

Adatbányászat: Adatok

2. fejezet

Tan, Steinbach, Kumar
Bevezetés az adatbányászatba
előadás-fóliák
fordította
Ispány Márton

Az adatelemzés csővezetéke

- Az adatbányászat nem az egyetlen lépés a folyamatban



- **Előfeldolgozás:** a valós adatok zajosak, hiányosak és inkonzisztensek. **Adattisztítás** is szükséges az adatok megértéshez
 - Módszerek: Mintavétel, Dimenzió csökkentés, Jellemző szelektálás.
 - Pizkos munka de gyakran a legfontosabb lépés az elemzésben.
 - **Utófeldolgozás:** Tegyük az eredményeket hasznossá és cselekvésre ösztönzővé a felhasználó számára
 - A kapott eredmény fontosságának statisztikai vizsgálata
 - Vizualizáció.
- Az elő- és utófeldolgozás gyakran maga is egy adatbányászati feladat

Az adatbányászat módszertana

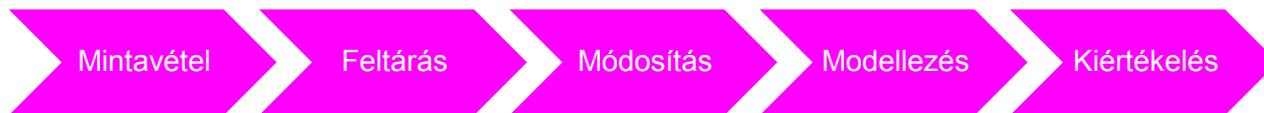
- Többféle (gyártótól is függő) módszertan:

CRISP-DM (SPSS-Clementine) <http://www.crisp-dm.org/>

SEMMA (SAS) <http://www.sas.com/>

- Az 5 lépcsős folyamat

- Mintavétel: az adatok előkészítése az adattárházból.
- Feltárás: új összefüggések, mintázatok keresése.
- Módosítás: attribútumok, rekordok, mezők módosítása, kitöltése.
- Modellezés: analitikus modellek illesztése.
- Kiértékelés: a modell(ek) jóságának, hasznosságának mérése.



Mit értünk adat alatt?

- Objektumok attribútumainak numerikusan jellemzett összessége.

- Attribútum: egy objektum tulajdonsága, jellemzője.
 - Példák: hajszín, hőmérséklet, stb.
 - Az attribútumot nevezik változónak, jellemzőnek (feature).

Objektumok

- Attribútumok értékeinek egy összessége ír le egy objektumot.
 - Az objektumot nevezik rekordnak, pontnak, esetnek, mintaelemnek, egyednek, entitásnak.

Attribútumok

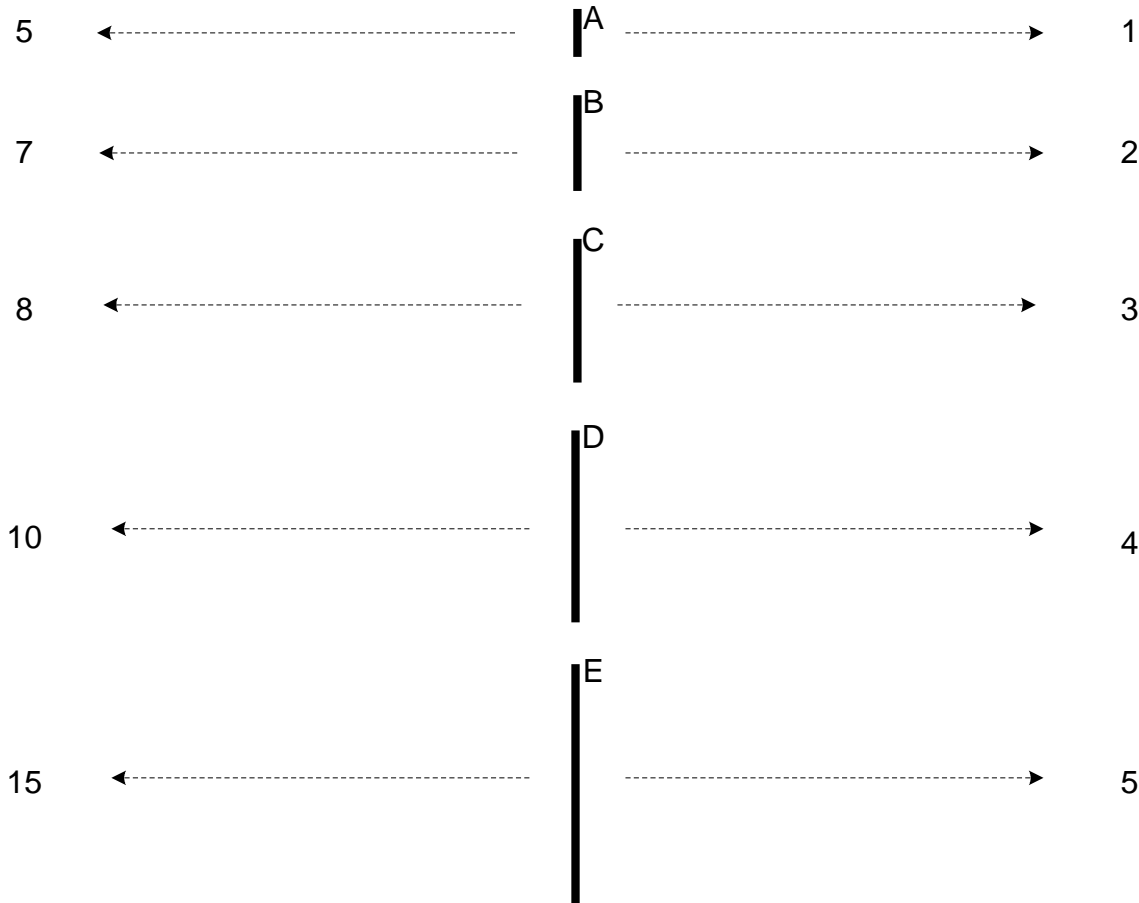
<i>Tid</i>	Vissza- térítés	Családi állapot	Jöve- delem	Csalás
1	Igen	Nőtlen	125K	Nem
2	Nem	Házass	100K	Nem
3	Nem	Nőtlen	70K	Nem
4	Igen	Házass	120K	Nem
5	Nem	Elvált	95K	Igen
6	Nem	Házass	60K	Nem
7	Igen	Elvált	220K	Nem
8	Nem	Nőtlen	85K	Igen
9	Nem	Házass	75K	Nem
10	Nem	Nőtlen	90K	Igen

