Învăţare automată

— Licenţă, anul III, 2018-2019, re-examinare, parţial II —

Nume student: **Grupa:**

1. **(Clusterizare ierarhică aglomerativă: aplicare pe date din $\mathbb{R}^2$, folosind măsurile de similaritate single-linkage, complete-linkage şi metrica lui Ward)**

Pe setul de date din $\mathbb{R}^2$

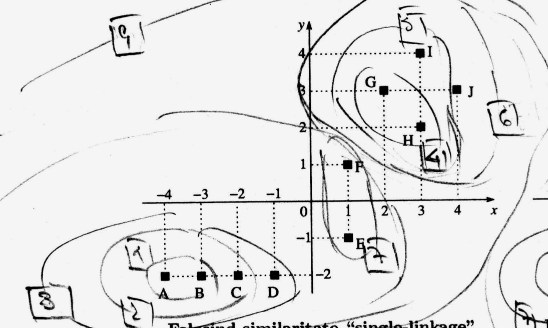
$$A:(-4,-2), B:(-3,-2), C:(-2,-2), D:(-1,-2), E:(+1,-1)$$
$$F:(+1,+1), G:(+2,+3), H:(+3,+2), I:(+3,+4), J:(+4,+3)$$

veţi aplica algoritmul de clusterizare ierarhică aglomerativă conform specificaţiilor de la fiecare din punctele următoare.

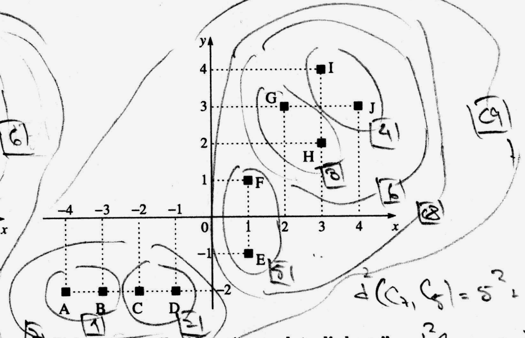
Precizare: **Dacă la o iteraţie a algoritmului de clusterizare distanţele (adică similarităţile) dintre două perechi de clustere au aceeaşi valoare, prioritatea la alcătuirea noului cluster este dictată de ordinea alfabetică.**

a. Folosind măsurile de similaritate single-linkage şi complete-linkage, reprezentaţi pe desenele de mai jos rezultatele aplicării algoritmului de clusterizare ierarhică aglomerativă, sub forma unor *dendrograme* (adică, ierarhii) *aplatizate* (engl., flat hierarchies).

Pentru fiecare cluster non-singleton veţi folosi câte o curbă închisă de formă elipsoidală pentru a încadra punctele din clusterul respectiv. Fiecărei elipse îi veţi asocia câte un index, scris sub forma $\boxed{1}$, $\boxed{2}$, $\boxed{3}$ ... (chiar pe conturul elipsei respective) pentru a indica ordinea în ~~care sunt~~ formate ~~clusterele~~.



Folosind similaritate "single-linkage"



Folosind similaritate "complete-linkage"

$d^2(C_7, C_6) = 5^2 + 3^2 = 34$

$d^2(C_5, C_6) = \max\{2^2+5^2 = 29,\ 3^2+4^2 = 25\}$

b. La curs am prezentat încă o funcţie de similaritate, numită *metrica lui Ward.* Potrivit acestei metrici, distanţa dintre două clustere disjuncte X şi Y se defineşte astfel:

$$\Delta(X,Y) = \sum_{x_i \in X \cup Y} \|x_i - \mu_{X\cup Y}\|^2 - \sum_{x_i \in X} \|x_i - \mu_X\|^2 - \sum_{x_i \in Y} \|x_i - \mu_Y\|^2 \qquad (1)$$

unde, spre exemplu, μ_X este centroidul [sau „centrul de greutate“ al] clusterului X, iar x_i este o instanţă generică dintr-un cluster [oarecare, fixat]. Prin definiţie, aici vom considera $\mu_X = \frac{1}{n_X}\sum_{x_i\in X} x_i$, unde n_X este numărul de elemente din X. (Similar sunt definiţi centroizii μ_Y şi $\mu_{X\cup Y}$.)

Se poate arăta că

$$\Delta(X,Y) \;=\; \frac{n_X n_Y}{n_X + n_Y}\|\mu_X - \mu_Y\|^2. \qquad (2)$$

Observaţii:

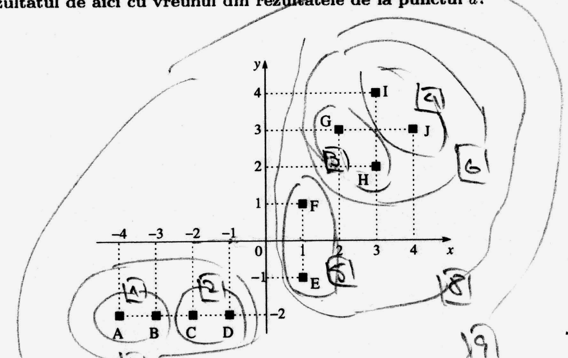
1. Pentru perechi de clustere (X,Y) şi (X',Y') astfel încât $n_Y = n_{X'}$ şi $n_Y = n_{Y'}$, formula (2) arată că la clusterizare ierarhică este „favorizată“ acea pereche pentru care centroizii (μ_X şi μ_Y, respectiv $\mu_{X'}$ şi $\mu_{Y'}$) sunt mai apropiaţi.

2. Invers, dacă $\|\mu_X - \mu_Y\| = \|\mu_{X'} - \mu_{Y'}\|$, atunci este favorizată perechea pentru care ponderea[1] (adică $\frac{n_X n_Y}{n_X + n_Y}$, respectiv $\frac{n_{X'} n_{Y'}}{n_{X'} + n_{Y'}}$) este mai mică.

Aceste două *observaţii* vă vor ajuta să simplificaţi / reduceţi foarte mult calculele pe care ar trebui să le faceţi la punctul *b*!

Aplicaţi algoritmul de clusterizare ierarhică aglomerativă pe acelaşi set de date ca mai sus, însă folosind de această dată metrica lui Ward. Ca şi la punctul *a*, veţi folosi elipse (şi indici) pentru a reprezenta noua *dendrogramă aplatizată*.

Coincide rezultatul de aici cu vreunul din rezultatele de la punctul *a*?



Folosind metrica lui Ward

identic cu complete-linkage

Observaţie:

La punctul *b*, la fiecare iteraţie a algoritmului *veţi justifica riguros* alegerea făcută scriind [doar] *calculele care sunt determinante*!

<u>Răspuns:</u> (pentru punctul *b*)

Iteraţia 1:

... $\min_{X,Y\in\{A,\dots,J\},\ X\neq Y}$ $\Delta(\{X\},\{Y\}) = \min_{X,Y\in\{A,\dots,J\},\ X\neq Y} \|X-Y\|^2$

...

$C_1 = \dots$ $\{A,B\}$

Pt. că

Iteraţia 2:

... $\Delta(\{A,B\},\{C\}) = \frac{2}{3}\cdot 1{,}5^2 = \frac{2}{3}\cdot\left(\frac{3}{2}\right)^2 = \frac{3}{2} > \frac{1}{2} = \Delta(\{C\},\{D\})$

$\Delta(\{A,B\},\{X\}) > \Delta(\{A,B\},\{C\})\ \forall X \in \{D,\dots,J\}$

[1] Această pondere este jumătate din *media armonică* a cardinalilor n_X şi n_Y.