



Mammal



Bird



Fish

# ÎNVĂȚAREA NESUPERVIZATĂ

LIMBOI SERGIU

# AGENDA

- Învățarea automată-recapitulare
- Învățarea nesupervizată-generalități
- Gruparea datelor (Clustering)
  - Generalități
  - Concepte de bază
  - Măsuri de similaritate
  - Algoritmi
  - Evaluarea rezultatelor
- Aplicații ale clustering-ului
- Concluzii



# ÎNVĂȚAREA AUTOMATĂ

- învățarea poate fi explicată ca fiind activitatea de a obține cunoștințe sau de a le înțelege, precum și abilitatea de a-ți însuși noțiuni prin studiu, instruire sau experiență [1].
- Învățarea automată (*Machine Learning*) [1] se referă la modificări din sisteme care realizează sarcini asociate cu diverse teme, concepte din Inteligența Artificială, sarcini precum diagnoză, predicție, planificare, control sau detectare.



# ÎNVĂȚAREA AUTOMATĂ

- Scop -> proiectarea și dezvoltarea unor algoritmi și metode utilizate pentru ca un sistem computațional să „învețe”.
- De ce să învețe sistemele informatice?
  - unele sarcini se pot defini doar prin exemple; de aceea trebuie să putem furniza perechi de intrări-ieșiri, în absența unei legături concrete între datele de intrare și rezultate;
  - este posibil ca anumite informații ascunse, nedescifrate, să reflecte corelații sau legături;
  - cantitatea mare de informații poate fi dificil de codificat de către mintea umană.



# ÎNVĂȚAREA AUTOMATĂ

- Pentru a defini conceptele utilizate în problematica învățării se folosește o funcție  $f$ , iar sarcina celui care învață este de a “ghici” această funcție.
- Tipuri de învățare
  - **Supervizată**- știm (uneori doar aproximativ) valorile funcției  $f$  pentru  $m$  cazuri ale unui set de antrenare; sistemul poate intui, după etapa de antrenare, care ar fi funcția și pentru un alt set de date
  - **Nesupervizată**- avem doar o mulțime de vectori (un set de date) pentru care nu știm funcția. Presupune divizarea (partiționarea) setului de date în submulțimi sau grupuri. În acest caz valoarea funcției va fi numele subgrupului (clasei) din care vectorul de intrare face parte
  - **Prin întărire (reinforcement)**- sistemul interacționează cu mediul și poate primi recompense sau penalizări



# ÎNVĂȚAREA NESUPERVIZATĂ

## GENERALITĂȚI

- Ideea de bază- sistemul primește informații (secvențe de forma  $x_1, x_2, \dots$  ), dar nu i se furnizează rezultate prestabilite, obținute anterior și nici nu primește răsplată din partea mediului
- Stilul nesupervizat este mai comun creierului decât cel supervizat.
- Oamenii și animalele învață să-și analizeze mediul și să identifice obiectele și evenimentele din jurul lor.



# ÎNVĂȚAREA NESUPERVIZATĂ

## GENERALITĂȚI

- Putem obține grupuri de fructe pe baza anumitor caracteristici?



# ÎNVĂȚAREA NESUPERVIZATĂ-EXEMPLE [2]

- Dar dacă avem fructe și legume?

