CLUSTERIZARE

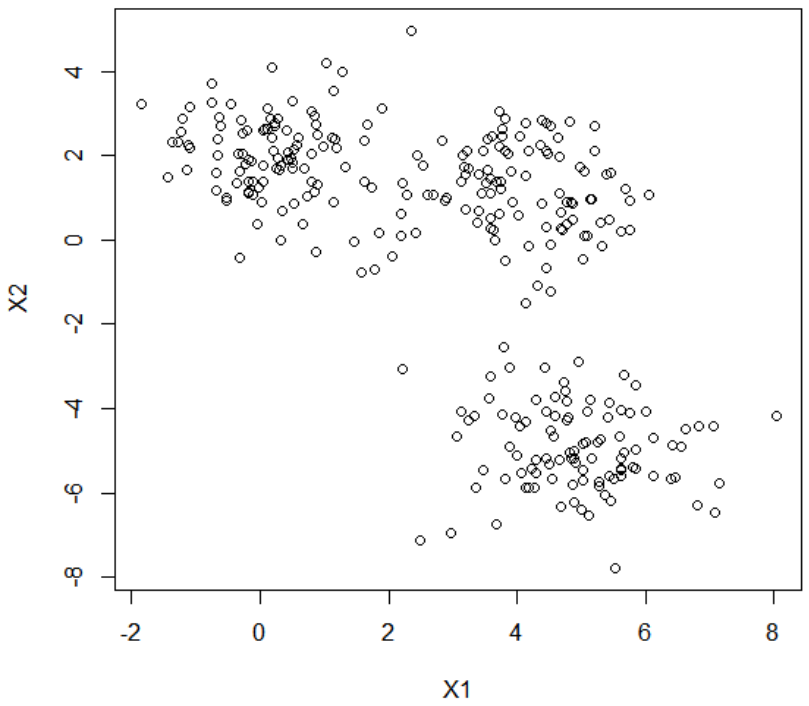
CLUSTERIZARE

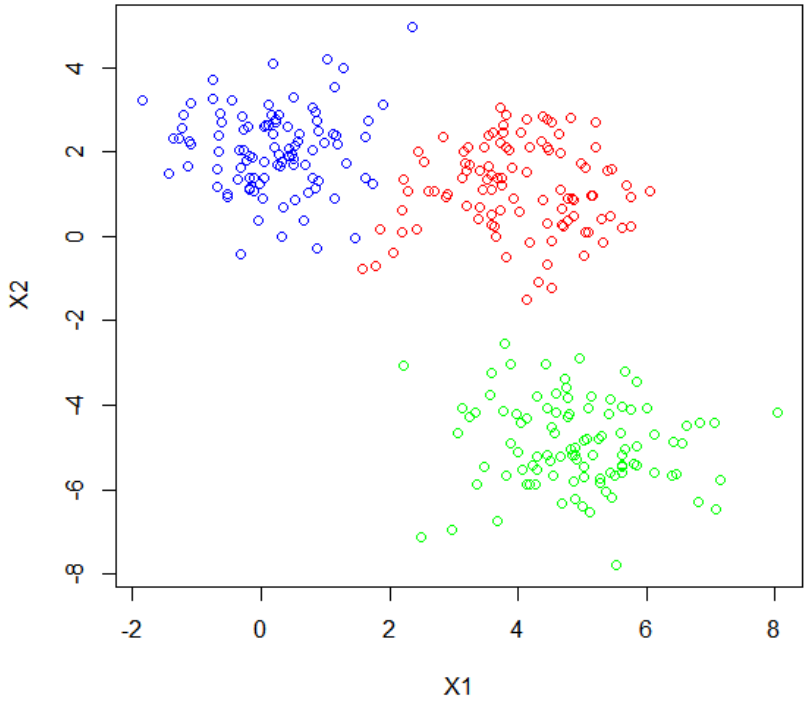
- Clusterizarea este procesul de divizare a unei baze de date in grupe de inregistrari similare astfel incat membrii aceleasi grupe sa fie cat se poate de apropriati unul de altul, iar grupurile sa fie cat se poate de departate unele de celelalte.

- Clusterizarea este o metoda nesupervizata (unsupervised).
- In clusterizare nu exista nici o mt de date preclasificate si nu se face nici o distinctie intre variabilele indep si cele dependente.
- Var (atributele) in fct de care se face operatia de clusterizare se numesc var de intrare, iar dimensiunea problemei este data de nr var de intrare.

