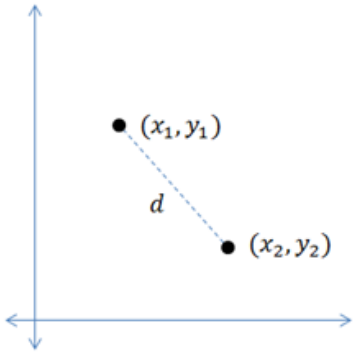
# Gruparea datelor folosind K-means

Gruparea (engl. c*lustering*) este una dintre cele mai frecvent utilizate tehnici de analiză exploratorie a datelor, ideea fiind de a deduce regulile și modalitatea de structurare a acestora. Metodele de clusterizare presupun identificarea unor configurații de grupare a datelor astfel încât grupurile formate să fie constituite din elemente cu un anumit grad de similaritate. Altfel spus, se dorește gruparea datelor în așa fel încât în cadrul aceluiași grup datele să fie silmilare într-o mare măsură, în timp ce datele din grupuri diferite să fie, la rândul lor, diferite. Gruparea presupune stabilirea unei modalități de evaluare a similarității a două elemente din setul de date ce se grupează. Măsurarea similarității se realizează prin stabilirea unei metrici corespunzătoare (o metodă de cuantificare a similarității a două elemente). Metrica se decide funcție de specificul datelor și de scopul grupării, de exemplu poate fi distanța euclidiană, o distanța bazată pe corelație, sau o formulă mai complexă care ia în calcul o multitudine de trăsături ale datelor.

Exemple de metrici de similaritate:

- distanța euclidiană dintre două puncte:




$$d = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$

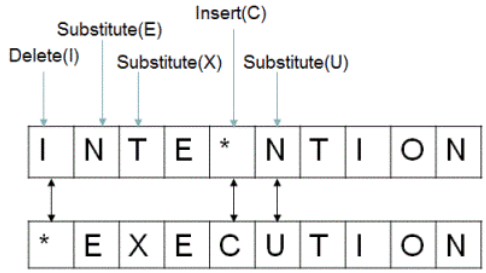
- similaritatea cosinus dintre doi vectori:

$$\cos\theta = \frac{\vec{a} \cdot \vec{b}}{\|\vec{a}\|\|\vec{b}\|}$$

$$\|\vec{a}\| = \sqrt{a_1^2 + a_2^2 + a_3^2 + \cdots + a_n^2}$$

$$\|\vec{b}\| = \sqrt{b_1^2 + b_2^2 + b_3^2 + \cdots + b_n^2}$$

- distanța Levenshtein dintre două șiruri de caractere (numărul de operații de inserare/ștergere/substituție necesare pentru a obține unul dintre șiruri pornind de la celălalt)




K-means este unul dintre cei mai simpli algoritmi de grupare. În cadrul gruparii K-means se încearcă partiționarea setului de date în K grupuri (clustere) distincte care nu se suprapun, unde fiecare element al setului de date aparține unui singur grup (există alte metode de grupare unde un element poate aparține mai multor grupuri cu anumite probabilități). Numărul de grupuri K este prestabilit. Scopul este de a crea grupuri în cadrul cărora elementele să fie cât mai similare posibil, în timp ce elementele din grupuri diferite să fie cât mai puțin similare posibil. Cu cât există mai puține variații în cadrul grupurilor, cu atât punctele de date sunt mai omogene (similare) în cadrul aceluiași grup. Cu cât variația în cadrul unui grup este mai redusă, cu atât crește omogenitatea grupului respectiv.
Astfel, punctele se atribuie unui grup astfel încât suma pătratelor distanțelor dintre elementele grupului să fie minimă. În practică însă, este ineficient calculul similarității dintre fiecare două elemente ale setului de date, de aceea problema se reformulează astfel: pentru fiecare grup se determină și se actualizează un centru (punct central) și se atribuie elementele grupului cu centrul cel mai apropiat (cu similaritate maximă).

Metoda presupune mai întâi stabilirea *a priori* (de la început) a numărului de grupuri K. Apoi, pentru fiecare element, se determină grupul cu centrul cel mai apropiat și se consideră că elementul aparține acestui grup. După ce se prelucrează toate elementele în această manieră, centrul fiecărui cluster se recalculează ca fiind *baricentrul* sistemului format din punctele identificate anterior ca aparținând clusterului. Baricentrul se determină în poziția ce minimizează distanțele până la elementele grupului. Odată recalculate pozițiile tuturor centrelor grupurilor, se determină din nou distanțele până la fiecare element și se regrupează elementele în mod corespunzător. Raționamentul se reia până când centrele grupurilor devin stabile (nu se mai modifică semnificativ). În Fig 1 se ilustrează un exemplu de grupare a unei mulțimi de puncte din plan.

