ADATSZEGMENTÁLÁS

KLASZTERANALÍZIS

Az eseteket homogén csoportokba (ú.n. klaszterekbe) soroljuk. A csoportosítás alapja egy adott metrika szerinti közelség, illetve egy adott hasonlósági mérték szerinti hasonlóság.

DISZKRIMINANCIAALÍZIS

Az esetek egy kategóriaváltozó értékei alapján osztályokba vannak tagolva. A feladat az, hogy a többdimenziós térben az osztályokat szeparáló felületekkel elválasszuk.

OSZTÁLYOZÁS

Ismert kategóriájú esetek segítségével (tananyag) döntésfüggvényt konstruálunk, amivel ismeretlen kategóriájú esetekhez is tudunk osztályokat rendelni.

ADATSZEGMENTÁLÁS

PÉLDÁK

KLASZTERANALÍZIS

- Milyen csoportok alakíthatók ki az *employee* állományban a fizetési adatok (*salary, salbegin*) alapján?

- Milyen csoportosulások keletkeznek az országok halmazában, ha az egészségügyi helyzetet jellemzo változókat tekintjük: *lifeexpf, lifeexpm, babymort, calories, aids_rt, b_to_d*

- Milyen csoportosulások keletkeznek az országok halmazában, ha a gazdasági helyzetet jellemzo változókat tekintjük: *gdp_cap, cropgrow, urban*

ADATSZEGMENTÁLÁS

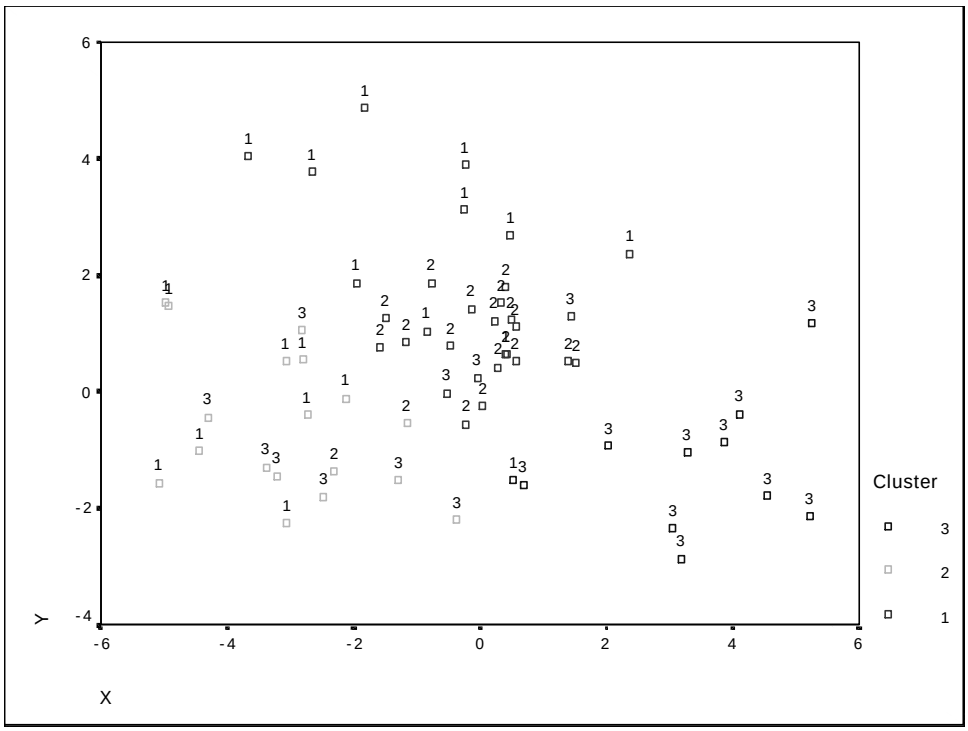
PÉLDÁK

DISZKRIMINANCIAALÍZIS

- A fizetési adatok mennyire választható szét a *jobcat*, *gender* illetve *minority* kategóriaváltozók alapján?
- Mennyire válnak szét az országok a gazdasági tömörülés (*region*) alapján?