## WHAT IS CLUSTERING?

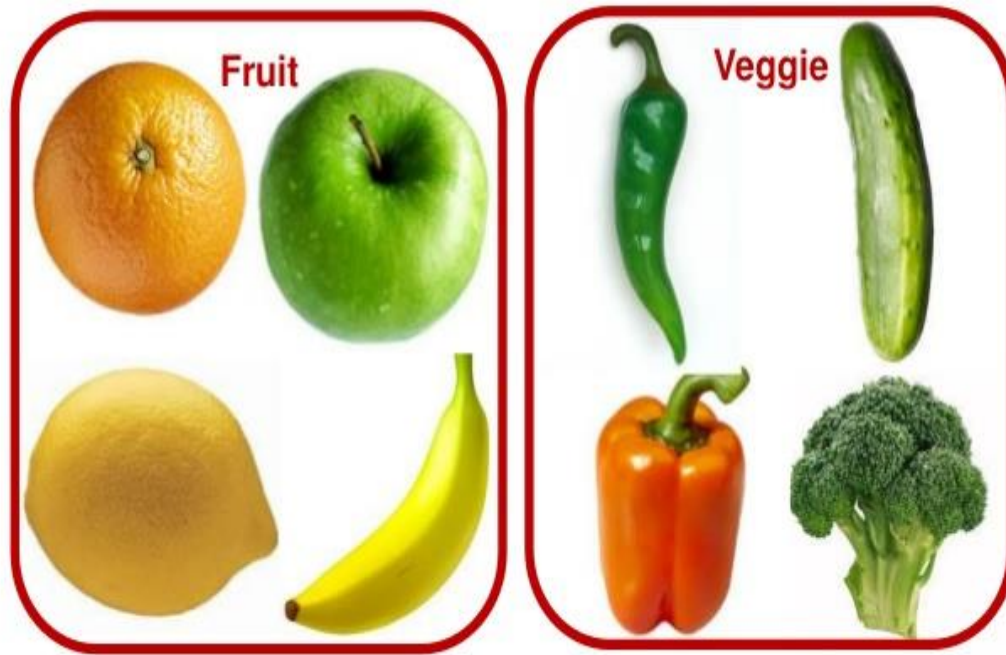
Grouping of objects





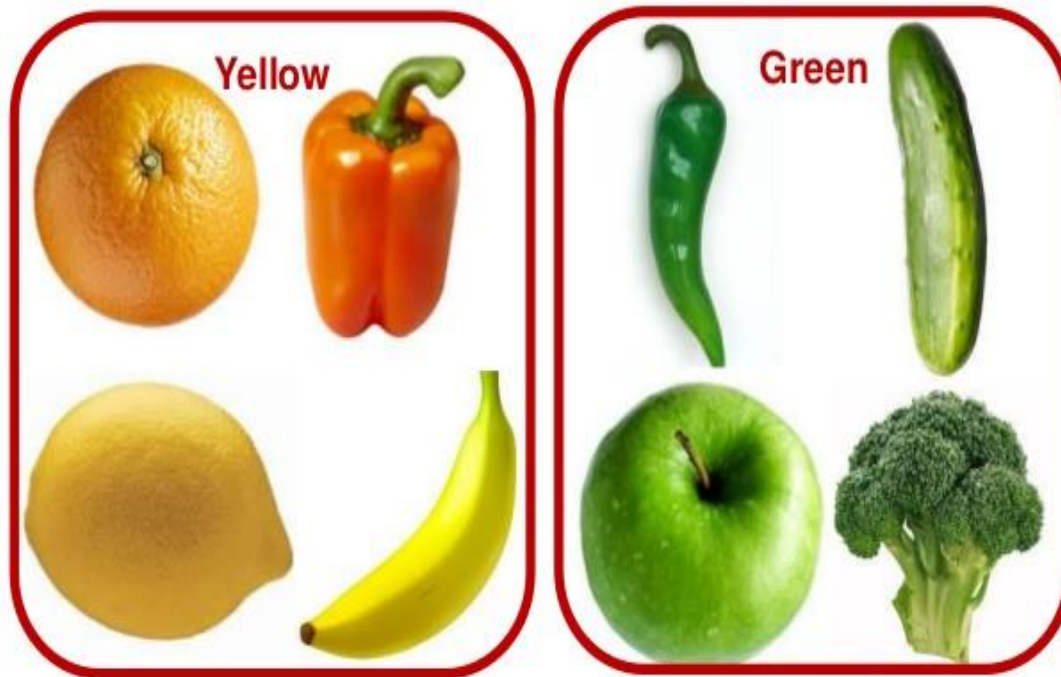
# ÎNVĂȚAREA NESUPERVIZATĂ-EXEMPLE [2]