Attribútum értékek

- Attribútum értékek alatt az attribútumhoz rendelt számokat vagy szimbólumokat értjük.
- Különbség az attribútumok és az attribútum értékek között:
 - Ugyanazt az attribútumot attribútum értékek különböző tartományaira képezhetjük le.
 - ◆ Példa: a magasságot méterben és lábban is mérhetjük.
 - Különböző attribútumokat attribútum értékek ugyanazon tartományára is le képezhetjük.
 - ◆ Példa: az ID és KOR változók attribútum értékei egészek.
 - ◆ Azonban az attribútum értékek tulajdonságai különfélék lehetnek:
 - ID-re nincs korlát, a KOR-nak van maximuma és minimuma.

Hosszúság mérése

- A mód, ahogy egy attribútumot mérünk részben eltérhet az attribútum tulajdonságaitól.



Attribútumok típusai

- A következő attribútum típusokat különböztetjük meg
 - **Névleges (nominális)**
 - ◆ Példák: ID, szemszín, irányítószám.
 - **Sorrendi (ordinális)**
 - ◆ Példák: rangsorolás (pl. a burgonyaszírom íze egy 1-10 skálán), fokozat, magasság mint {magas, átlagos, alacsony}.
 - **Intervallum**
 - ◆ Példák: dátum, hőmérséklet Celsiusban vagy Fahrenheitben.
 - **Hányados**
 - ◆ Példák: abszolút hőmérséklet (Kelvin), hosszúság, idő.

Attribútum értékek tulajdonságai

- Egy attribútum típusa attól függ, hogy milyen tulajdonságokkal rendelkeznek.
 - Egyezőség, különbözőség: = ≠
 - Rendezés: < >
 - Összeadás, kivonás: + -
 - Szorzás, osztás: * /

 - Névleges attribútum: egyezőség
 - Sorrendi attribútum: egyezőség és rendezés
 - Intervallum attribútum: egyezőség, rendezés és összeadás
 - Hányados attribútum: mind a 4 tulajdonság

Attribútum értékek tulajdonságai

Attribútum típusa	Leírás	Példák	Műveletek
Névleges (nominális)	Egy névleges attribútum értékei csak különböző nevek, azaz csak ahhoz nyújt elegendő információt, hogy egy objektumot megkülönböztessünk egy másiktól. (=, ≠)	irányítószám, dolgozó azonosító, szemszín, nem: { <i>férfi</i> , <i>nő</i> }	módusz, entropia, kontingencia korreláció, χ^2 érték
Sorrendi (ordinális)	Egy rendezett attribútum értékei ahhoz nyújtanak elegendő információt, hogy rendezzük az objektumokat. (<, >)	ásványok keménysége { <i>jó</i> , <i>jobb</i> , <i>legjobb</i> }, fokozat, házszám	medián, percentilis, rang korreláció, széria próba, előjel ill. előjeles rangösszeg próba
Intervallum	Egy intervallum attribútumnál az értékek közötti különbségek is jelentéssel bírnak. (+, -)	naptári dátumok, hőmérséklet Celsiusban ill. Fahrenheitben	átlag, szórás, Pearson féle korreláció, <i>t</i> és <i>F</i> próba
Hányados	Hányados változónál a különbségnek és a hányadosnak egyaránt van értelme. (*, /)	abszolút hőmérséklet, pénzügyi mennyiség, kor, tömeg, hossz, elektromos áram	mértani és harmónikus közép, százalék variáció

Attribútum értékek tulajdonságai

Attribútum szintje	Transzformáció	Megjegyzés
Névleges (nominális)	Az értékek bármilyen permutációja	Okoz-e bármilyen különbséget ha az alkalmazottak azonosítóit átrendezzük?
Sorrendi (ordinális)	Az értékek rendezés tartó transzformációja, azaz $új_érték = f(régi_érték)$, ahol f egy monoton függvény.	Egy attribútum melyet a jó, jobb és legjobb fokokkal írhatunk le egyaránt reprezentálható az $\{1, 2, 3\}$ vagy a $\{0.5, 1, 10\}$ számokkal.
Intervallum	$új_érték = a * régi_érték + b$ ahol a és b konstansok	Így a Fahrenheit és Celsius skálák abban különböznek hogy hol van a zéró érték és mekkora az egység (fok).
Hányados	$új_érték = a * régi_érték$	A hosszúság méterben és lábban is mérhető.

Diszkrét és folytonos attribútumok

- Diszkrét attribútumok

- Véges vagy megszámlálható végtelen sok értéke lehet.
- Példák: irányítószám, darabszám, szavak száma dokumentumokban.
- Gyakran egész értékű változókkal reprezentáljuk.
- Megjegyzés: a bináris attribútumok a diszkrét attribútumok egy speciális esete.

- Folytonos attribútumok

- Az attribútum értékek valós számok.
- Példák: hőmérséklet, magasság, súly.
- Gyakorlatban a valós értékek csak véges sok tizedesjegyig mérhetőek és ábrázolhatóak.
- A folytonos attribútumokat általában lebegőpontos változókkal reprezentáljuk.

Adatállományok típusai

- Rekord

- Adatmátrix (adatbázisok)
- Dokumentum mátrix (szövegbányászat)
- Tranzakciós adatok

- Gráf

- World Wide Web (webgráf)
- Molekula szerkezetek

- Rendezett

- Térbeli adatok
- Időbeli adatok
- Szekvenciális adatok
- Génszekvenciák adatai

Strukturált adatok fontos jellemzői

- Dimenzió
 - ◆ Dimenzió probléma

- Ritkaság
 - ◆ Csak az előforduló esetek elemezhetőek

- Felbontás
 - ◆ A mintázat függ a skálától

Rekordokból álló adatok

- Olyan adatok, amelyek rekordok egy halmazából állnak, ahol mindegyik rekord attribútum értékek egy adott halmazából áll.