OSZTÁLYOZÁS

- Betufelismerés
- Muholdképpontok osztályozása
- Banki rizikóelemzés: kapjon hitelt? Ne kapjon?
- Orvosi diagnosztika: Beteg? Nem beteg?
- Repülésirányítás: Felszálljon? Töröljék?



KLASZTERANALÍZIS

A *k*-közép módszer (*K-Means Cluster Analysis)***

Olyan dinamikus klaszterezo eljárás, amikor elore meg kell adni a klaszterek számát. A klaszter-középpontok térbeli helyzetét iterációban állandóan változtatjuk, amíg egy stabil állapot ki nem alakul. Az esetvektorok a legközelebbi klaszterközépponthoz lesznek rendelve.

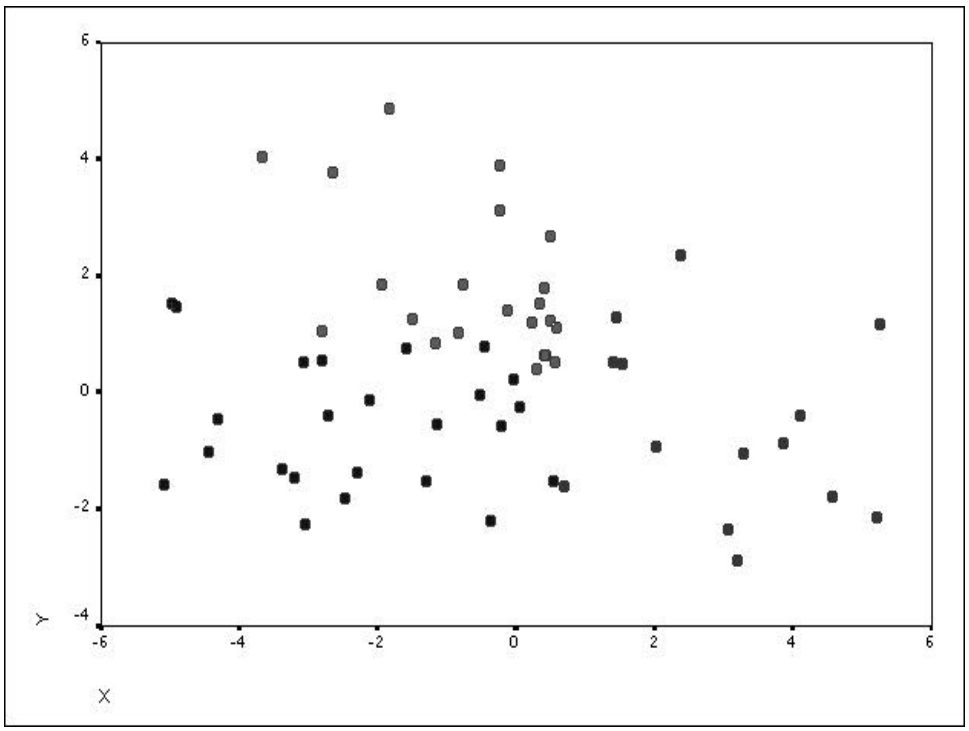
Elonye:

Nagy esetszámú adatmátrix feldolgozható vele.

Hátránya:

A metrika beépített, körülményes a koordinátasúlyozás

Elore meg kell adni a klaszterek számát



KLASZTERANALÍZIS

A hierachikus klaszterezés (*Hierarchical Cluster Analysis)*

Egyelemu klaszterekbol kiindulva, minden lépésben a két legközelebb fekvo klasztert összevonva csökkentjük a klaszterek számát, amíg minden eset egyetlen klaszterbe nem kerül. A folyamatot regisztráló dendogrammot utólag kielemezve, azt a köztes állapotot fogadjuk el, amikor az összevonás eroltetett volt, azaz az összevont klaszterek elég távol vannak egymástól.