- Data mining inseamna analiza automata a bazelor de date in scopul de a gasi sabloane, tipare, informatii
- De multe ori informatiile pot fi prea multe sau greu de detectat. Ca o analogie, daca ascultam la radio uneori auzim clar un post de radio, alteori anumite semnale sunt perturbate de alt post de radio, deci daca undele sonore sunt prea complexe este mai greu sa gasesti informatiile, daca in schimb sunt mai simple sunt mai usor de detectat.
- Cand avem de rezolvat o problema complexa de obicei o impartim in probleme mai mici si mai simple pe care le rezolvam mai usor.
- Cand trebuie sa intelegem o intrebare mai complexa, de obicei o impartim in intrebari mai simple pe care incercam sa le intelegem.
- Daca suntem intrebati care este culoarea frunzelor arborilor dintr-o padure raspundem ca aceasta depinde in functie de tipul arborilor (daca le cad frunzele sau nu) si in functie de anotimp. Atunci o baza de date ce ar cuprinde informatii despre arbori: tipul, culoarea in fct de anotimp, varsta, inaltimea ar putea fi impartita in grupuri dupa tip si anotimp, grupuri pentru care culoarea frunzelor ar fi similara.

CLUSTERIZARE

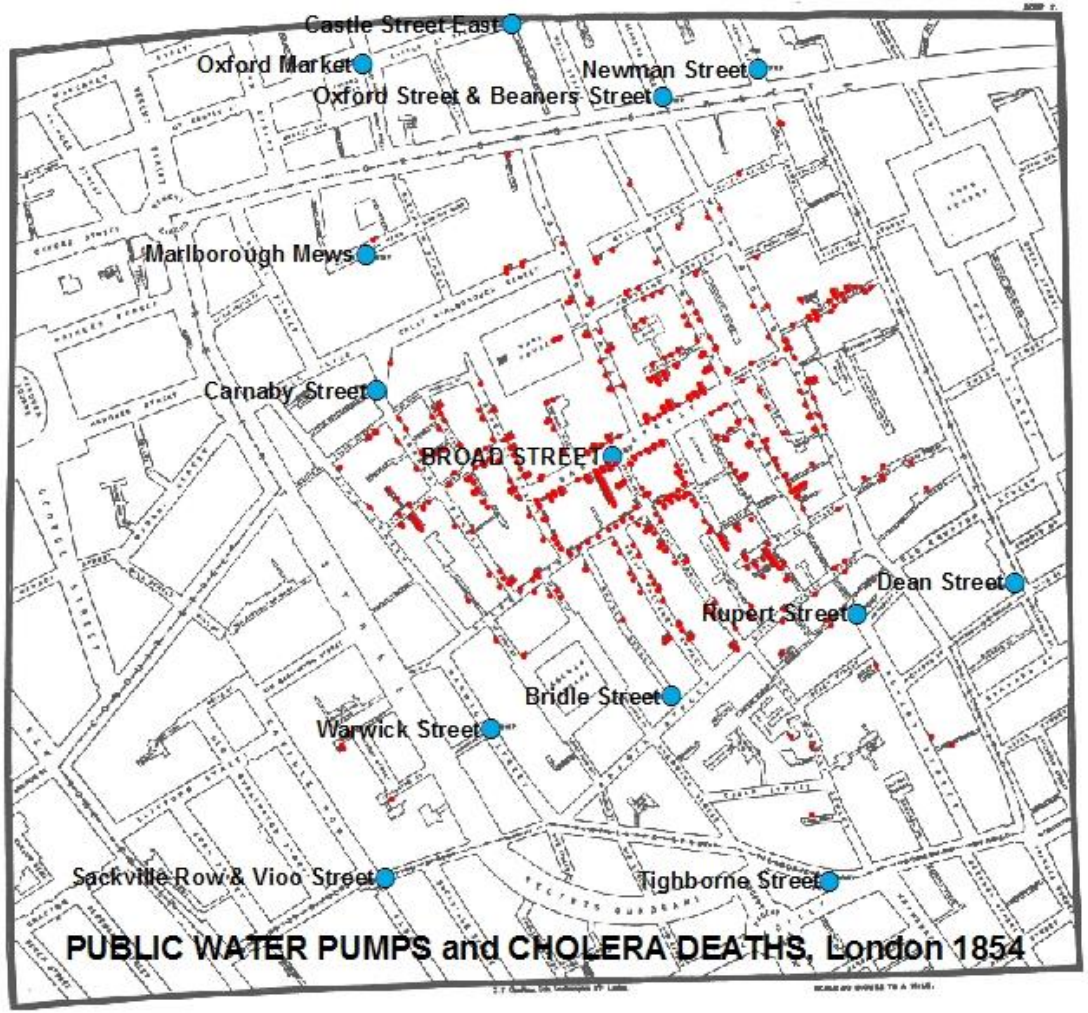




- In timpul epidemiei de holera din Londra in 1854, un medic John Snow credea ca boala se transmitea prin apa infestata. A trasat o harta a strazilor Londrei si a localizat zonele in care fusesera descoperite cazurile de holera si pozitiile tuturor puturilor de apa. A descoperit ca toate cazurile se concentrau in jurul a trei puturi de apa care s-au dovedit apoi ca era infestate.Teoria lui a fost astfel demonstrata.