<i>Tid</i>	Vissza- térítés	Családi állapot	Jöve- delem	Csalás
1	Igen	Nőtlen	125K	Nem
2	Nem	Házias	100K	Nem
3	Nem	Nőtlen	70K	Nem
4	Igen	Házias	120K	Nem
5	Nem	Elvált	95K	Igen
6	Nem	Házias	60K	Nem
7	Igen	Elvált	220K	Nem
8	Nem	Nőtlen	85K	Igen
9	Nem	Házias	75K	Nem
10	Nem	Nőtlen	90K	Igen

Adatmátrix

- Ha az objektumokat leíró adatok numerikus attribútumok egy adott halmazából állnak, akkor gondolhatunk rájuk úgy, mint pontokra a többdimenziós térben, ahol minden egyes dimenzió egy attribútumot reprezentál.
- Az ilyen adatokat egy $n \times p$ –es mátrixszal reprezentálhatjuk, amelynek n sora az objektumoknak, p oszlopa pedig az attribútumoknak felel meg.

X vetület	Y vetület	Távolság	Súly	Vastagság
10.23	5.27	15.22	2.7	1.2
12.65	6.25	16.22	2.2	1.1

Dokumentum mátrix

- Minden dokumentumot kifejezések egy vektorával írunk le.
 - Minden kifejezés egy attribútuma a vektornak.
 - Minden attribútum érték annak a száma, hogy az attribútumhoz tartozó kifejezés hányszor fordul elő a dokumentumban.

	csapat	edző	meccs	labda	pont	játék	győzelem	vereség	szezon
1. Doc	3	0	5	0	2	6	0	2	2
2. Doc	0	7	0	2	1	0	0	3	0
3. Doc	0	1	0	0	1	2	2	0	0

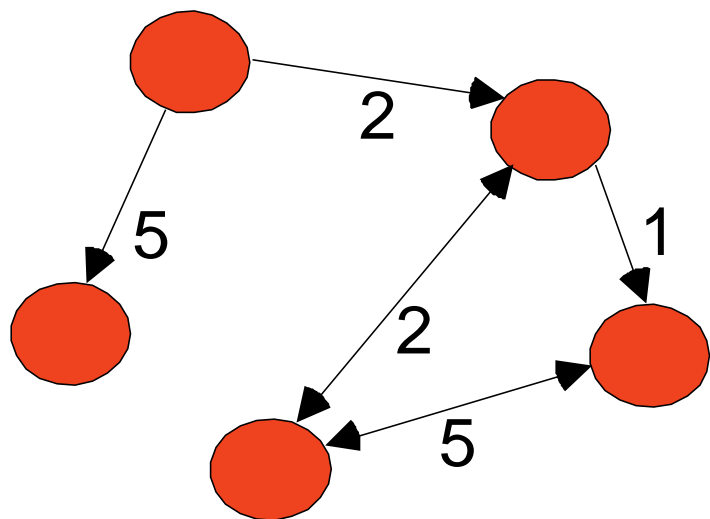
Tranzakciós adatok

- Speciális rekord típusú adatok, ahol
 - minden rekord (tranzakció) tételek egy halmazát tartalmazza.
 - Pl.: tekintsünk egy élelmiszerboltot. A tranzakció azon árucikkekből áll, amelyeket a vásárló vesz egy vásárlás során, míg a tételek a vásárolt árucikkek.

<i>TID</i>	<i>Tételek</i>
1	Kenyér, Kóla, Tej
2	Sör, Kenyér
3	Sör, Kóla, Pelenka, Tej
4	Sör, Kenyér, Pelenka, Tej
5	Kóla, Pelenka, Tej

Gráf adatok

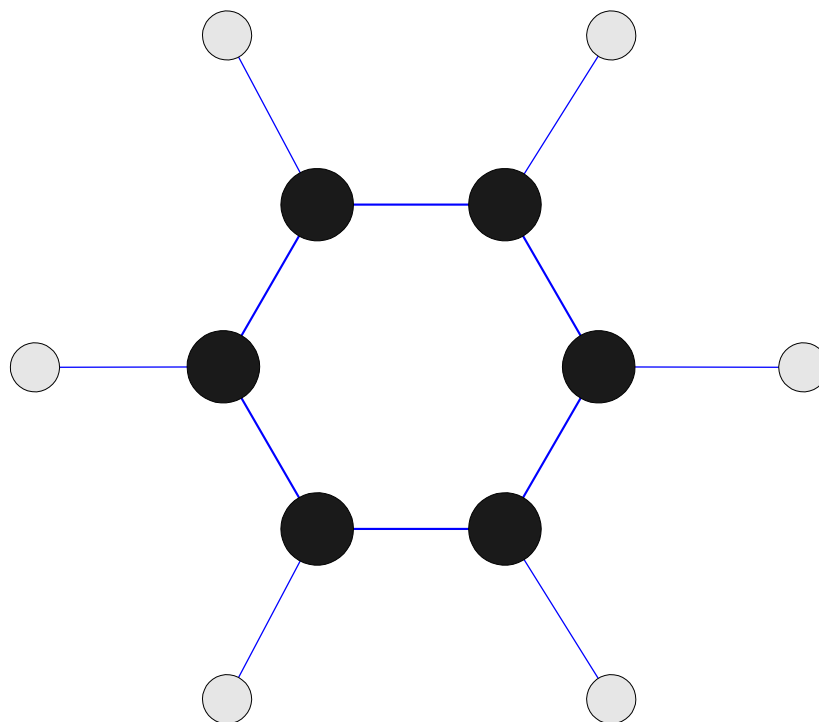
- Példák: általános gráf, HTML linkek



```
<a href="papers/papers.html#bbbb">  
Data Mining </a>  
<li>  
<a href="papers/papers.html#aaaa">  
Graph Partitioning </a>  
<li>  
<a href="papers/papers.html#aaaa">  
Parallel Solution of Sparse Linear System of Equations </a>  
<li>  
<a href="papers/papers.html#ffff">  
N-Body Computation and Dense Linear System Solvers
```

