf{4.6} \Rightarrow Merge$
- $d(C,(A+B)) = max(d(A,C), d(B,C)) = \mathbf{6} \Rightarrow Merge$

Finally, merge (C+A+B) and (D+E+F)

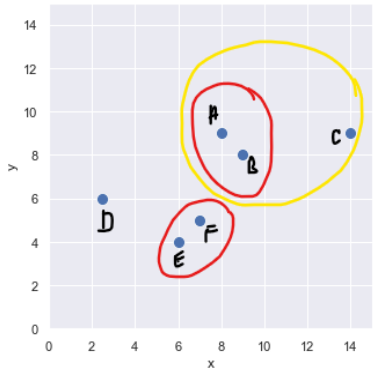

Figure 11: Merge-1

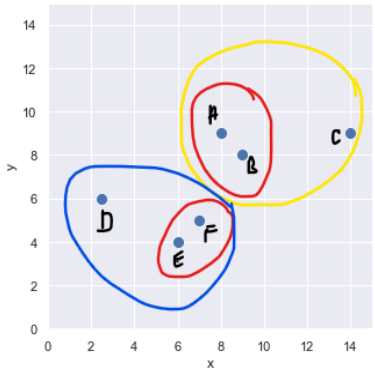

Figure 12: Merge-2

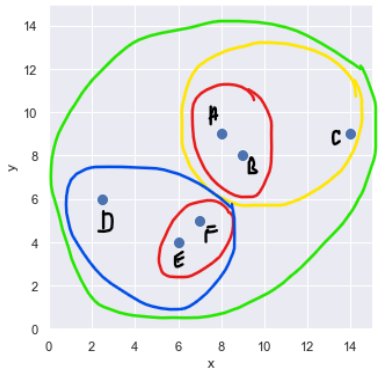

Figure 13: Merge-3

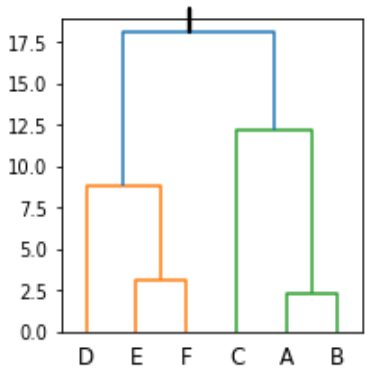

Figure 14: Dendogram-MAX link

Observations

- As an example, if customer segments are more or less as spherical shaped (globular shaped) then MIN linkage is not good as it is easily affected by outliers and tends to produce long, "loose" clusters. They control **nearest-neighbor** similarity. SLINK clustering has a tendency to produce a *chaining* of objects: a pair forms, then an object rejoins the pair, and another, and so on.
- MAX link controls **farthest-neighbor** similarity. Cluster built is of *circle*-shaped. Such clusters are "compact" contours by their borders, but they are not necessarily compact inside. *(generally preferred)*
- Centroid link (UPGMC) can be used in cases where groups of objects representing different situations (and thus likely to form different groups) are represented by *unequal* numbers of objects.