- http://www.math.uah.edu/stat/data/Snow.html
- http://www.ncgia.ucsb.edu/pubs/snow/snow.html
- http://www.udel.edu/johnmack/frec480/cholera/cholera2.html

HARTA LUI JOHN SNOW



- Pp ca avem n variabile de intrare:

$$X_1, X_2, \ldots, X_n$$

- Atunci fiecare inreg ce contine cate o valoare pt cele n var reprezinta un punct in spatial n-dimensional.

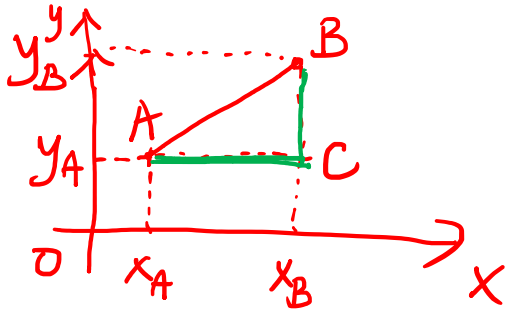
$$X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n$$

EXEMPLU

CYLINDER	DISPLACEMENT	HORSEPOWER	WEIGHT	ACCELERATION	MPG
8	307	130	3504	12	18
8	350	165	3693	11.5	15
8	318	150	3436	11	18
8	304	150	3433	12	16
8	302	140	3449	10.5	17

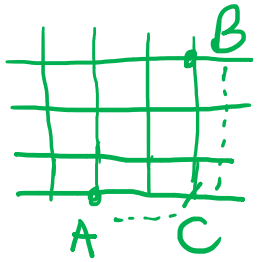
- Estimarea consumului de combustibil al unei masini in functie de cateva caracteristici ale sale:
- greutate
- marimea motorului
- puterea motorului
- nr. cilindri
- acceleratie
- displacement

$$d_M(A,B) = |x_A - x_B| + |y_A - y_B|$$



- Obs. Alg functioneaza numai pt atribuite cu valori numerice. O posibila fct distant care sa descrie notiunea de cel mai apropriat este fct distanta euclidiana intre doua pct:

$$d(A,B) = \sqrt{(x_A - x_B)^2 + (y_A - y_B)^2}$$



$$X = (x_1, x_2, \ldots, x_n)$$

$$Y = (y_1, y_2, \ldots, y_n)$$

$$d(X,Y) = \sqrt{(x_1 - y_1)^2 + \ldots + (x_n - y_n)^2}$$

ALEGEREA LUI K=NUMAR DE CLUSTERE

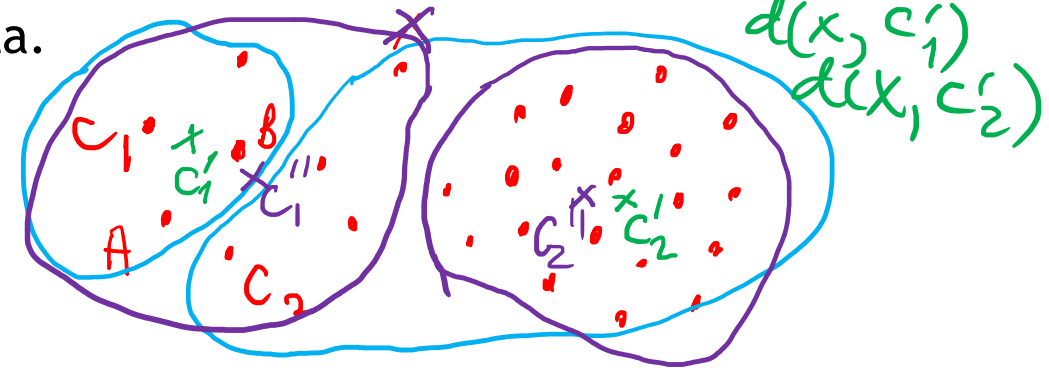
- In cele mai multe cazuri nu ex un motiv apriori pt selectarea lui k. De obicei se alege o valoare a lui k, se aplica alg, se evalueaza rezultatele,apoi se incearca o alta valoare si se analizeaza.

