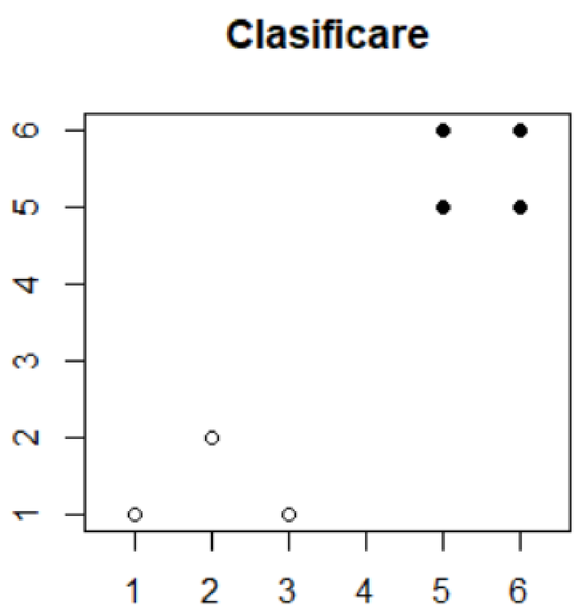
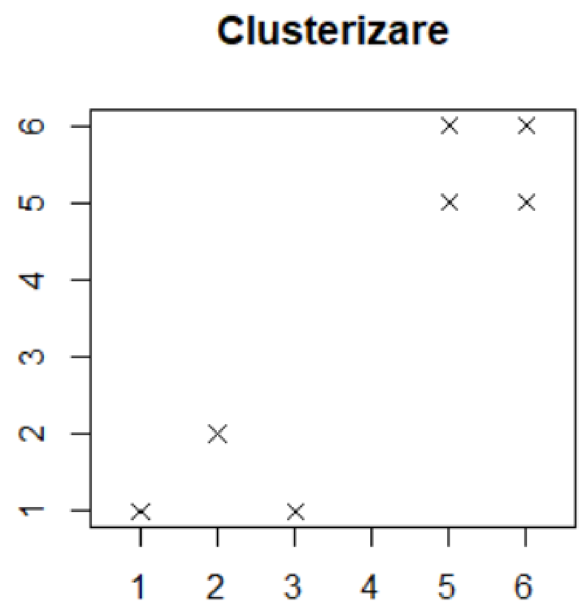
# Notițe Seminar 9

November 30, 2019

**Intro:** Trecem la un nou mare capitol care va fi studiat până la sfârșitul semestrului: **Clusterizare**. Dacă până acum am făcut învățare supervizată de tip clasificare, de acum facem învățare **nesupervizată** de tip **clusterizare**. Mai concret, dacă până acum (la clasificare) un set de date arăta astfel:




la clusterizare va arăta astfel:




Diferența este că nu mai avem instanțe etichetate (pentru că de acum am zis că facem învățare **ne**supervizată). Totuși ce putem face cu aceste date? Dacă vă uitați atent la ultimul desen, deși nu avem etichete, observăm că instanțele s-ar grupa în două grupuri (sau clustere): unul în stânga jos și unul în dreapta sus. De fapt, asta vom face: vom încerca să grupăm datele (= să le clusterizăm).

[După ce le-am grupat, dacă vrem, celor din stânga jos le putem pune o etichetă, iar celor din dreapta sus, o altă etichetă. Astfel, am etichetat setul de date și de acum putem rula, dacă vrem, algoritmi de clasificare.]

Dar hai să vedem cum se poate face clusterizarea asta...