## CLUSTERING I. (BY TYPE)



# ÎNVĂȚAREA NESUPERVIZATĂ-EXEMPLE [2]

## CLUSTERING II. (BY COLOR)



# ÎNVĂȚAREA NESUPERVIZATĂ-EXEMPLE [2]

## CLUSTERING III. (BY SHAPE)

Bushy



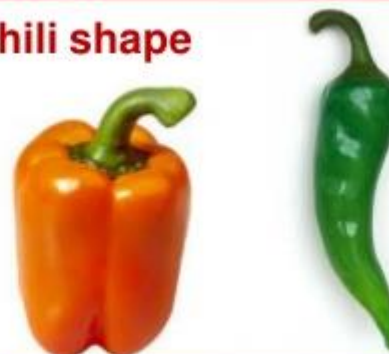
Longish



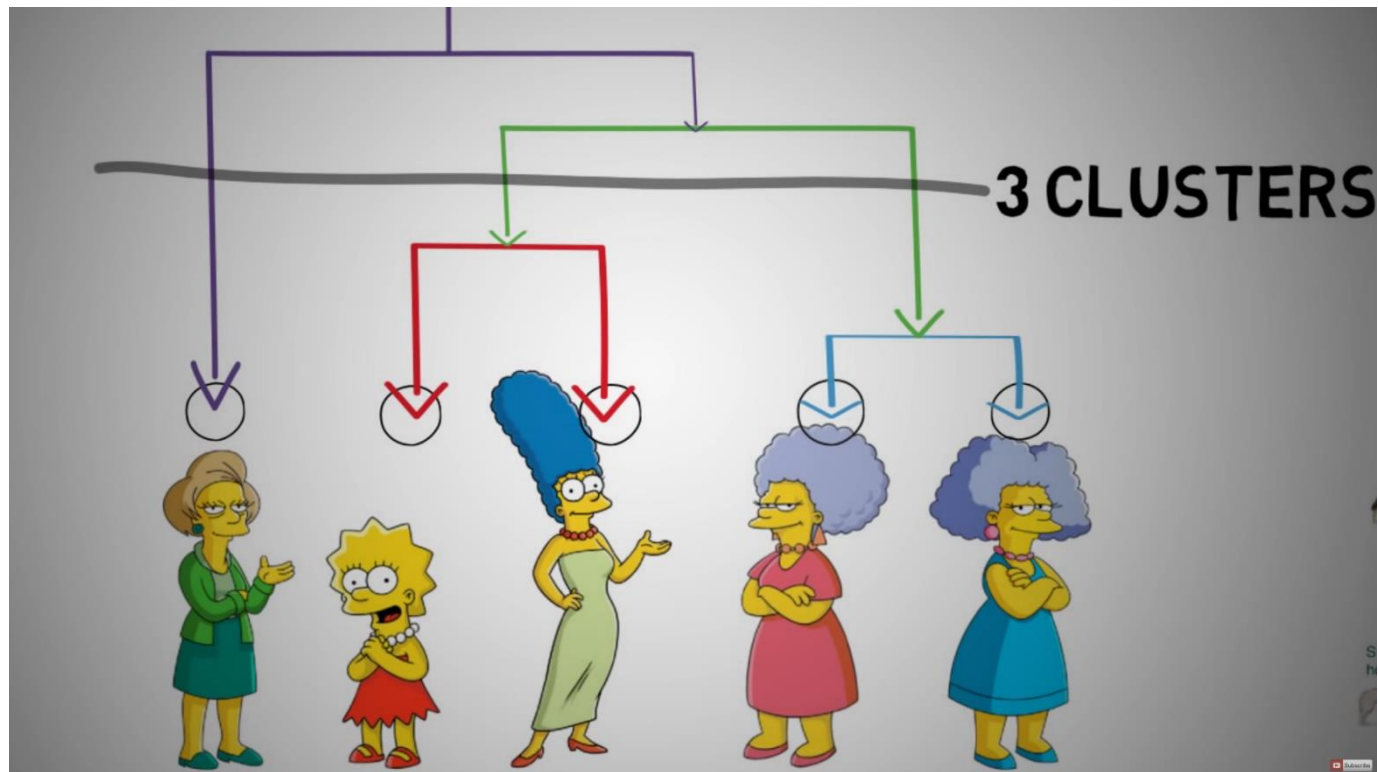
Ball



Chili shape



# ÎNVĂȚAREA NESUPERVIZATĂ-EXEMPLE



# ÎNVĂȚAREA NESUPERVIZATĂ

## GENERALITĂȚI

- “Ce dorim să învețe sistemul, dacă nu îi livrăm informație externă?”
  - descoperirea unor grupuri (clase) din setul de date inițial;
  - extragerea unor attribute ce caracterizează datele de intrare într-un mod mai compact;
  - identificarea unor coincidențe (similitudini) naturale în cadrul datelor primite.



# ÎNVĂȚAREA NESUPERVIZATĂ

## GENERALITĂȚI

- Utilitate în clasificarea de obiecte neetichetate.
- Plecând de la o mulțime de instanțe, se dorește găsirea unor mecanisme de a grupa obiectele pe baza unor similitudini, astfel încât obiectele din același grup să aibă o similitudine maximă.



# ÎNVĂȚAREA NESUPERVIZATĂ

## GENERALITĂȚI

### ○ Abordări

- gruparea datelor (*clustering*);
- rețele cu auto-organizare (*self-organizing maps*);
- reguli de asociere;
- algoritmi de maximizare a așteptărilor;
- analiza componentelor independente;
- analiza componentelor principale;
- descompunerea în valori singulare.
- metoda momentelor



# GRUPAREA DATELOR

## CLUSTERING

- *Clustering-ul* (gruparea datelor) [3] este procesul de grupare a obiectelor (instanțe, observații) pe baza caracteristicilor lor.
- Scopul acestei activități este ca în cadrul aceluiasi grup, obiectele să fie similare unele cu altele și diferite de obiectele din alte grupuri.
- Omogenitatea (similaritatea) din cadrul unui grup, denumit și *cluster*, precum și diferența cât mai mare între grupuri caracterizează acest proces.





# GRUPAREA DATELOR

## CLUSTERING

- În clasificarea supervizată se dă o colecție de date etichetate (preclasificate), problema fiind de a putea eticheta un nou set de date.
- În cazul *clustering-ului*, problema este de a grupa o colecție de date neetichetate în grupuri cu o anumită semnificație.
- Într-o anumită măsură, etichetele sunt asociate grupurilor, dar aceste etichete sunt date pe seama informațiilor inițiale (nu se oferă etichete din sursă externă), așa numitele etichete *data driven*.



# GRUPAREA DATELOR

## CLUTERING

- Componentele (pașii) procesului de grupare a datelor [4] :
  - stabilirea datelor de intrare (opțional selectarea atributelor—*feature selection*);
  - definirea unei măsuri de proximitate (vecinătate) adecvată domeniul datelor;
  - gruparea propriu-zisă (*clustering*);
  - abstractizare (dacă este cazul);
  - evaluarea rezultatelor.



# CLUSTERING

## CONCEPTE

- Stabilirea datelor se referă la identificarea numărului de clase (*clusteri*), numărului și tipurilor atributelor din cadrul setului de date.
- În cadrul acestei etape se poate aplica și selectarea atributelor (*feature selection*), proces de identificare a submulțimii, din mulțimea de attribute care va fi efectiv folosită pentru grupare.
- Tipuri de date
  - După nr de attribute- binare, discrete, continue
  - După tipul valorilor- calitative și cantitative



# CLUSTERING CONCEPTE

- O măsură de proximitate sau de similaritate reprezintă, de fapt, o funcție distanță, pentru a determina cât de similari/diferiți sunt doi clusteri
- Abstractizarea este procesul de extragere a unei reprezentări simple și compacte a setului de date. Se dorește simplitate (să fie ușor de înțeles), reliefată printr-o descriere concisă a fiecărui grup de obiecte, cu ajutorul unor termeni corespunzători acestui domeniu de *clustering* (ex: medoizi, centroizi).



# CLUSTERING CONCEPTE

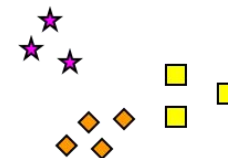
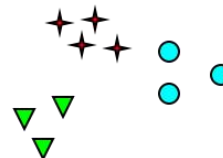
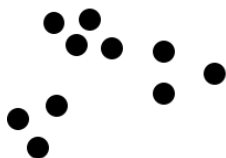
- Evaluarea rezultatelor
  - Se poate folosi o examinare externă care presupune compararea structurii obținute cu cea *a priori* sau o examinare internă care evaluează structura intrinsecă potrivită pentru setul de date inițial.



# CE ESTE UN CLUSTER?

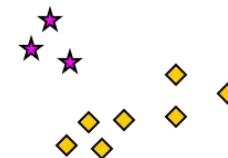
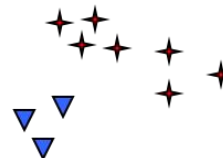
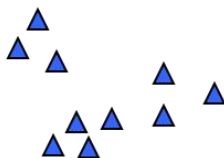
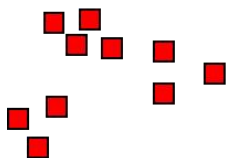


So tell me how many clusters do you see?