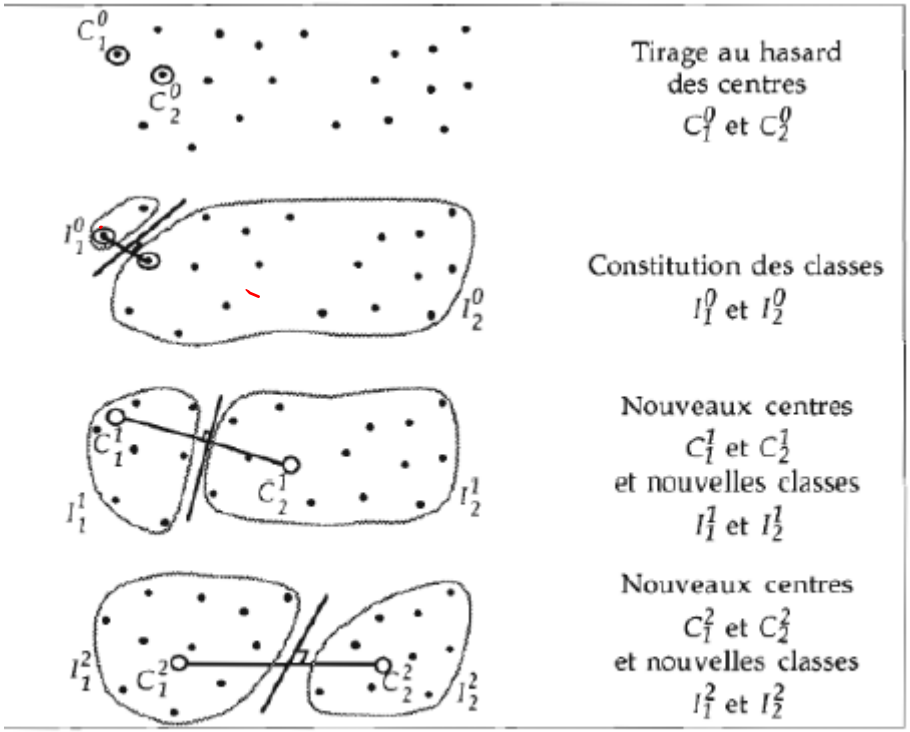
top 10 Algoritmi de data mining

ALG K-MEANS (MACQUEEN 1967)

- Fie k fixat, k=nr clustere.
- Se aleg la intamplare k puncte (inregistrari) ca fiind centerele initiale ale celor k clustere. (MacQueen propune alegerea primelor k inreg)
- Pt fiecare inreg determina cel mai apropriat centru si atribuie inregistrarii clusterul asociat centrului.
- Pt fiecare cluster, calculeaza media inreg din cluster. Muta centrul clusterului in pct coresp mediei.
- Repeta pasii 2 si 3 pana cand se obtine convergenta adica pana cand nr de reatribuiri ale clusterelor este mai mic decat o valoare ε data.