# 1 Remember

1. **produsul scalar**:

$$\begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} \cdot \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = x_1y_1 + \cdots + x_ny_n$$

   Exemplu:

$$\begin{bmatrix} 1 \\ 2 \end{bmatrix} \cdot \begin{bmatrix} 3 \\ 4 \end{bmatrix} = 1 \cdot 3 + 2 \cdot 4 = 11$$

2. **norma** $p$, $p \geq 1$:

$$\left\| \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} \right\|_p = (|x_1|^p + \cdots + |x_n|^p)^{\frac{1}{p}}$$

Exemple:

$p = 1$

$$\left\| \begin{bmatrix} 1 \\ -2 \end{bmatrix} \right\|_1 = |1| + |-2| = 3$$

$p = 2$

$$\left\| \begin{bmatrix} 1 \\ -2 \end{bmatrix} \right\|_2 = \left\| \begin{bmatrix} 1 \\ -2 \end{bmatrix} \right\| = \sqrt{1^2 + (-2)^2} = \sqrt{5} \text{ - norma euclidiană}$$

**Implicit**, dacă nu se specifică o altă normă, semnul $\|...\|$ se va referi la norma **euclidiană**.

$p = \infty$

$$\left\| \begin{bmatrix} 1 \\ -2 \end{bmatrix} \right\|_\infty = \max\{|1|, |-2|\} = 2$$

3. distanța $p$ indusă de norma $p$ (vezi *Notițe Seminar 8*)

$$d_p(x, y) = \|x - y\|_p$$

4. $\|x\|_2^2 = x \cdot x$ (o puteți verifica imediat)

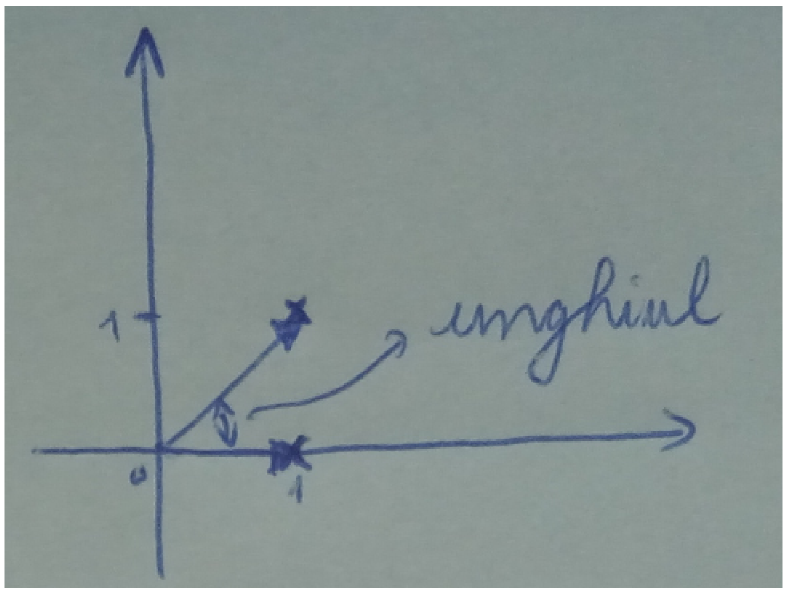
5. $x^2 \overset{\text{not.}}{=} x \cdot x$

6. **unghiul** dintre 2 vectori

$$\cos(x, y) = \frac{x \cdot y}{\|x\| \|y\|}$$

Exemplu:

$$\cos\left(\begin{bmatrix}1\\0\end{bmatrix}, \begin{bmatrix}1\\1\end{bmatrix}\right) = \frac{\begin{bmatrix}1\\0\end{bmatrix} \cdot \begin{bmatrix}1\\1\end{bmatrix}}{\left\|\begin{bmatrix}1\\0\end{bmatrix}\right\| \left\|\begin{bmatrix}1\\1\end{bmatrix}\right\|} = \frac{1}{1 \cdot \sqrt{2}} = \frac{1}{\sqrt{2}}$$




## 2 Clusterizare

- nu există etichete/coloană output
- nu există date de antrenare/de test, ci doar date (deși unii algoritmi pot adaptați pentru a include o nouă instanță într-un cluster...)
- coeziune(cluster) $\stackrel{\text{intuitiv}}{=}$ "cât de legate/unite sunt punctele în cluster"
- separare(cluster$_1$, cluster$_2$) $\stackrel{\text{intuitiv}}{=}$ "cât de bine distanțate sunt punctele din clusterul$_1$ față de punctele din clusterul$_2$"

