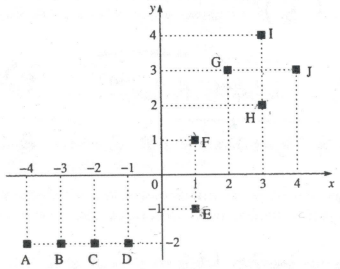
2. (Clusterizare ierarhică aglomerativă: exemplificare pe date din $\mathbb{R}^2$, folosind tipurile de similaritate "single-linkage" şi "average-linkage")

Considerăm setul de date din figura alăturată.

a. Realizaţi clusterizarea ierarhică a datelor în maniera bottom-up, folosind similaritate de tip "single-linkage" şi distanţa euclidiană între puncte.

b. Realizaţi clusterizarea ierarhică a datelor, folosind de această dată similaritate de tip "average-linkage".



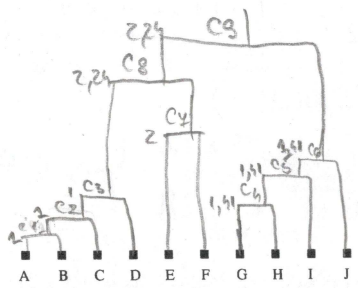
Pentru uşurinţa calculelor, în tabelul alăturat sunt listate distanţele dintre perechile de instanţe date mai sus.

	A	B	C	D	E	F	G	H	I	J
A	0.00	1.00	2.00	3.00	5.10	5.83	7.81	8.06	9.22	9.43
B	1.00	0.00	1.00	2.00	4.12	5.00	7.07	7.21	8.49	8.60
C	2.00	1.00	0.00	1.00	3.16	4.24	6.40	6.40	7.81	7.81
D	3.00	2.00	1.00	0.00	2.24	3.61	5.83	5.66	7.21	7.07
E	5.10	4.12	3.16	2.24	0.00	2.00	4.12	3.61	5.39	5.00
F	5.83	5.00	4.24	3.61	2.00	0.00	2.24	2.24	3.61	3.61
G	7.81	7.07	6.40	5.83	4.12	2.24	0.00	1.41	1.41	2.00
H	8.06	7.21	6.40	5.66	3.61	2.24	1.41	0.00	2.00	1.41
I	9.22	8.49	7.81	7.21	5.39	3.61	1.41	2.00	0.00	1.41
J	9.43	8.60	7.81	7.07	5.00	3.61	2.00	1.41	1.41	0.00

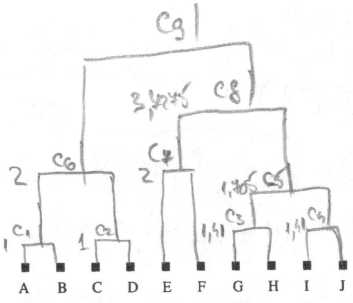
Observaţie (1): Dacă la o iteraţie a algoritmului de clusterizare „distanţele" (adică similarităţile) dintre două perechi de clustere au aceeaşi valoare, prioritatea la alcătuirea noului cluster este dictată de ordinea alfabetică.

Observaţie (2): La construcţia dendrogramelor, înălţimea [nodului rădăcină] pentru fiecare cluster C_i va fi dată de distanţa (adică măsura de similaritate) dintre clusterele C_j şi C_k din care a fost construit clusterul C_i.

Veţi construi cele două dendrograme pornind de la reprezentările de mai jos.



Dendrograma *single-linkage*



Dendrograma *average-linkage*

5.

$d(AB, CD) = \frac{2+3+1+2}{4} = 2$

$d(F,G,H) = \frac{2,24+2,24}{2} = 2,24$

$d(G,H,I) = \frac{1,41+2+1,41+2}{4} = \frac{6,82}{4} = 1,705$

$d(AB,CD,EF) = \frac{5,10+4,12+3,16+2,24+5,83+5+4,24+3,61}{8} =$

$= 4,1625$

$d(EF,GH,I) = \frac{4,12+3,61+5,39+5+2,24+2,24+3,61+3,61}{8} =$

$= 3,7275$

$d(ABCD, EFGF) = \ldots = 6,3908$

6.

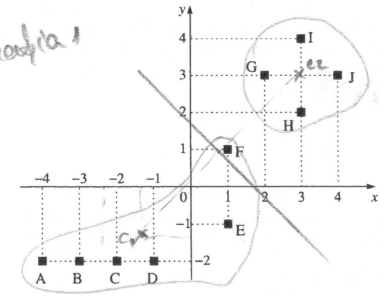
c. La curs am afirmat că una dintre metodele de iniţializare pentru algoritmul K-means este următoarea: clusterele iniţiale sunt cele K clustere de la vârful dendrogramei (engl., top clustere) obţinute cu ajutorul unei metode de clusterizare ierarhică.

Executaţi algoritmul K-means pe datele din enunţ, pentru fiecare dintre cele două variante de iniţializare, folosind rezultatele de la punctele precedente. (Veţi lucra cu $K = 2$.) Folosiţi pentru aceasta reprezentările de mai jos. Indicaţi la fiecare iteraţie coordonatele centroizilor şi componenţa fiecărui cluster.