K=2





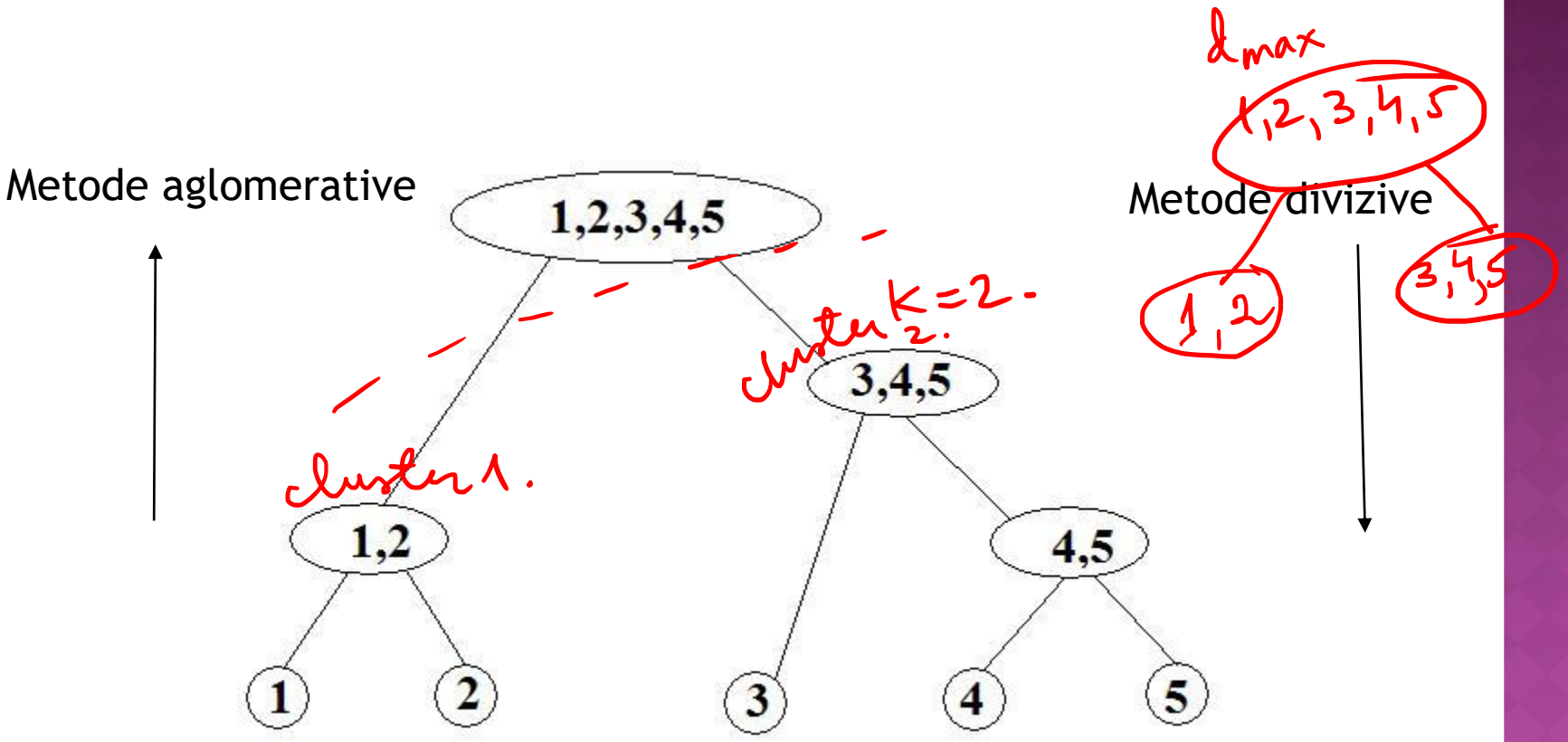
Lebart et al., 1995 ; page 149.

CLUSTERIZAREA IERARHICĂ

- Metode aglomerative de clusterizare – presupun o serie de fuziuni a inregistrarilor initiale din n clustere (cate inreg sunt) in grupuri din ce in ce mai putine de inreg, pana cand se obtine nr de clustere dorit.
- Metode divizive de clusterizare – pp ca initial toate inreg fac parte dintr-un singru cluster pe care apoi il vor imparti in grupuri.

- Metodele de clusterizare ierarhice se pot reprezenta printr-o diagrama 2D numita dendograma care prezinta fuziunile sau divizarile facute.

DENDOGRAMA:


Metode aglomerative
1,2,3,4,5
3,4,5
1,2
4,5
1
2
3
4
5
cluster 1.
cluster k=2.
Metode divizive
dmax
1,2,3,4,5
1,2
3,4,5

METODE AGLOMERATIVE:

- Produc o serie de partitii ale inreg: $P_n, P_{n-1}, \ldots P_1$
- unde n=nr total de inreg
- $P_n = \{\{i_1\}, \{i_2\}, \ldots, \{i_n\}\}$ n mt (n clustere)
- .
- .
- .
- . $P_1 = \{i_1, i_2, \ldots, i_n\}$ 1 cluster
- La fiecare pas metoda alege cate doua clustere – mai exact doua clustere care sunt cele mai apropriate:
- Alegem R, S a.i. . Unim R si S.
- Diferenta dintre metode apare datorita modurilor diferite de a masura dist intre clustere.