- poate fi împărțită astfel:

  1. **ierarhică**
     - vom forma un arbore care se numește **dendrogramă**
     - în funcție de cum construim dendrograma (adică de jos în sus sau de sus în jos), clusterizarea ierarhică se împarte în:
       (a) clusterizare ***bottom up*** (sau aglomerativă) - majoritatea exercițiilor vor fi de acest tip
       (b) clusterizare ***top down*** (sau divizivă) - doar ex. 6/pag. 477 este de acest tip
  2. **neierarhică**/plată/aplatizată
     (a) cu asignare **hard** a instanțelor la cluster, adică având o instanță, spunem că ea aparține unui singur cluster și atât: algoritmul **k-means**
     (b) cu asignare **soft** a instanțelor la cluster, adică având o instanță, spunem că ea aparține tuturor clusterelor: cu probabilitatea $p_1$ aparține clusterului 1, cu probabilitatea $p_2$ aparține clusterului 2 etc.: algoritmul **EM/GMM**

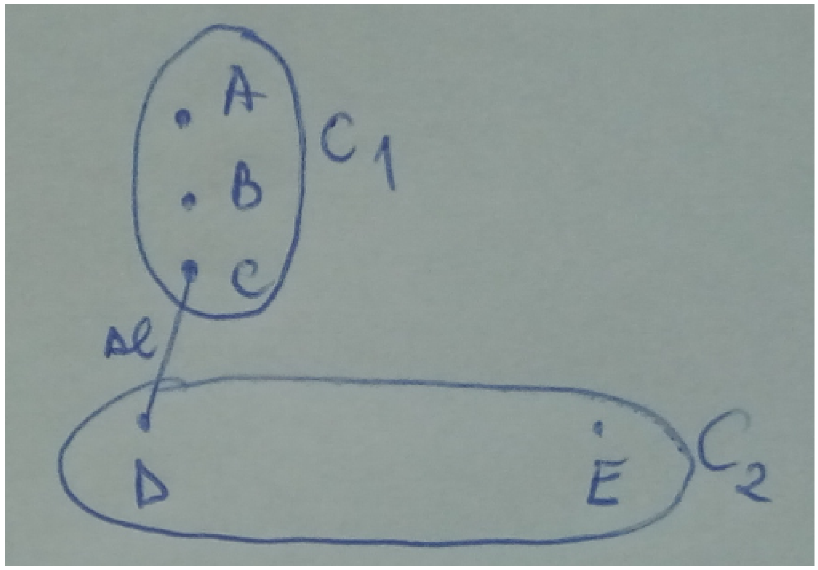
# 3 Clusterizare ierarhică

În acest context apare noțiunea de **similaritate**. Noi vom lucra, de obicei, cu similaritatea dintre doi vectori definită ca *inversul distanței* dintre acei doi vectori. Într-un exercițiu aveți și o similaritate care nu pleacă de la o distanță: este vorba de ex. 32/pag. 539 unde se vorbește despre similaritatea cosinus (care are sens dacă vă gândiți că $\cos \in [-1, 1]$, $\cos = 1$ dacă unghiul dintre vectori este de zero grade [deci, sunt similare], iar $\cos = -1$ dacă unghiul este de 180 de grade [deci, nu sunt similare]).