How many clusters?

Six Clusters



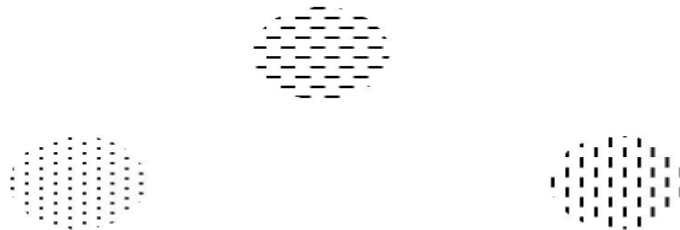
Two Clusters

Four Clusters



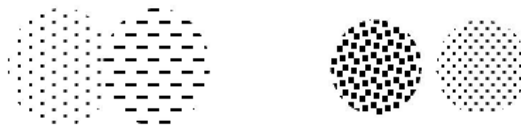
# CE ESTE UN CLUSTER?

- Cormack (1971) și Gordon (1999) [5] au stabilit că un *cluster* poate fi definit în funcție de două proprietăți interne: omogenitate și izolare externă (separare).
- Vipin Kumar [3] prezintă câteva definiții ale unui *cluster*:
  - definiție bazată pe delimitări/separări: *cluster* = o mulțime de puncte în care orice punct este mai apropiat de un altul din mulțime, decât de un punct din afara ei



# CE ESTE UN CLUSTER?

- definiție bazată pe centru: *cluster* = o mulțime de obiecte în care un obiect este mai apropiat de centrul mulțimii decât de centrul unei alte mulțimi



- definiție bazată pe similaritate: *cluster* = mulțime de obiecte care sunt “similare”, dar diferite de obiectele din alte grupuri





# MĂSURI DE SIMILARITATE

- O funcție de similaritate “măsoară” cât de bun este un anumit grup. În general termenul utilizat pentru astfel de măsuri este proximitate sau vecinătate. Două instanțe sunt “apropiate”, atunci când disimilaritatea (distanța) este mică sau similaritatea este mare.
- Distanța-minimă, similaritatea-maximă
- Exemple:
- Distanța euclideană  $d(i,j) = \sqrt{\sum_{k=1}^n |x_{ik} - x_{jk}|^2}$
- Distanța *Manhattan*:  $d(i,j) = \sum_{k=1}^n |x_{ik} - x_{jk}|$ .



# MĂSURI DE SIMILARITATE

- Similaritatea cosinus

- $\cos(A,B) = (A \cdot B) / (||A|| \cdot ||B||)$ ,
  - unde  $\cdot$  este produsul vectorial, iar  $||A||$  lungimea vectorului A.

- Similaritate Jaccard – unde A, B mulțimi

- între 0 și 1

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$



# TEHNICI DE CLUSTERING

- metode ierarhice
  - algoritmi aglomerativi;
  - algoritmi divizivi.
- metode partiționale
  - *k-medoids*;
  - *k-means*;
  - algoritmi bazați pe densitate.
- metode bazate pe grilă /rețea
- algoritmi bazați pe scalabilitate
- algoritmi pentru date de dimensiuni mari
  - *clustering* de subspațiu;
  - tehnici de proiecție;



# TEHNICI DE CLUSTERING

- Jain [4] consideră că metodele de grupare a datelor pot fi organizate în funcție de anumite criterii:
  - **aglomerativ** vs. **diviziv**- abordare ce se referă la modul de formare a grupurilor;
  - ***hard*** (strictă) vs ***fuzzy*** (maleabilă/permisivă) –se referă la modul de atribuire a instanțelor la un anumit grup.
    - Atribuirea *hard* indică faptul că un obiect poate aparține unui singur grup, de-a lungul procesului. Atribuirea *fuzzy* (*soft*) definește grade de apartenență la diferite grupuri, pentru fiecare instanță;