MĂSURAREA DISTANȚEI DINTRE CLUSTERE

- ***Clusterizare cu leg simpla***
- Dist intre clustere= cel mai scurt drum intre clustere
- $$D(r,s) = \min\{d(i,j) / i \text{ este in clusterul } r \text{ si } j \text{ in clusterul } s\}$$
- ***Clusterizare cu leg completa***
- Dist intre clustere= distanta dintre cele mai departate inreg din clustere
- $$D(r,s) = \max\{d(i,j) / i \text{ este in clusterul } r \text{ si } j \text{ in clusterul } s\}$$
- ***Clusterizare cu leg medie***
- D(r,s)=media{d(i,j)/i in clusterul r ,j sunt in clusterul s}= $\frac{\sum d(i,j)}{card(r) \cdot card(s)}$
- ***Clusterizare cu leg medie de grup*** – r si s se unesc a.i. dupa unire dist medie din interiorul fiecarui cluster sa fie minima. Pp ca noul cluster format prin unirea lui r cu s este t. Atunci:
- D(r,s)=media{d(i,j)/i ,j sunt in clusterul t format prin fuzionarea lui r cu s}
- ***Clusterizare cu legatura Ward:*** Dist = cresterea in suma patratelor erorii ESS dupa fuzionarea celor doua clustere intr-unul singur. Se aleg pasi succesivi care sa minimizeze cresterea in ESS la fiecare pas.
- $$ESS(X) = \sum_{i=1}^{n} \|x_i - media\|^2$$
- X=mt, n = nr elem ale lui X
- D(r,s)=ESS(t=cluster obt prin fuzionarea lui r si s) –(ESS(r) +ESS(s))

METODE DIVIZIVE

- Algoritm:
- Initial toate inreg sunt intr-un singur cluster.
- Se decide un prag pt dist.
- Se calc dist dintre orice 2 inreg si se det perechea cu cea mai mare dist.
- Dist max se compara cu pragul pt dist.
- Daca dist max > prag pt dist atunci grupul se imparte in doua. Cele 2 inreg se pun in clustere diferite iar celelalte pcte se pun in clusterul cel mai apropriat. Si se repeta procedeul.
- Daca dist max < prag pt dist atunci stop.

ALEGEREA LUI K - METODA COTULUI

- Se compara WSS = suma patratelor distantelor din fiecare cluster pentru diferite valori ale lui k= nr de clustere
- $WSS(within\ cluster\ squared\ distance) = \sum_{j=1}^{k} d^2(exemplul\ i, Centroid\ cluster\ j)$

  unde d= distanta dintre 2 puncte (de ex. Dist euclidiana)
- K=2,3,4,…
- Pe masura ce k creste, WSS descreste
- De la o valoare a lui k, rata de scadere lui WSS nu este semnificativa. Aceea va fi valoarea aleasa pentru k.

EXEMPLU

- Baza de date:data_1024.xls
- Atribute:
  - Driver_ID
  - Distance_Feature = Distanta parcursa de sofer in medie pe zi
  - Speeding_Feature =Durata medie pe zi, exprimata in procente, in care soferul merge cu o viteza mai mare decat viteza legala
- 4000 de exemple
- Alg: k-means

TANAGRA 1.4.50 - [K-Means 1]

File Diagram Component Window Help

Default title

- Dataset (data_1024.xls)
  - Scatterplot 1
    - Define status 1
      - K-Means 1
      - K-Means 2
      - K-Means 3
      - K-Means 4
      - K-Means 6
      - K-Means 7
      - HAC 1
        - Scatterplot 3
      - K-Means 8

R-Square	0.9076

Cluster size and WSS

Clusters	4		
Cluster	**Description**	**Size**	**WSS**
cluster n°1	c_kmeans_1	427	185.5014
cluster n°2	c_kmeans_2	104	165.5692
cluster n°3	c_kmeans_3	695	206.1267
cluster n°4	c_kmeans_4	2774	181.9562

R-Square for each attempt

Number of trials	5
Trial	**R-square**
1	0.907606
2	0.907606
3	0.907606
4	0.907606
5	0.907606

Cluster centroids

Attribute	Cluster n°1	Cluster n°2	Cluster n°3	Cluster n°4
Distance_Feature	50.404824	177.835096	180.434863	50.016637
Speeding_Feature	32.365340	70.288462	10.529496	5.204037