Presupunem că se dorește gruparea unei mulțimi de elemente în K grupuri. Algoritmul implică următorii pași:

1. Se stabilesc cele K centre în spațiul elementelor ce trebuie grupate. Pozițiile centrelor pot fi, de exemplu, generate aleator.

2. Pentru fiecare element, se determină grupul căruia îi aparține ca fiind cel cu centrul cel mai apropiat.

3. Odată grupate toate elementele, se actualizează centrul fiecărui grup (de exemplu, în cazul unor puncte din plan se poate face media aritmetică a coordonatelor).

4. Se repetă pașii 2, 3 până când pozițiile centrelor grupurilor nu se mai modifică (în practică, se stabilește o valoare prag sub care trebuie să se situeze diferența dintre pozițiile centrelor grupurilor determinate în cadrul iterației curente și cele determinate în iterația anterioară).

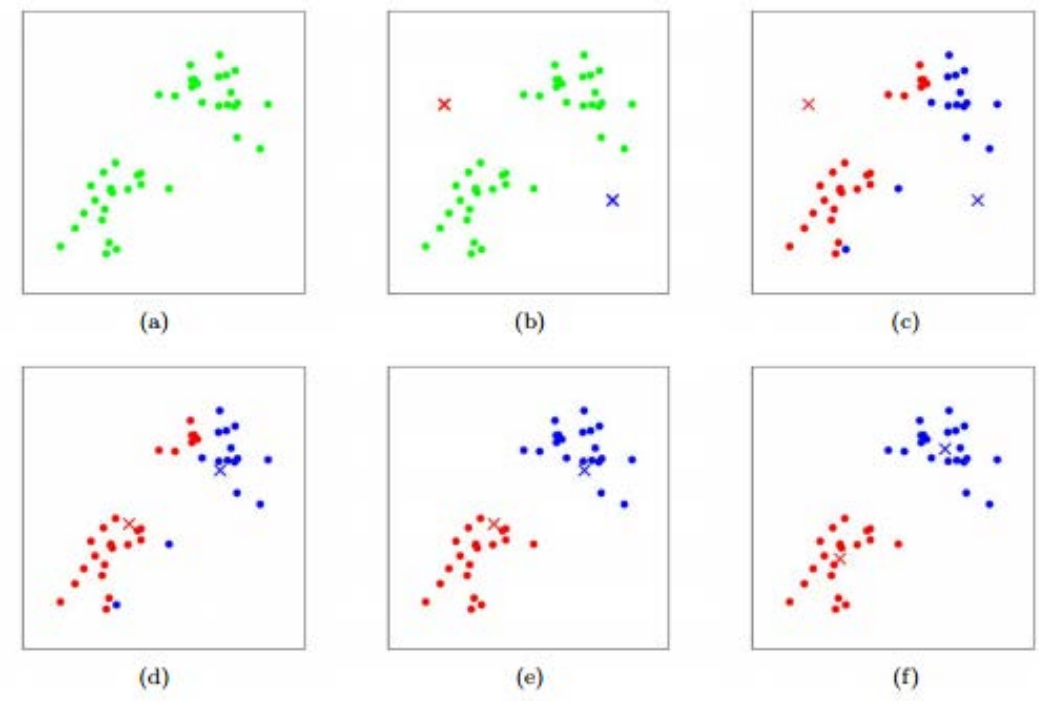
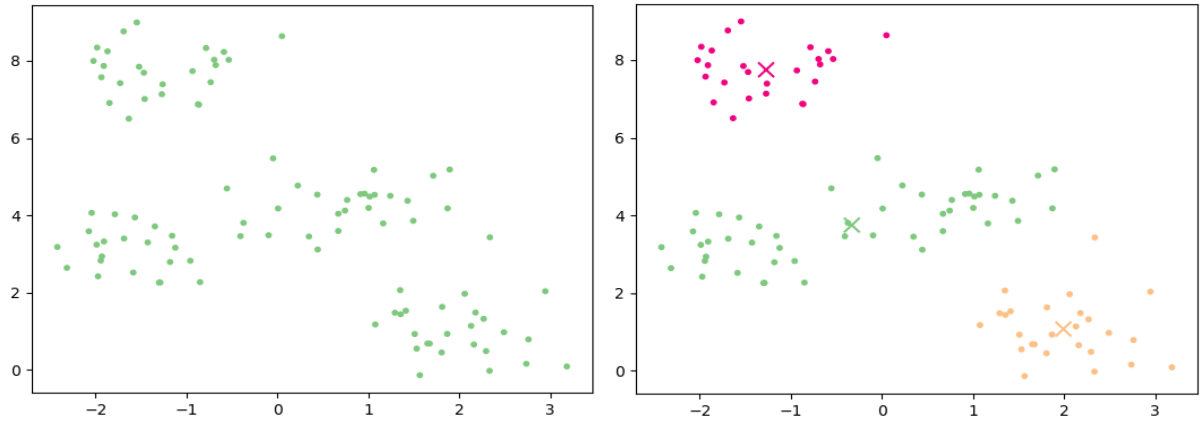