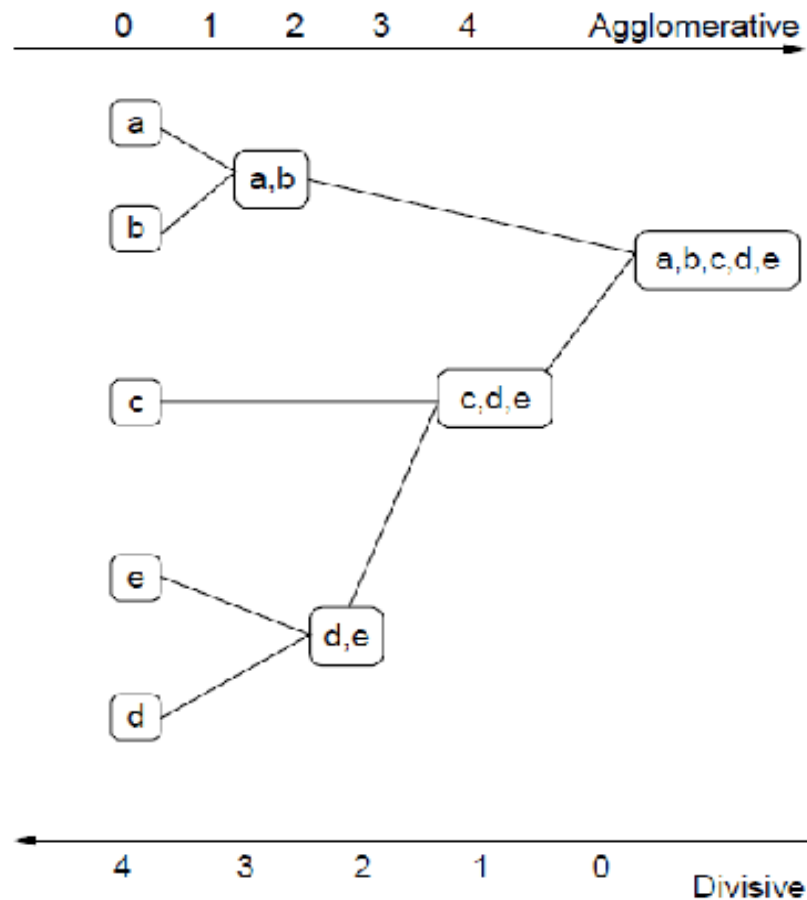


# CLUSTERING IERARHIC

- În clasificarea ierarhică datele nu sunt partiționate într-un anumit număr de grupuri dintr-un singur pas.
- Tehnicile ierarhice se împart în metode aglomerative și divizive.
- Cele aglomerative procesează o serie de fuziuni succesive, a  $n$  instanțe, în grupuri. Cele divizive separă, succesiv,  $n$  indivizi în grupuri fine, mai precise.



# CLUSTERING IERARHIC [5]



# CLUSTERING IERARHIC

- Aglomerativ: Se pornește cu obiecte individuale. Se unifică, gradual, obiecte care au similaritate maximă (distanță minimă). Se continuă până când toate obiectele vor fi conținute într-un singur grup sau până când se ajunge la număr dorit de *clusteri*.
- Diviziv: Se pornește cu un grup ce conține toate obiectele. Se separă, gradual, grupul în două, atribuind obiectele celor doi noi *clusteri* astfel încât să se maximizeze similaritatea din cadrul fiecărui grup. Se continuă divizarea până când se obțin *clusteri* ce conțin un singur obiect sau până când se obține numărul dorit de *clusteri*.



# CLUSTERINGUL PARTIȚIONAL

- Un algoritm de *clustering* partițional obține o partiție a setului de date, comparativ cu *clustering-ul* ierarhic care produce o structură
- O problemă a algoritmilor partiționali este alegerea numărului de *clusteri* doriți a fi obținuți în urma procesului.
- Metodele partiționale au ca scop construirea sau găsirea unei partiții formate din  $k$  *clusteri*, plecând de la un set de date cu  $n$  instanțe.
- Aceste tehnici cuprind și câteva metode euristice: *k-means* și *k-medoids* [5]. Algoritmul *k-means* presupune că fiecare grup este reprezentat de centrul lui, iar algoritmul *k-medoids* implică faptul că fiecare grup este reprezentat de unul dintre obiectele lui.

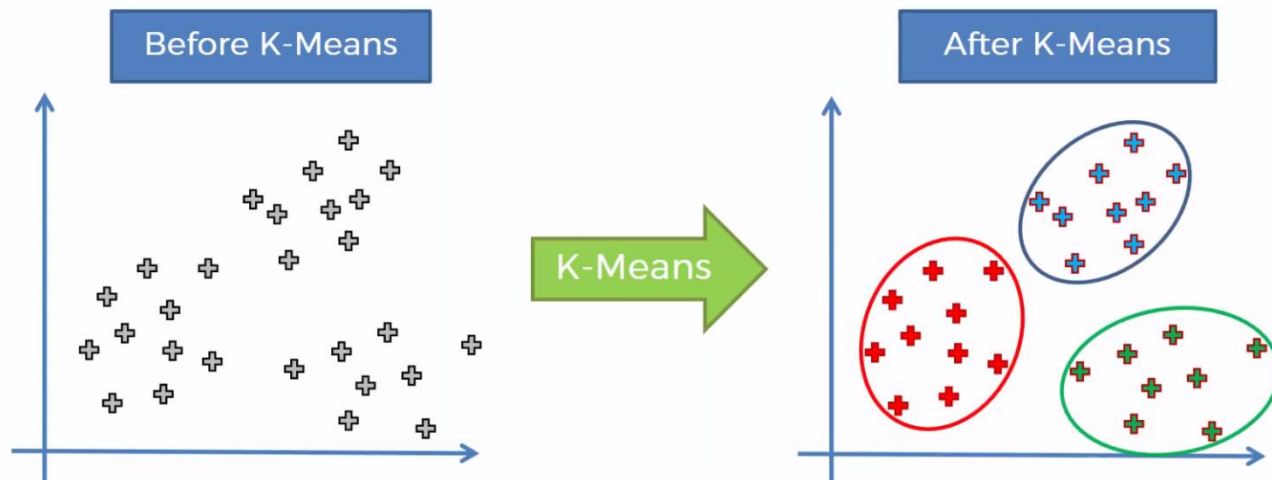




# K-MEANS

- Denumirea provine de la reprezentarea a  $k$  *clusteri* pe baza mediei obiectelor din cadrul lor.
- Un astfel de obiect, definit ca fiind media instanțelor unui *cluster*, poartă denumirea de centroid [5]

## What K-Means does for you



# K-MEANS

## ○ Algoritmul

- selectează  $k$  instanțe ca fiind centroizii inițiali;
- atribuie obiectele la cel mai apropiat centroid;
- recalculează centroidul fiecărui grup;
- revenire la pasul 2, oprirea realizându-se atunci când nu mai au loc modificări.