Use GROUP CHARACTERIZATION for detailed comparisons

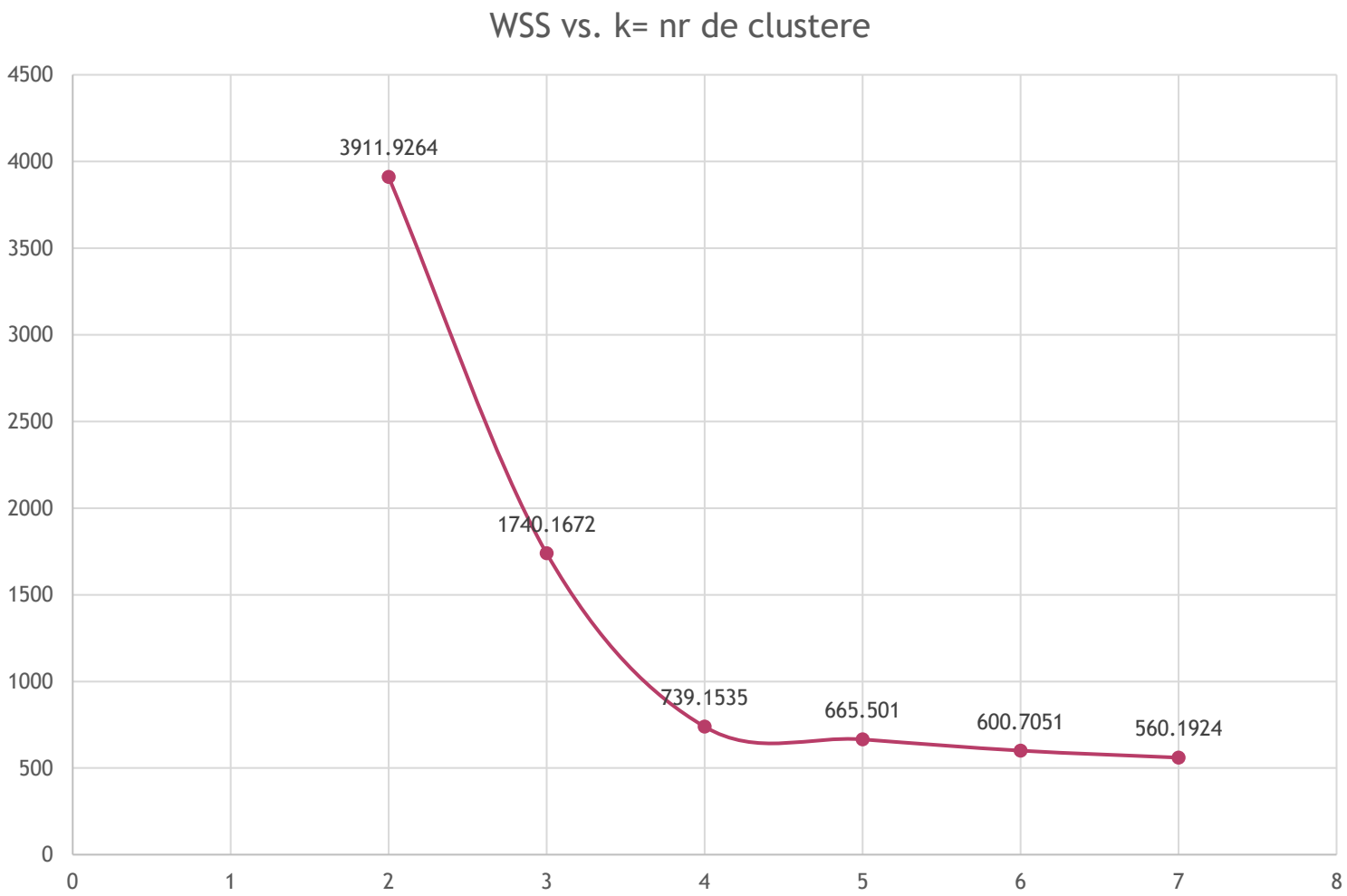
Computation time : 94 ms.
Created at 4/15/2019 4:22:20 PM