Kémiai adatok

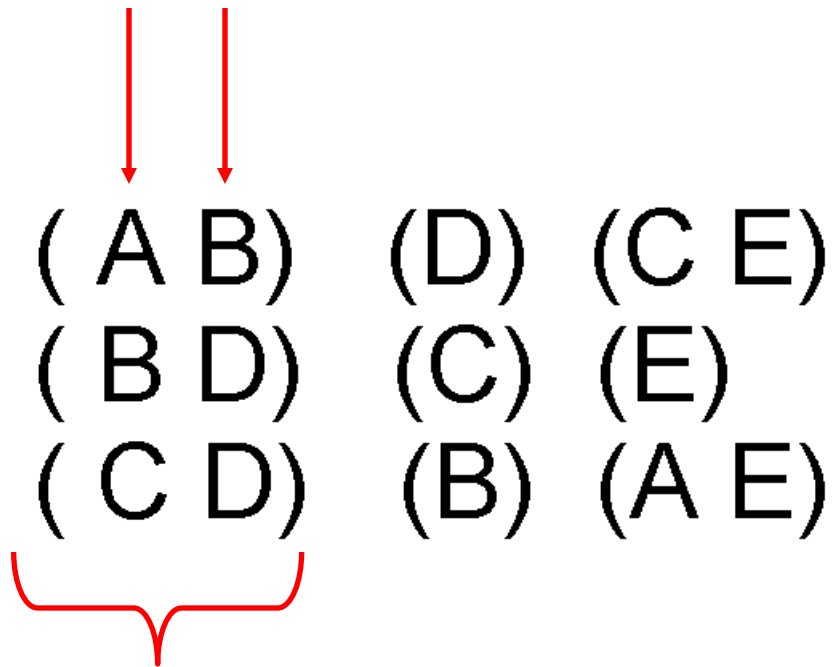
- Benzin molekula: C_6H_6



Rendezett adatok

- Tranzakciók sorozatai

Tételek/Események



A sorozat egy
eleme

Rendezett adatok

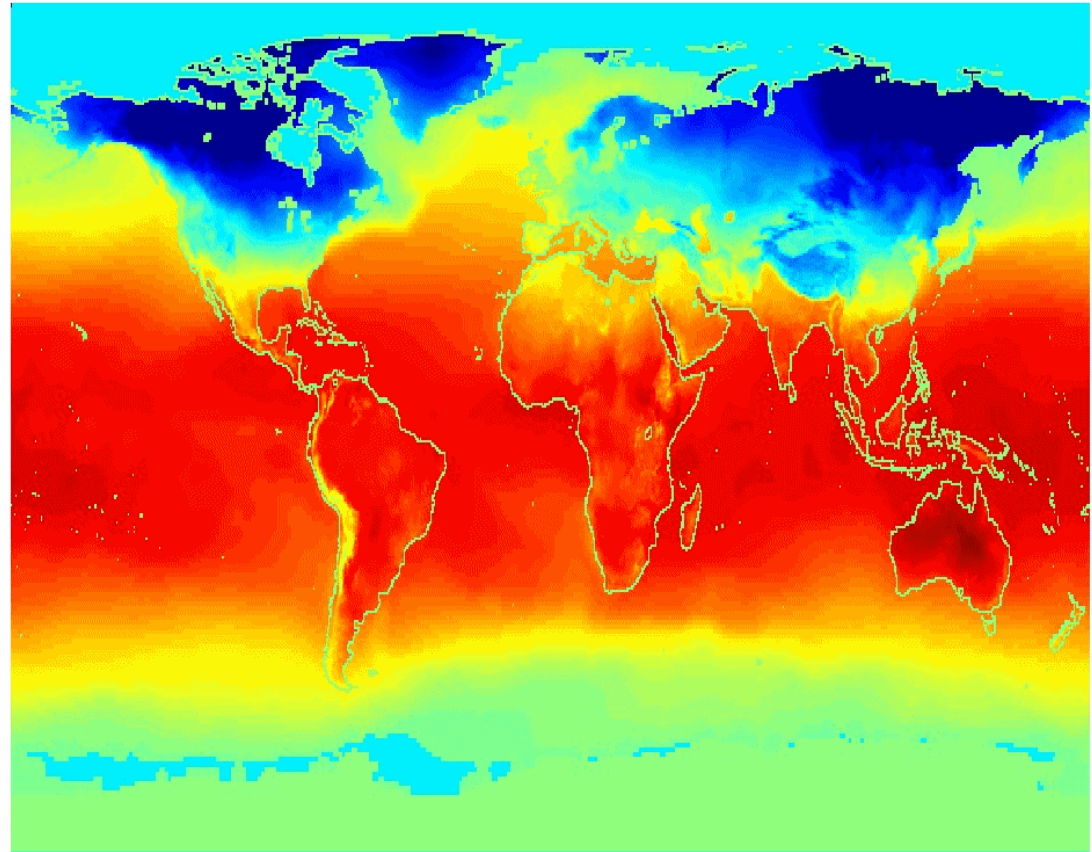
- Génszekvenciák

```
GGTTC CGCCTTCAGCCCCGCGCC  
CGCAGGGCCCGCCCCGCGCCGTC  
GAGAAGGGCCCGCCTGGCGGGCG  
GGGGGAGGCGGGGCCGCCCGAGC  
CCAACCGAGTCCGACCAGGTGCC  
CCCTCTGCTCGGCCTAGACCTGA  
GCTCATTAGGCGGCAGCGGACAG  
GCCAAGTAGAACACGCGAAGCGC  
TGGGCTGCCTGCTGCGACCAGGG
```

Rendezett adatok

- Tér és időbeli adatok

Jan



A földrészek és óceánok átlagos havi középhőmérséklete

Adatminőség

- Milyen adatminőségi problémák léphetnek fel?
- Hogyan ismerhetjük fel ezeket a problémákat az adatainkon?
- Hogyan kezelhetjük ezeket a problémákat?

- Példák adatminőségi problémákra:
 - zaj (hiba) és kiugró adatok
 - hiányzó adatok
 - duplikált adatok

Adatminőség

- Példák adatminőségi problémákra:
 - zaj (hiba) és kiugró adatok
 - hiányzó adatok
 - duplikált adatok

Hiba vagy milliomos?