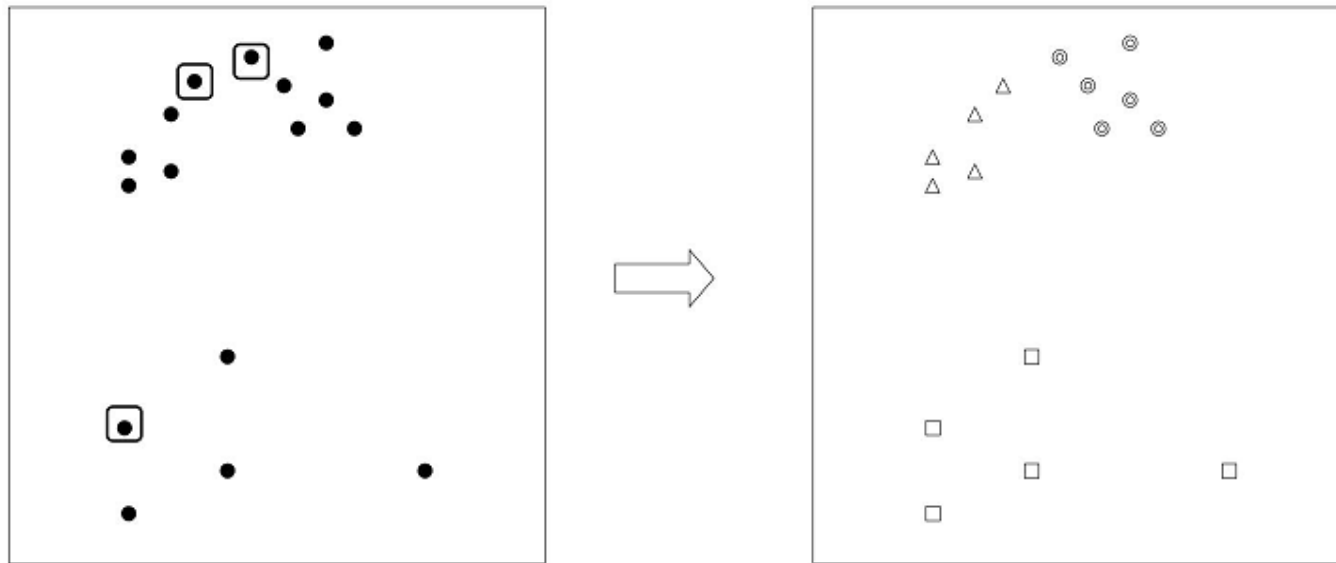
# K-MEANS

- Alegerea centroizilor inițiali
  - etapa de bază a procedurii *k-means*. Este foarte ușor să alegi aleator  $k$  obiecte care să fie centroizi, dar rezultatele, de regulă, nu sunt mulțumitoare.
  - Este posibilă și abordarea ce implică mai multe rulări, fiecare cu un număr diferit de obiecte alese aleator ca reprezentând centroizi



# K-MEANS

- Alegerea naturală a centroizilor [5]



- centroizii inițiali sunt aleși, adesea, din regiuni dense, în așa fel încât obiectele să fie foarte bine separate, pentru ca doi centroizi să nu fie aleși din cadrul aceluiași grup.

# K-MEANS EXEMPLU

- Se dau punctele  $P1(2,3)$ ,  $P2(3,1)$ ,  $P3(4,2)$ ,  $P4(11,5)$ ,  $P5(12,4)$ ,  $P6(12,6)$ ,  $P7(7,5)$ ,  $P8(8,4)$ ,  $P9(8,6)$
- Aplicați k-means pornind de la centroizii  $K1=P2$  și  $K2=P8$
- Se folosește distanța euclideană

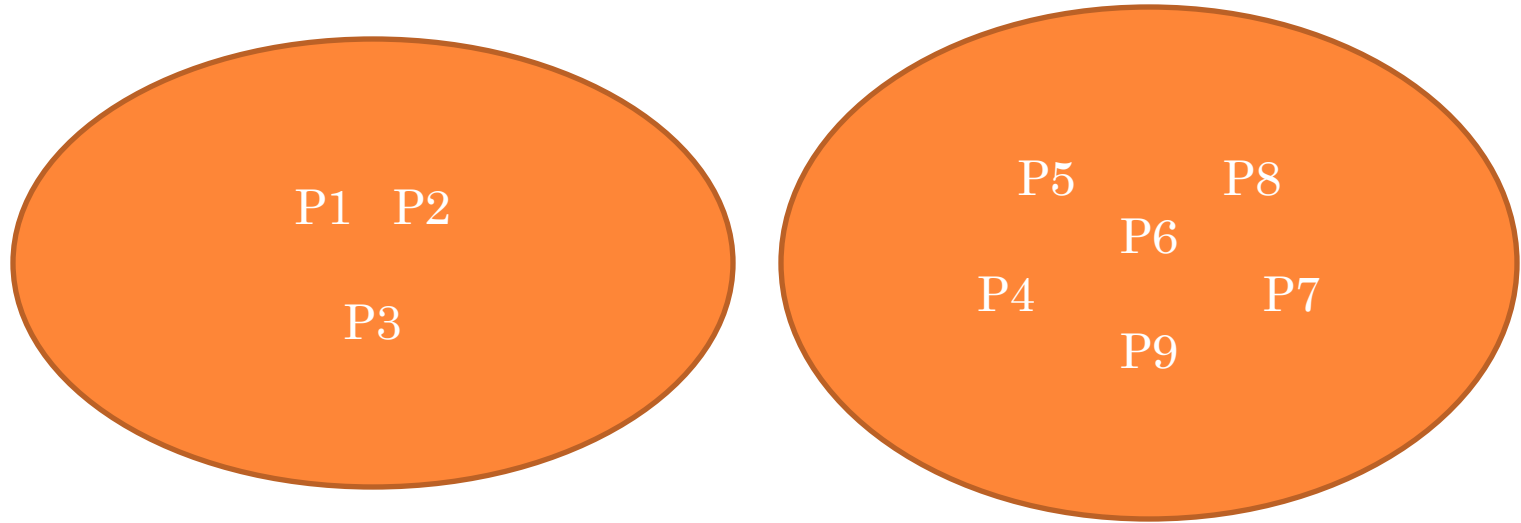


# REZOLVARE

- Calculăm distanța dintre fiecare obiect neselectat și centroizi
- $d(P1, K1)=2,23$  și  $d(P1, K2)=6,08 \rightarrow P1$  se atribuie lui  $K1$
- $d(P3, K1)=1,41$   $d(P3, K2)=4,47$  deci  $P3$  în  $K1$
- $d(P4, K1)=8,94$   $d(P4, K2)=3,16$  ,  $P4$  în  $K2$
- .....