Elonye:

Nem kell elore tudni a klaszterek számát

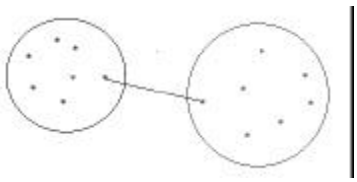
Változtatható a távolság- és hasonlósági-mérték

Hátránya:

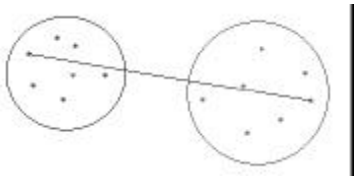
Kis dimenziószám esetén indítható el

HIERARCHIKUS KLASZTEREZÉS

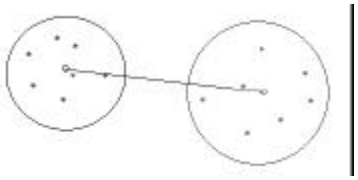
KLASZTEREK d(C_1, C_2) TÁVOLSÁGA



A legközelebbi-társ távolság



A legtávolabbi-társ távolság



Klasztercentrumok távolsága

KLASZTERANALÍZIS

ESETEK d(x, y) TÁVOLSÁGAI

Minkowsky $d(\underline{x},\underline{y}) = \left(\sum_{i=1}^{p} |x_i - y_i|^r\right)^{\frac{1}{r}}$

City-blokk $d(\underline{x},\underline{y}) = \sum_{i=1}^{p} |x_i - y_i|$

Euklideszi $d(\underline{x},\underline{y}) = \left(\sum_{i=1}^{p} |x_i - y_i|^2\right)^{\frac{1}{2}}$

Mahalanobis

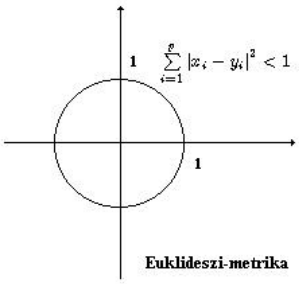
$$d(\underline{x},\underline{y}) = (\underline{x} - \underline{y})^T \underline{\underline{Q}} (\underline{x} - \underline{y}) = \sum_{i=1}^{p}\sum_{j=1}^{p} q_{ij} (x_j - y_j)(x_i - y_i)$$

Canberra $d(\underline{x},\underline{y}) = \sum_{i=1}^{p} \frac{|x_i - y_i|}{|x_i| + |y_i|}$

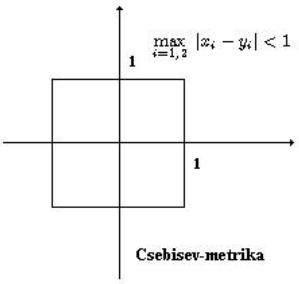
Csebisev $d(\underline{x},\underline{y}) = \max_{i=1,\ldots,p} |x_i - y_i|$