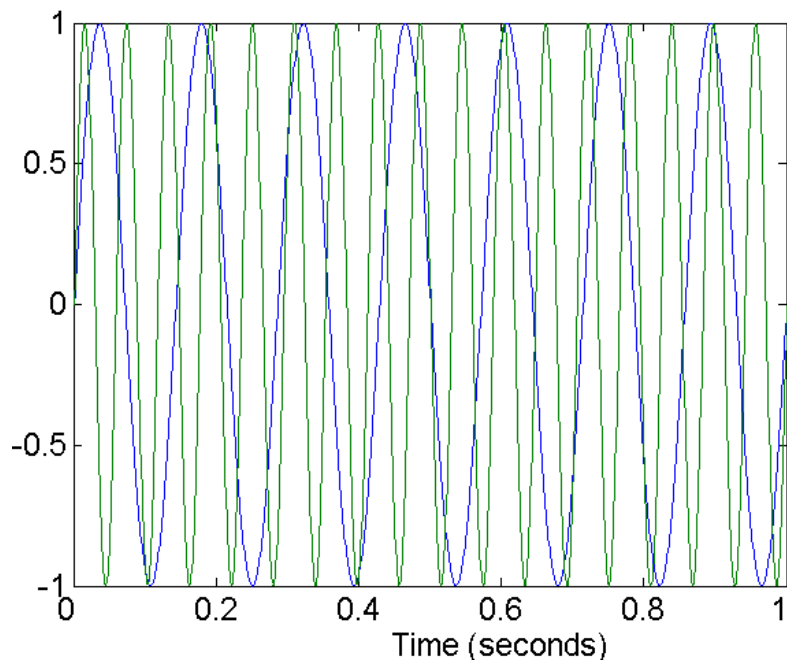
Hiányzó érték (NULL)

Inkonzisztens duplikátumok

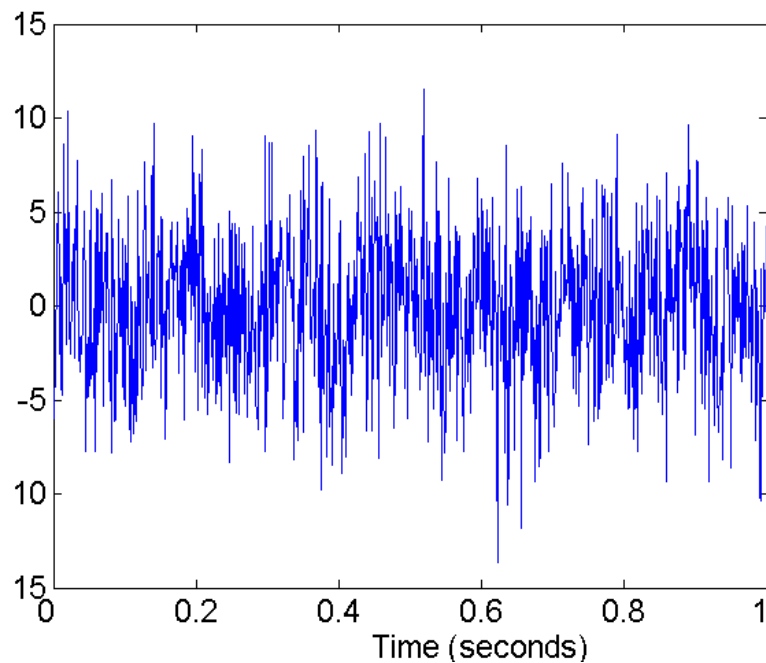
<i>Tid</i>	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	10000K	Yes
6	No	NULL	60K	No
7	Yes	Divorced	220K	NULL
8	No	Single	85K	Yes
9	No	Married	90K	No
9	No	Single	90K	No

Zajos adatok

- Zaj alatt az eredeti (igazi) érték módosulását értjük
 - Példák: az emberi hang torzulása ha rossz telefonon beszélünk, szemcsésedés a képernyőn.



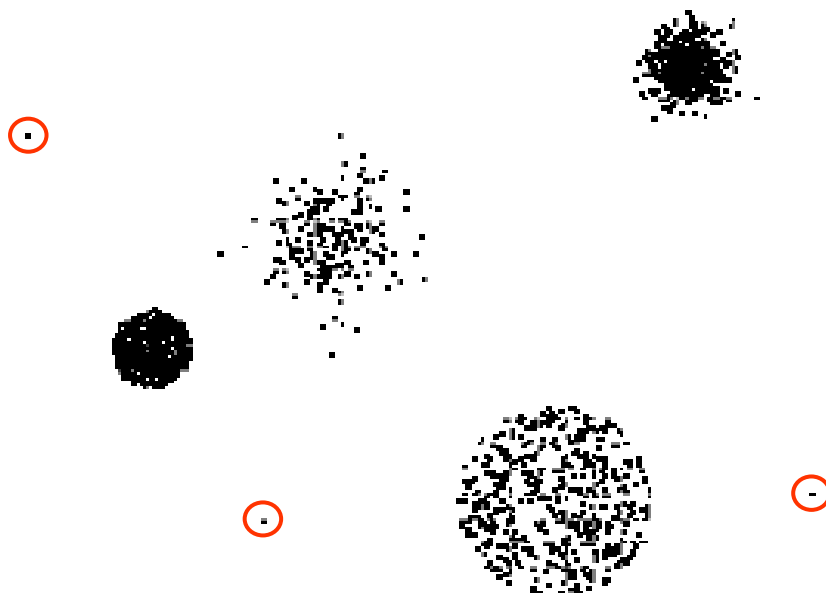
Két szinusz hullám



Két szinusz hullám + Zaj

Kiugró adatok

- A kiugró adatok olyan objektumok adatai, amelyek jellemzői jelentősen eltérnek az adatállományban lévő más objektumok adataitól.



Hiányzó adatok

- Hiányzó adatok okai:
- Az információt nem gyűjtöttük össze (pl. az emberek visszautasították a koruk és súlyuk megadását).
 - Egyes attribútumok nem alkalmazhatóak minden esetben (pl. a gyerekeknek nincs jövedelme).
- Hiányzó adatok kezelése:
 - Objektumok (rekordok) törlése.
 - Hiányzó adatok becslése.
 - A hiányzó értékek figyelmen kívül hagyása az elemzésnél.
 - Helyettesítés az összes lehetséges értékkel (a valószínűségek alapján).