# REZOLVARE



- Calculăm noii centroizi
  - $K1=(3,2)$   $K1x=9/3$   $K1y=6/3$
  - $K2=...$



# K-MEDOIDS

1. se selectează  $k$  obiecte reprezentative numite medoizi
2. se înlocuiește unul dintre obiectele selectate (medoizi) cu unul dintre obiectele neselectate. Se calculează distanța dintre fiecare obiect neselectat și cel mai apropiat medoid candidat, iar apoi distanța este însumată pentru toate punctele/ instanțele. Această distanță reprezintă costul configurației curente
3. se selectează configurația cu costul cel mai mic. Dacă există o nouă configurație, se repetă pasul 2;
4. dacă nu există o configurație nouă, se atribuie fiecare obiect neselectat celui mai apropiat medoid și algoritmul se oprește.





# EVALUAREA REZULTATELOR

- De ce să evaluăm procesul de *clustering*?
  - pentru a compara diferiți algoritmi de *clustering*;
  - pentru a compara două grupuri;
  - pentru a compara mulțimi de *clusteri*.
- Aspecte legate de validarea *clustering-ului*:
  - determinarea numărului ideal de *clusteri*;
  - găsirea unor legături între structurile identificate în urma procesului, din cadrul setului de date, și informații externe legate de datele de intrare;
  - evaluează cât de bine se potrivesc rezultatele produse de analiza de *clustering* cu setul de date, fără a referi informații externe (utilizează doar datele).



# EVALUAREA REZULTATELOR

- Există două abordări referitoare la activitatea de evaluare a *clustering-ului*:
  - externă-bazată pe informații anterioare despre date (posibil să cunoaștem etichetele datelor);
  - internă-bazată doar pe informații intrinseci (doar setul de date), fără să utilizăm surse externe.



# EVALUAREA INTERNA

- Indexul Dunn [6] -acest index determină raportul dintre cea mai mică distanță *intercluster* și cea mai mare distanță *intracluster*, în cadrul partiției.

$$D = \frac{\min_{1 \leq i < j \leq n} d(i, j)}{\max_{1 \leq k \leq n} d'(k)} ,$$

- unde  $d(i, j)$  reprezintă distanța *intercluster* (distanța între *clusterul*  $i$  și *clusterul*  $j$ ), iar  $d'(k)$  reprezintă *distanța intracluster* (distanța în *clusterul*  $k$ ).
- Ca distanță între doi *clusteri* se poate folosi orice distanță, de exemplu distanța între centroizii grupurilor. Distanța *intracluster* ( $d'$ ) poate fi cea mai mare distanță între orice pereche de elemente din *clusterul*  $k$ . Valorile mari pentru indexul Dunn, indică soluții bune.



# EVALUAREA INTERNA

- Coeficientul Silhouette [6]

$$s(i) = \frac{(b(i) - a(i))}{\text{Max}\{a(i), b(i)\}}$$

- unde  $a(i)$  este distanța medie între obiectul  $i$  și celelalte obiecte din  $X_j$ , iar  $b(i)$  reprezintă cea mai mică distanță medie între obiectul  $i$  și obiectele din ceilalți *clusteri*, din care  $i$  nu face parte.
- Valorile pentru  $s(i)$  sunt din  $[-1,1]$ , soluții bune când indexul este cât mai aproape de 1



# EVALUAREA INTERNA

- Indexul Davies-Bouldin [6]

$$BD = \frac{1}{c} \sum_{i=1}^c \max_{i \neq j} \left\{ \frac{d(X_i) + d(X_j)}{d(c_i, c_j)} \right\}$$

- unde  $c$  reprezintă numărul de *clusteri*,  $d(X_i)$  și  $d(X_j)$  sunt distanțele între toate obiectele din *clusterul*  $X_i$ , respectiv  $X_j$ , și centroizii acelor *clusteri*.  $D(c_i, c_j)$  este distanța între centroizii celor doi *clusteri*,  $c_i$  și  $c_j$ . Cu cât acest index produce valori mai mici, cu atât se obțin soluții mai bune.



# EVALUARE EXTERNĂ

- Acuratețea
- Precizia
- Specificitatea
- Senzitivitatea (recall)



# APLICAȚII ALE CLUSTERINGULUI

## ○ Marketing și comerț

- segmentare=modalitatea de organizare a clienților în grupuri, în funcție de preferințe pentru anumite produse, trăsături sau așteptări