Dacă până acum sunteți deja obișnuiți să calculați distanțe între vectori (vezi *Notițe Seminar 8*), în contextul clusterizării ierarhice trebuie să știm să calculăm **distanțe între clustere** având setată o anumită distanță între vectori. Astfel, setând o distanță $d$ (nu neapărat cea euclidiană) între vectori, vom putea calcula distanța între vectori în mai multe moduri:

1. **single-linkage**: $d_{sl}(C_1, C_2) \stackrel{\text{def.}}{=} \min\{d(x, y) | x \in C_1, y \in C_2\}$

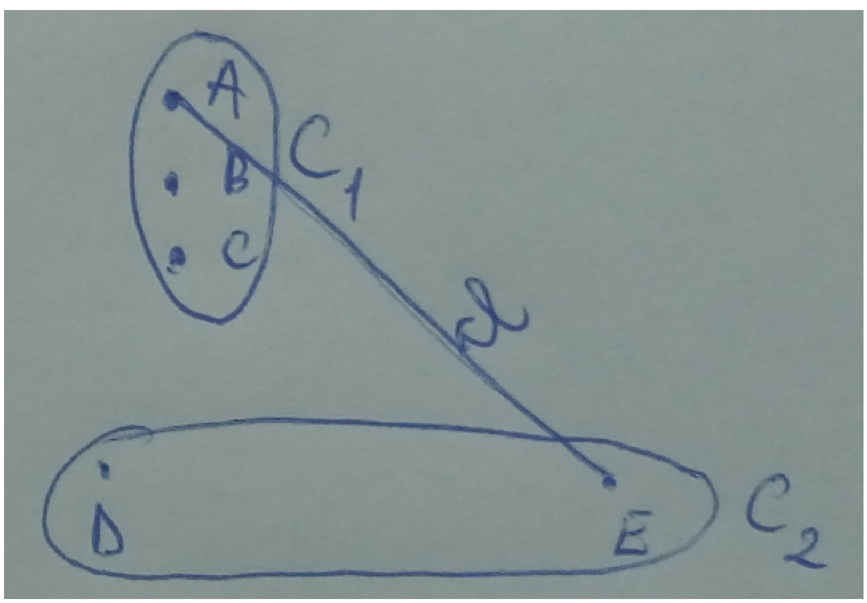
   Exemplu vizual având setată distanța euclidiană:




$d_{sl}(C_1, C_2) = d(C, D)$

2. **complete-linkage**: $d_{cl}(C_1, C_2) \stackrel{\text{def.}}{=} \max\{d(x, y) | x \in C_1, y \in C_2\}$

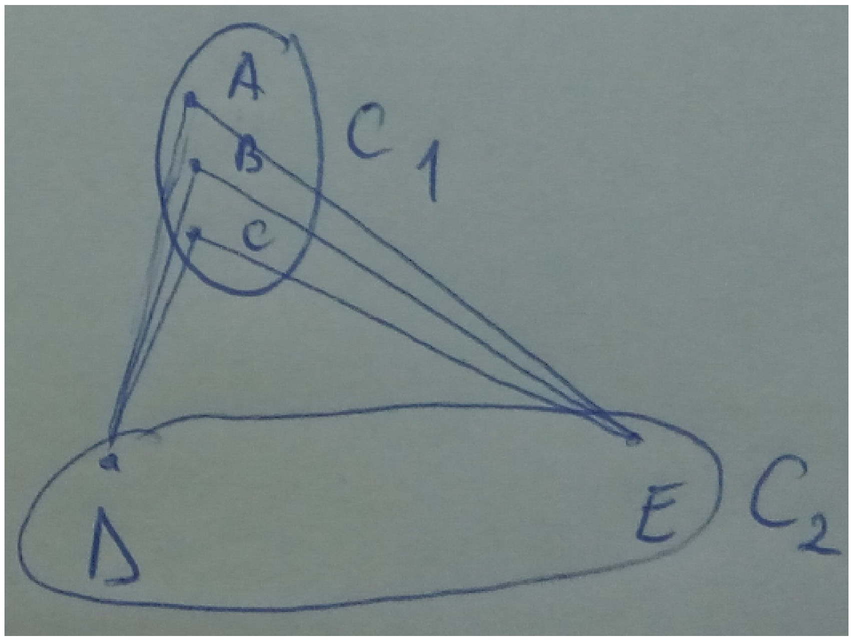
   Exemplu vizual având setată distanța euclidiană:




$d_{cl}(C_1, C_2) = d(A, E)$

3. **average-linkage**: $d_{sl}(C_1, C_2) \stackrel{\text{def.}}{=} avg\{d(x,y) | x \in C_1, y \in C_2\} = \frac{1}{|A||B|} \sum_{x \in C_1, y \in C_2} d(x,y)$

   Exemplu:




$$d_{al}(C_1, C_2) = \frac{d(A,D) + d(B,D) + d(C,D) + d(A,E) + d(B,E) + d(C,E)}{6}$$

4. **metrica lui Ward**:

   - aici vom lucra DOAR cu distanța euclidiană (deci, nu putem seta $d$ să fie altă distanță)
   - avem nevoie de noțiunea de **centroid al unui cluster**:

$$\mu_{\text{Cluster}} \stackrel{\text{def.}}{=} \frac{\sum_{x \in \text{Cluster}} x}{|\text{Cluster}|}$$

   De exemplu: Dacă Cluster $= \left\{ \begin{bmatrix} 1 \\ 2 \end{bmatrix}, \begin{bmatrix} 3 \\ 4 \end{bmatrix}, \begin{bmatrix} 5 \\ 6 \end{bmatrix} \right\}$, atunci

$$\mu_{\text{Cluster}} = \frac{\begin{bmatrix} 1 \\ 2 \end{bmatrix} + \begin{bmatrix} 3 \\ 4 \end{bmatrix} + \begin{bmatrix} 5 \\ 6 \end{bmatrix}}{3} = \begin{bmatrix} \frac{1+3+5}{3} \\ \frac{2+4+6}{3} \end{bmatrix} = \begin{bmatrix} 3 \\ 4 \end{bmatrix}$$

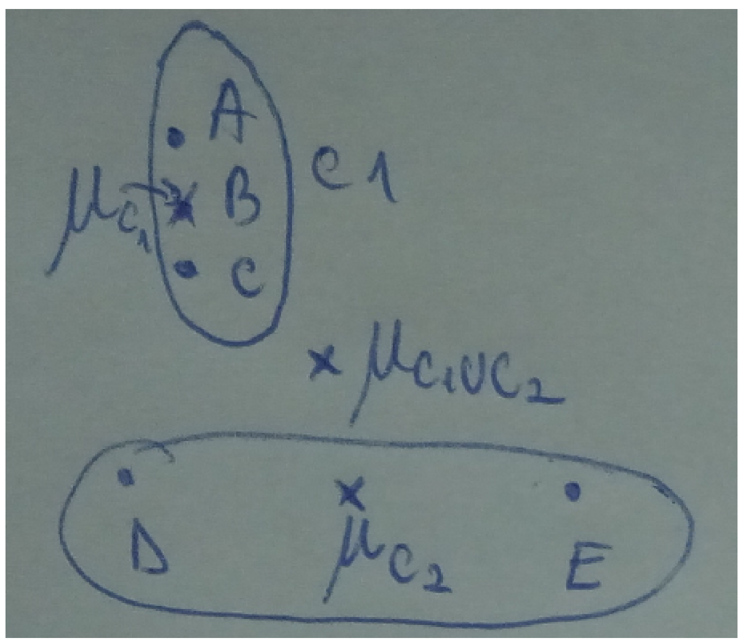
Următoarea formulă (pe care o puteți verifica imediat) vă va fi de folos în unele demonstrații:

$$\mu_{A \cup B} = \frac{|A|\mu_A + |B|\mu_B}{|A| + |B|}$$

Revenim la metrica lui Ward și o definim:

$$d_{\text{Ward}}(C_1, C_2) \stackrel{\text{def.}}{=} \sum_{x \in C_1 \cup C_2} d^2(x, \mu_{C_1 \cup C_2}) - \sum_{y \in C_1} d^2(y, \mu_{C_1}) - \sum_{z \in C_2} d^2(z, \mu_{C_2})$$
$$\stackrel{\text{dem. ex.30/538-vezi rez.din slide-uri}}{=} \frac{|C_1||C_2|}{|C_1| + |C_2|} d^2(\mu_{C_1}, \mu_{C_2})$$