Duplikált adatok

- Az adatállomány tartalmazhat olyan rekordokat, amelyek más rekordok pontos ill. kevésbé pontos ismétlődései.
 - Főként akkor merül fel ha heterogén forrásokból egyesítjük az adatokat.
- Példa:
 - Ugyanaz az ember többféle e-mail vagy lakcímmel.
- Adattisztítás
 - Az a folyamat, mely során az ismétlődő adatokat kezeljük.

Adatok előfeldolgozása

- Aggregálás
- Mintavétel
- Dimenzió csökkentés
- Jellemzők (features) részhalmazainak szelekciója
- Új jellemzők, attribútumok létrehozása
- Diszkretizáció és binarizálás
- Attribútum transzformáció

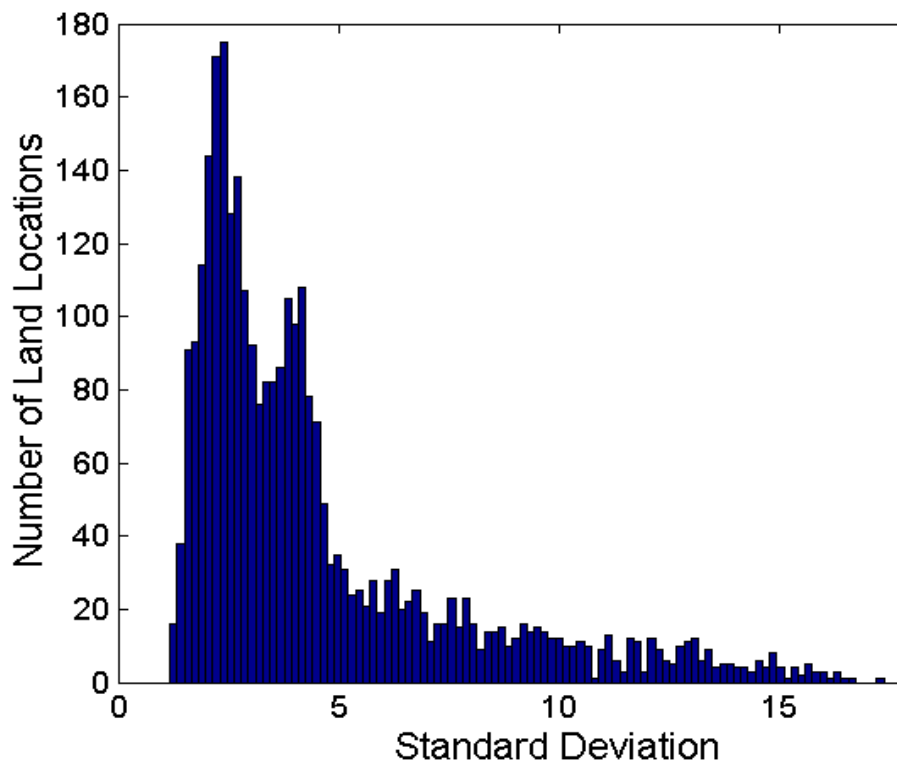
Aggregálás

- Kettő vagy több attribútum (objektum) kombinálása egy attribútummá (objektummá).

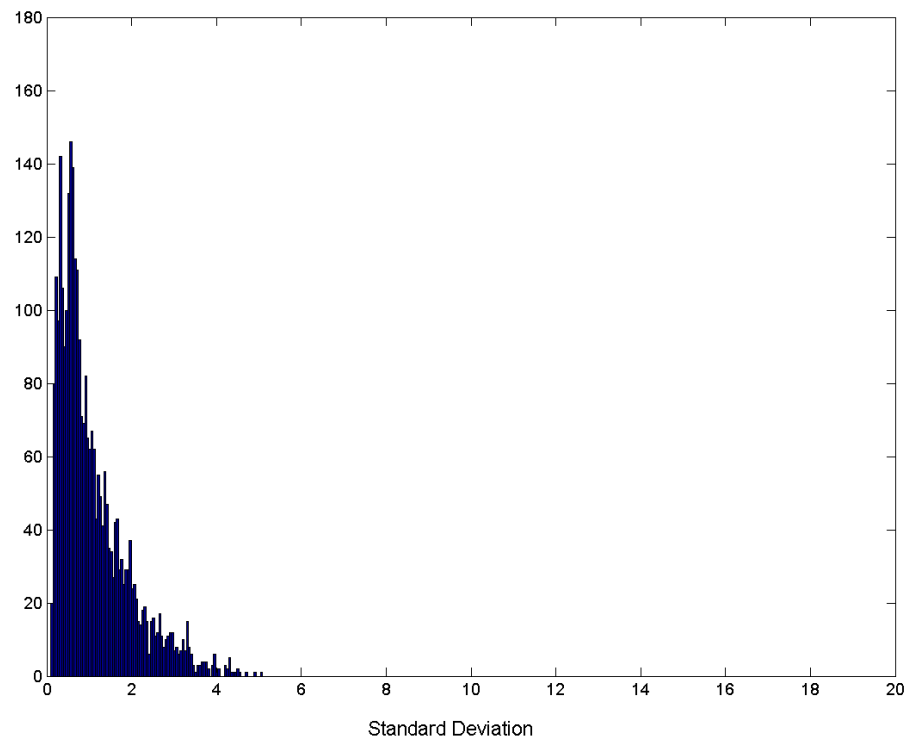
- Cél:
 - Adatcsökkentés
 - ◆ Csökkentsük az attribútumok vagy az objektumok számát.
 - A skála megváltoztatása
 - ◆ A városokat régiókba, megyékbe, országokba fogjuk össze.
 - Az adatok stabilitásának növelése
 - ◆ Az aggregált adatok ingadozása csökken (simítás).

Aggregálás

A csapadék szórása Ausztráliában



Havi átlagos csapadék szórása



Évi átlagos csapadék szórása

Mintavétel

- Az adatszelekció fő módszere
 - Egyaránt használatos az adatok előzetes vizsgálatánál és a végső adatelemzésnél.
- A statisztikusok azért használnak mintavételezést mivel a teljes populáció **megfigyelése** túl drága vagy túl időigényes.
- Az adatbányászok azért használnak mintavételezést mivel a teljes adatállomány (adat-tárház) **feldolgoása** túl drága vagy túl időigényes.