Fig. 1. Exemplu de grupare K-means: (a) Punctele în configurația inițială, negrupate; (b) Se stabilește K=2 și se inițializează doue centre, marcate cu X, în poziții aleatoare; (c) Se realizează o primă grupare: fiecare punct aparține grupului cu centrul cel mai apropiat; (d, e, f) în cadrul mai multor iterații se recalculează pozițiile centrelor (media aritmetică a pozițiilor punctelor) și se regrupează punctele folosind noile centre.

**Cerințe:**

1. Implementați algoritmul K-means pentru o mulțime de puncte din plan și reprezentați rezultatul grupării (Fig. 2). Se poate utiliza ca punct de plecare codul sursă care însoțește documentația laboratorului. Testați implementarea folosind punctele din fișierele knnpoints3.txt, knnpoints4.txt, knnpoints_uniform.txt



(a) (b)

Fig. 2. (a) Punctele inițiale, negrupate; (b) Punctele grupate în trei grupuri (K=3). Centrele grupurilor sunt marcate cu X

2. Determinați cea mai bună valoare a lui K (cea pentru care grupurile sunt cât mai compacte și mai bine separate) folosind metoda coeficientului siluetă.

   Coeficientul siluetă (*Silhouette Coefficient,* notat SC) se calculează astfel:
   Pentru fiecare punct i se determină:

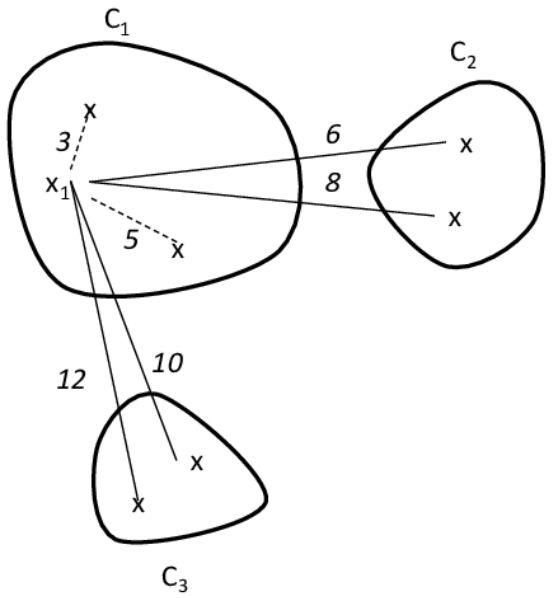
   a(i) = distanța medie dintre i și celelalte puncte din grupul din care face parte i

   b(i) = minimul distanțelor medii dintre i și punctele din celelalte grupuri

   s(i) = (b(i) – a(i)) / max(a(i), b(i))  s(i) – silueta punctului i

   În Fig. 3. se prezintă un exemplu de calcul al siluetei unui punct.

   SC = s(i) mediu (silueta medie determinată pentru toate punctele, pentru o anumită valoare a lui K)
   Cea mai bună valoare a lui K = cea pentru care SC este maxim.




$$a(x_1) = \frac{3+5}{2} = 4$$

$$b(x_1) = min\left(\frac{6+8}{2}, \frac{10+12}{2}\right) = 7$$

$$s(x_1) = \frac{7-4}{\max(7,4)} = \frac{3}{7}$$

Fig. 3. Exemplu de calcul al siluetei $s(x_1)$ a punctului $x_1$ , pentru K = 3