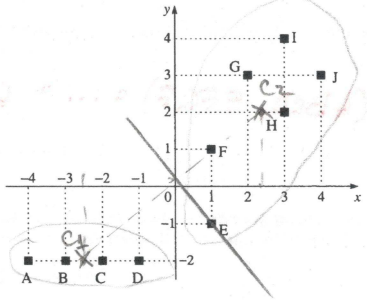
Setul de date din acest exerciţiu este oarecumva particular. Judecând pe un caz [mai] general, care dintre cele două tipuri de funcţii de similaritate folosite la punctele a şi b vi se pare mai adecvată pentru pasul de iniţializare a algoritmului K-means.?

d. Considerând cele două *clustere* de "top" obţinute de către algoritmul de clusterizare ierarhică care foloseşte similaritate de tip "single-linkage" ca fiind nişte *clase*, trasaţi pe desenul din stânga de mai jos, *separatorul decizional* produs de algoritmul 1-NN.

c) Iteraţia 1

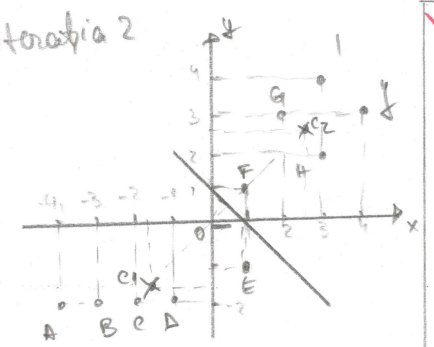


K-means cu clusterele iniţiale cf. clusterizării ierarhice *single-linkage*

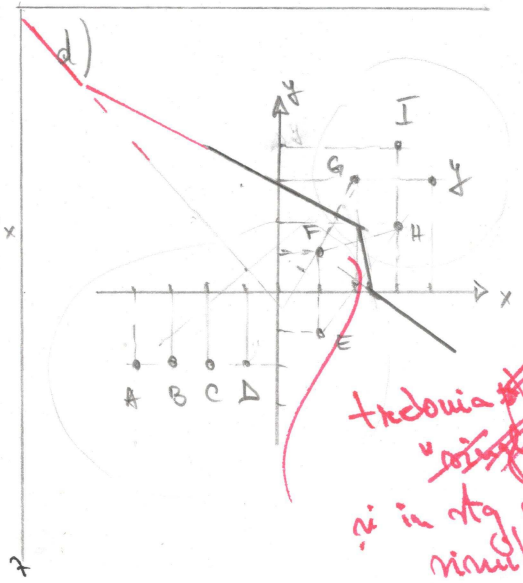


K-means cu clusterele iniţiale cf. clusterizării ierarhice *average-linkage*

Iteraţia 2



d)



trebuia ~~o linie~~ și în stg mus, nimic!

c) Cazul single linkage!

$x_{c_1} = \frac{-4-3-2-1+1}{6} \approx -1(3)$

$y_{c_1} = \frac{-2\cdot 4 - 1 + 1}{6} \approx -1(3)$

$c_1(-1,3; -1,3)$

$x_{c_2} = \frac{2+2\cdot 3+4}{4} = 3$

$y_{c_2} = \frac{2+2\cdot 3+4}{4} = 3$

$c_2(3,3)$

$d(F,c_1) = \sqrt{(-1,3+1)^2+(-1,3+1)^2} = 3,25$

$d(F,c_2) = \sqrt{(3-1)^2+(3-1)^2} = \sqrt{8} = 2,82$ → F ∈ Cluster 2

Iteraţia 2:

$x_{c_1} = \frac{-4-3-2-1+1}{5} = -1,8$

$y_{c_1} = \frac{-2\cdot 4 - 1}{5} = -1,8$

$c_1(-1,8; -1,8)$

$x_{c_2} = \frac{1+2+2\cdot 3+4}{5} = 2,6$

$y_{c_2} = \frac{1+2+2\cdot 3+4}{5} = 2,6$

$c_2(2,6; 2,6)$

Clusterele nu se mai schimbă, centroizii rămân la fel, algoritmul s-a stabilizat:

$c_1: \{A, B, C, D, E\}$, $c_2: \{F, G, H, I, J\}$

Cazul Average-linkage

$x_{c_1} = \frac{-4-3-2-1}{4} = -2,5$

$y_{c_1} = -2$

$c_1(-2,5; -2)$

$x_{c_2} = \frac{2\cdot 1+2+2\cdot 3+4}{6} = 2(3)$

$y_{c_2} = \frac{-1+1+2+2\cdot 3+4}{6} = 2$

$c_2(2,(3); 2)$

$d(E,c_1) = \sqrt{(1+2,5)^2+(-1+2)^2} = \sqrt{12,25+1} = 3,64$

$d(E,c_2) = \sqrt{(1-2,(3))^2+(-1-2)^2} = \sqrt{1,68+9} = 3,28$ → E ∈ Cluster

⇒ Clusterele şi centroizii rămân neschimbate, algoritmul s-a stabilizat

$c_1: \{A, B, C, D\}$, $c_2: \{E, F, G, H, I, J\}$