# APLICAȚII ALE CLUSTERINGULUI

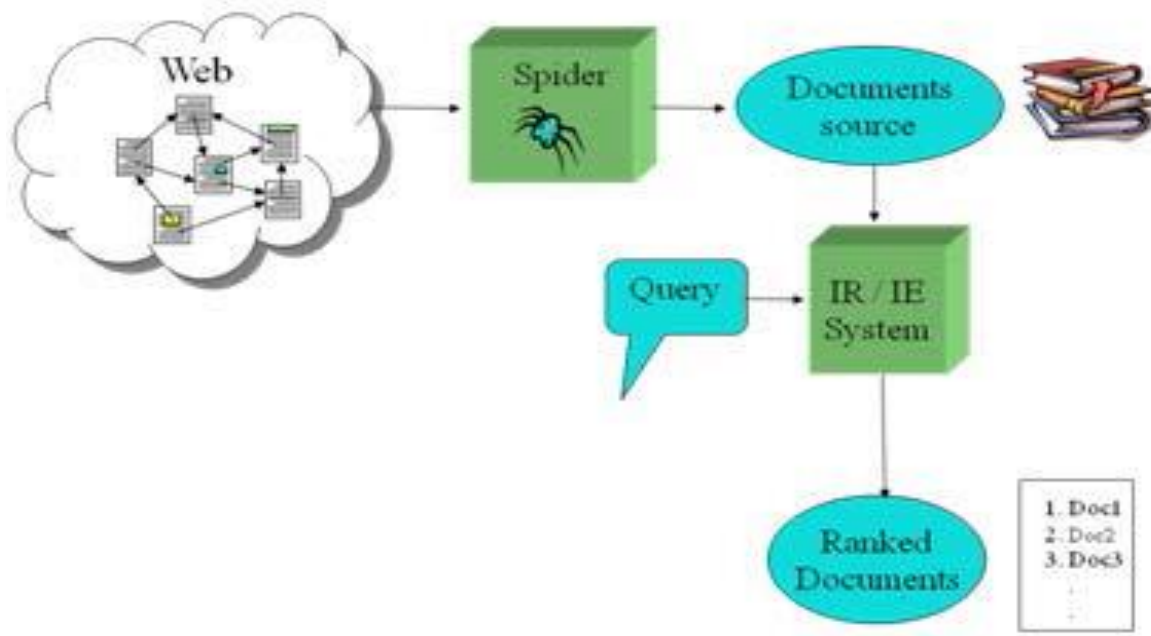
- Geico [7] - planificări pentru personalizarea ofertelor de asigurări auto și voia să înțeleagă exact ce dorințe și așteptări au clienții de la furnizorul de asigurări
- Respondenții implicați în chestionare au fost grupați în clustere/segmente
- acțiunile luate de clienții din același segment ilustrează răspunsuri similare, dar diferite de cele date de clienții din restul segmentelor. Așadar, aspectele privind asigurarea auto pe care un client le consideră importante vor fi importante și pentru restul clienților din același *cluster/segment*





# APLICAȚII ALE CLUSTERINGULUI

- Motoare de căutare [9]




# APLICAȚII ALE CLUSTERINGULUI

- Analiza crimelor [8]
- Pentru aplicarea *clustering-ului* în analiza crimelor s-au utilizat seturi de date înregistrare de poliția din Anglia și Țara Galilor în perioada 1990-2011. Seturile de date se refereau la omucideri (uciderea unei alte persoane).



# APLICAȚII ALE CLUSTERINGULUI

- *k-means* a avut drept obiectiv urmărirea ratei criminalității
  - s-au luat în considerare omuciderile și s-au format *clusteri* ce ofereau informații legate de numărul de omucideri din diferiți ani.
  - *Clusterul 0* cuprindea între 0 și 15 crime, *clusterul 1* între 0 și 60, *clusterul 2* între 0 și 600, etc.
  - De pe graficele corespunzătoare fiecărui *cluster* se observa anul cu număr de omucideri minim și anul cu număr maxim. La finalul analizei s-a observat că omuciderile au scăzut din 1990 până în 2011.
  - tendințele legate de crime, infracționalități pentru următorii ani și se pot proiecta tehnici de prevenție și de reducere a acestor acțiuni.
- 

# APLICAȚII ALE CLUSTERINGULUI

- K-Means în predicția performanței studenților [10]
  - monitorizarea performanțelor studenților
  - rezultatele studenților de la un institut privat din Nigeria
- Economie
- Finanțe – identificarea unor categorii de clienți
- Bio-arheologie



# CONCLUZII

- *Clustering-ul* reprezintă o ramură importantă a învățării nesupervizate, aplicarea lui având rezultate semnificative în multe domenii
- *clustering-ul* simplifică mult o muncă manuală care, uneori, se dovedește a fi anevoioasă.



## BIBLIOGRAFIE

- [1] Nils J.Nilsson, *Introduction to Machine Learning*, Stanford University, Stanford, 1996
- [2] Gabor Veress, Clustering training, 2013, <https://www.slideshare.net/gveress/cluster-training-2013>
- [3] Vipin Kumar, *An introduction to cluster analysis for data mining*, course, 2000
- [4] A. K. Jain, M. N. Murty and P. J. Flynn, *Data clustering: A review.*, ACM Comput.Surv., 31(3):264-323, 1999
- [5] Brian S.Everitt, Sabine Landau, Morven Leese, Daniel Stahl, *Cluster Analysis*, 5th edition, Wiley, Londra, 2011



# BIBLIOGRAFIE

- [6] E.Rendón, I.Abundez, A.Arizmendi and E. M. Quiroz, *Internal versus External cluster validation indexes*, International Journal of Computers and Communications, Vol.5: 27-34, 2011
- [7] R. Venkatesan, *Cluster analysis for segmentation*, Darden Business Publishing, University of Virginia, 2007
- [8] Jyoti Agarwal, Renuka Nagpal, Rajni Sehgal, *Crime Analysis using K-Means Clustering*, International Journal of Computer Applications, 83(4):1-4, 2013



# BIBLIOGRAFIE

- [9] L.V. Bijuraj, *Clustering and its Applications*, National Conference on New Horizons in IT, 169-172, 2013
- [10] O.J. Oyelade, O.O. Oladipupo, I.C. Obagbuwa, *Application of k-means clustering algorithm for prediction of students' academic performance*, International Journal of Computer Science and Information Security, 1:292-295, 2010