Exemplu:



$$
\begin{aligned}
d_{\text{Ward}}(C_1, C_2) &= d^2(A, \mu_{C_1 \cup C_2}) + d^2(B, \mu_{C_1 \cup C_2}) + d^2(C, \mu_{C_1 \cup C_2}) + \\
&+ d^2(D, \mu_{C_1 \cup C_2}) + d^2(E, \mu_{C_1 \cup C_2}) - \\
&- (d^2(A, \mu_{C_1}) + d^2(B, \mu_{C_1}) + d^2(C, \mu_{C_1})) - \\
&- (d^2(D, \mu_{C_2}) + d^2(E, \mu_{C_2})) \\
&\overset{\text{dem.}}{=} \frac{|C_1||C_2|}{|C_1| + |C_2|} d^2(\mu_{C_1}, \mu_{C_2}) \\
&= \frac{3 \cdot 2}{3 + 2} d^2(\mu_{c_1}, \mu_{C_2}) \\
&= \frac{6}{5} d^2(\mu_{C_1}, \mu_{C_2})
\end{aligned}
$$

Acum aveți noțiunile de bază ca să aplicați **algoritmul care formează dendrograma de jos în sus (*bottom up*)**. Setăm o distanță între vectori (pentru sl, cl, al) sau nu (pentru Ward). Setăm o distanță între clustere (sl, cl, al sau Ward). În continuare vom lucra cu aceste setări.

Inițial, fiecare punct va fi într-un cluster separat (cluster *singleton*) și va fi o frunză în dendrogramă [deci, inițial avem atâtea clustere câte puncte avem].

La o iterație:

- Calculăm distanțele între oricare două clustere.
- Luăm distanța **minimă** și combinăm clusterele corespunzătoare în dendrogramă trecând astfel de la un nivel inferior în arbore la un nivel superior.
- Dacă avem mai mulți candidați pentru distanța minimă, se va specifica o convenție din care să reiasă care clustere vor fi combinate în această iterație.

Executăm iterațiile până când ajungem la rădăcină (adică ajungem să punem toate punctele într-un singur cluster).

Exemplu: vezi ex. 1a/pag. 468

**Mențiuni**:

- prin tăierea dendrogramei cu o linie orizontală, vom obține o clusterizare **plată** (vezi desen final din rezolvarea ex. 1b)

- având o dendrogramă, putem împărți datele în oricâte clustere dorim (de la 2 la numărul de instanțe); pentru a afla totuși în câte clustere ar fi bine să împărțim datele, ne putem uita la înălțimi; **înălțimile** în dendrogramă sunt importante!; în exerciții se specifică modul în care se calculează înălțimile; de obicei, ele vor fi invers proporționale cu coeziunea noului cluster SAU direct proporționale cu separarea dintre cele două clustere tocmai unite; din acest motiv vom folosi înălțimile ca să obținem **numărul natural de clustere**, adică în câte clustere trebuie să împărțim datele (vezi ex. 1b/pag. 468)

- clusterizarea ierarhică asignează ***hard*** instanțele la clustere

## Schemă de final

1. Clasificare vs. Clusterizare
2. Remember
3. Clusterizare
   (a) coeziune, separare
      i. ierarhică
         A. bottom up
         B. top down (vezi ex.6/pag.477)
      ii. neierarhică
         A. cu asignare hard a instanțelor la clustere
         B. cu asignare soft a instanțelor la clustere
4. Clusterizare ierarhică
   (a) distanță între vectori
   (b) similaritate între vectori
   (c) distanță între clustere
      - single linkage
      - complete linkage
      - average linkage
      - Ward
   (d) similaritate între clustere
      - cos
   (e) dendrogramă, înălțimi în dendrogramă
   (f) număr natural de clustere