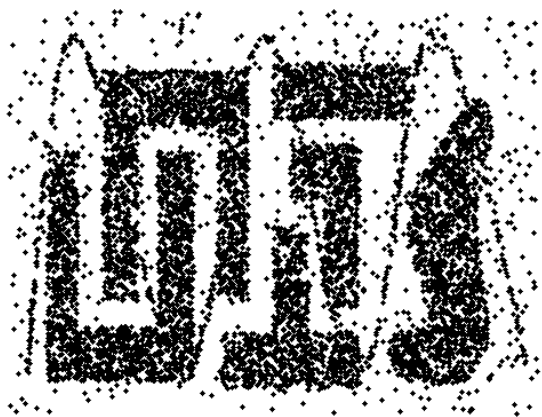
Mintavétel

- A hatékony mintavétel alapelve:
 - A mintával ugyanolyan jól tudunk dolgozni mint a teljes adatállománnyal, amennyiben a minta reprezentatív.
 - A minta akkor reprezentatív ha a számunkra fontos tulajdonságok szempontjából ugyanúgy viselkedik mint a teljes adatállomány.

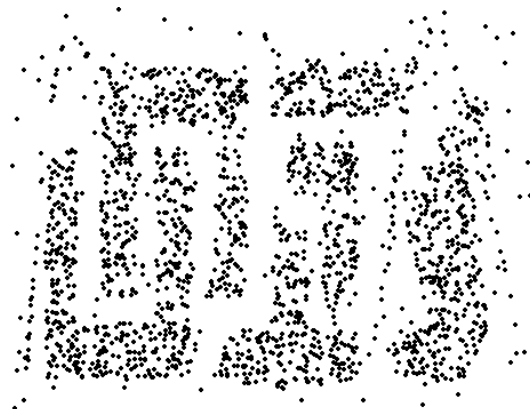
Mintavételi módok

- Egyszerű véletlen minta
 - Ugyanakkora valószínűséggel választunk ki minden objektumot.
- Visszatevés nélküli mintavétel
 - Ha egy objektumot már kiválasztottunk, akkor azt töröljük az adatállományból.
- Visszatevéses mintavétel
 - Az objektumot nem töröljük az adatállományból akkor sem ha a mintavétel kiválasztotta.
 - ◆ Ekkor egy objektumot többször is kiválaszthatunk.
- Rétegzett mintavétel
 - Osszuk fel az adatállományt részekre, majd vegyünk véletlen mintákat minden részből.