Components

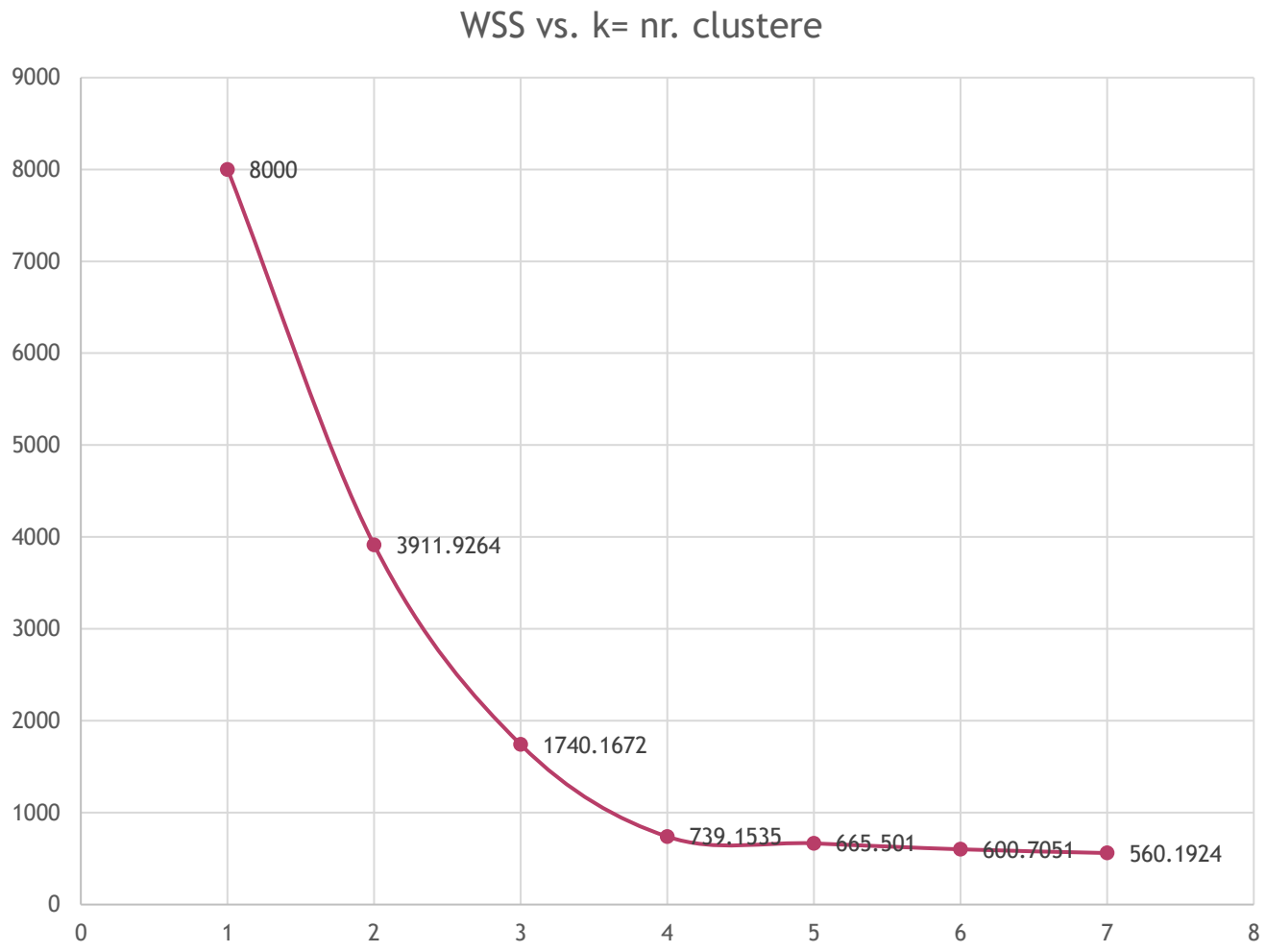
Data visualization	Statistics	Nonparametric statist	Instance selection	Feature construction	Feature selection	Regression	Factorial analysis	PLS	Clustering
Spv learning	Meta-spv learning	Spv learning assessme	Scoring	Association					

CatVARHCA
CT
CTP
EM-Clustering
EM-Selection
HAC
K-Means
K-Means Strengthening
Kohonen-SOM
LVQ
Neighborhood Graph
VARCLUS
VARHCA
VARKMeans

K=4


WSS vs. k= nr de clustere
4500
4000
3500
3000
2500
2000
1500
1000
500
0
0
1
2
3
4
5
6
7
8
3911.9264
1740.1672
739.1535
665.501
600.7051
560.1924

K=4


WSS vs. k= nr. clustere
9000
8000
7000
6000
5000
4000
3000
2000
1000
0
0
1
2
3
4
5
6
7
8
8000
3911.9264
1740.1672
739.1535
665.501
600.7051
560.1924

APLICATII

- Clusterizarea este folosita in numeroase aplicatii:
  - Analiza pietei
  - Recunoasterea formelor
  - Procesarea de imagini
- In business, clusterizarea poate ajuta specialistii in marketing sa descopere grupuri distincte de clienti cu acelasi comportament, ca de exemplu grupuri de clienti cu acelasi trend de cumparaturi, pe care sa le trateze la fel, sa le caracterizeze si sa aplice metodele lor de marketing in mod eficient.

APLICATII

- Clusterizarea poate fi folosita pentru detectarea anomaliilor (a valorilor singulare, indepartate de toate clusterele) care pot fi uneori mai interesante decat in cazurile comune.
- Detectarea anomaliilor are aplicatii in:
  - detectarea atacurilor in retelele informationale,
  - detectarea infractiunilor cu cardurile bancare

- De obicei clusterizarea poate fi un doar o etapa premergatoare aplicarii altor algoritmi si tehnici de data mining: clasificare, asociere.
- http://dataminingarticles.com/