KLASZTERANALÍZIS

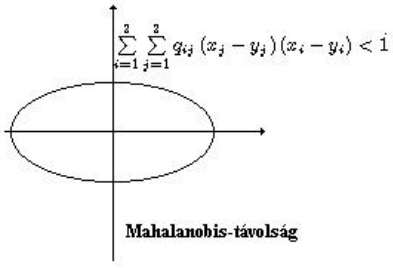
ESETEK d(x, y) TÁVOLSÁGAI



Euklideszi-metrika

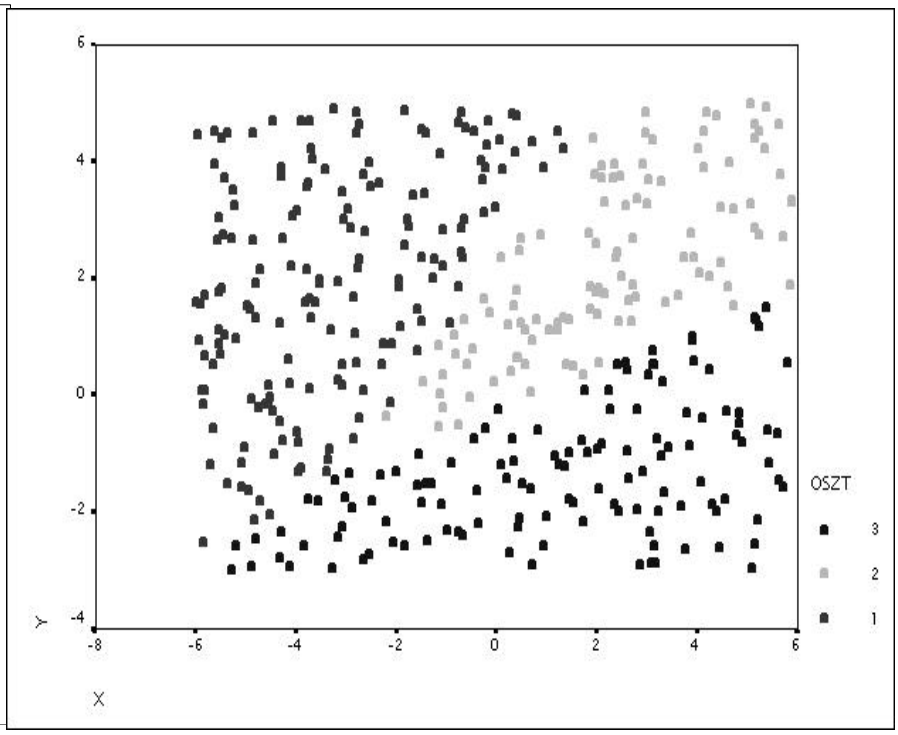


Csebisev-metrika



Mahalanobis-távolság

DISZKRIMINANCIAANALÍZIS



TÖBBDIMENZIÓS SKÁLÁZÁS

Egy p-dimenziós sokaságot lehet egy k=1,2 vagy 3 dimenziós Euklideszi ponthalmazzal vizualizálni.
A ponthalmaz távolságviszonyai az eredeti sokaság eseteinek távolságviszonyaival nagymértékben egyezik.

A vizualizálás révén tanulmányozható a statisztikai sokaság térbeli struktúrálódása. Jellegzetes tömörülések, irányok fedezhetok fel az elkészült scatter-grafikonon.

$$\underline{\underline{X}} = \begin{pmatrix} x_1^1 & x_1^2 & \cdots & x_1^p \\ x_2^1 & x_2^2 & \cdots & x_2^p \\ \vdots & \vdots & \ddots & \vdots \\ x_n^1 & x_n^2 & \cdots & x_n^p \end{pmatrix}$$

$\underline{x}^{(1)}$ 1. eset

$\underline{x}^{(2)}$ 2. eset

.

$\underline{x}^{(n)}$ n. eset

TÖBBDIMENZIÓS SKÁLÁZÁS

$\underline{\underline{D}} = (d(\underline{x}^{(i)}, \underline{x}^{(j)}))$ **Az esetvektorok egymástól vett *nxn*-es távolságmátrixa**

Megkonstruálhatók olyan $\underline{z}_1, \underline{z}_2, \ldots, \underline{z}_n$ ***k*=1,2 vagy 3 dimenziós vektorok, melyek *nxn*-es Euklideszi távolság-mátrixa nagymértékben hasonló** $\underline{\underline{D}}$ **-hez.**

$\sum_{i=1}^{n}\sum_{j=1}^{n}\left(d(\underline{x}^{(i)}, \underline{x}^{(j)}) - \rho(\underline{z}_i, \underline{z}_j\right)^2$ **„kicsi”**