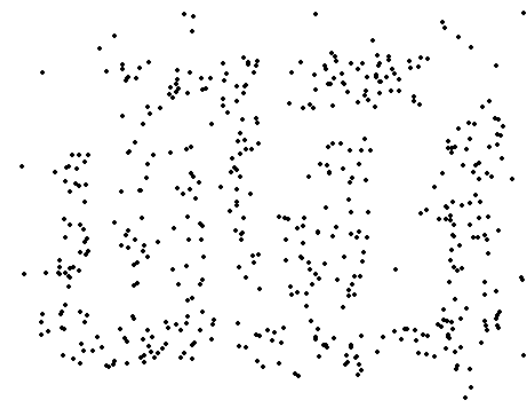
Mintanagyság



8000 pont



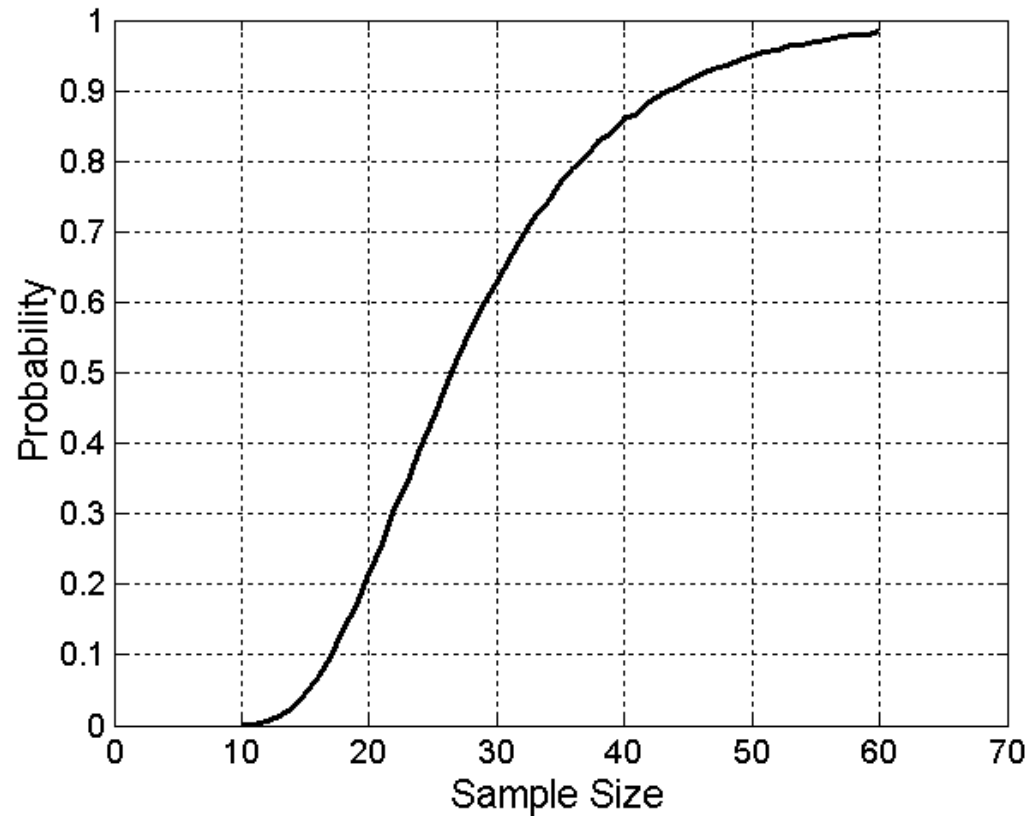
2000 pont



500 pont

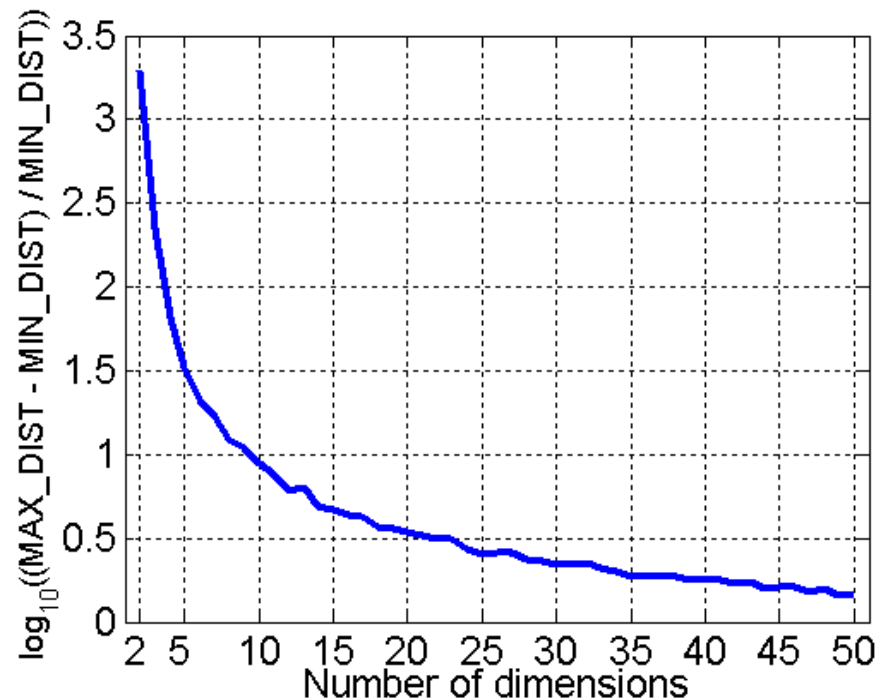
Mintanagyság

- Mekkora mintanagyság szükséges, hogy 10 csoport mindegyikéből kiválasszunk legalább egy objektumot?



Dimenzió probléma

- Amikor a dimenzió nő a rekordok (pontok) egyre ritkábbak lesznek a térben, ahol elhelyezkednek.
- A rekordok (pontok) közötti távolság és sűrűség, melyek alapvetőek csoportosításnál és kiugró adatok meghatározásánál, fontossága csökken.



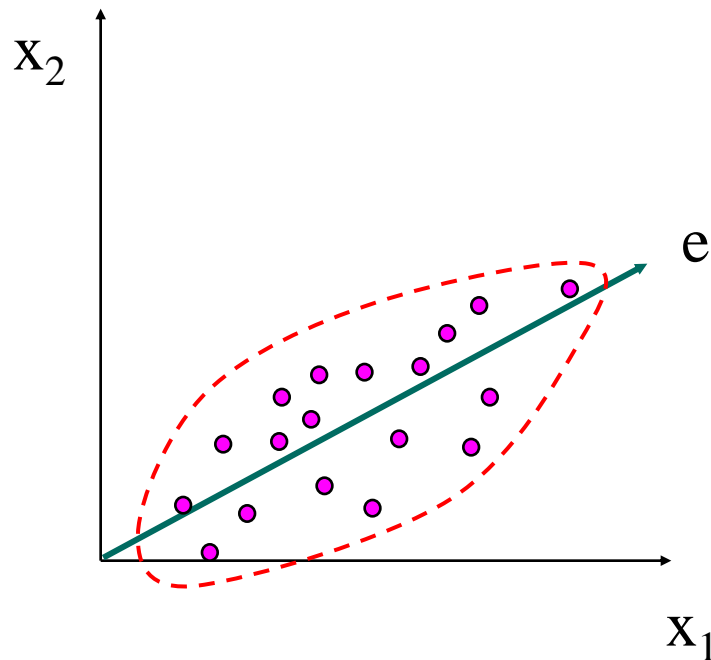
- **Generáljunk 500 véletlen pontot**
- **Számítsuk ki az összes pontpár közötti távolság maximuma és minimuma különbségét**

Dimenzió csökkentés

- Cél:
 - Elkerülni a dimenzió problémát.
 - Csökkenteni az adatbányászati algoritmusokhoz szükséges időt és memóriát.
 - Segíteni az adatok könnyebb megjelenítését.
 - Segíteni a hiba csökkentését és a lényegtelen jellemzők meghatározását majd elhagyását.
- Módszerek
 - Főkomponens analízis (PCA)
 - Szinguláris felbontás (SVD)
 - Egyéb felügyelt és nemlineáris módszerek, pl. többdimenziós skálázás (MDS)

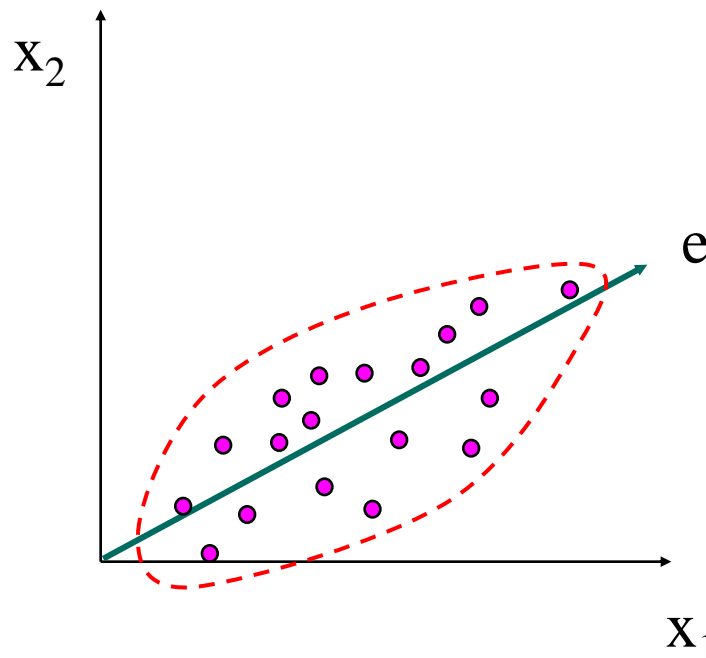
Dimenzió csökkentés: PCA

- Célja olyan vetítés (projekció) meghatározása, amely leginkább megőrzi az adatokban lévő variációt, sokszínűséget.



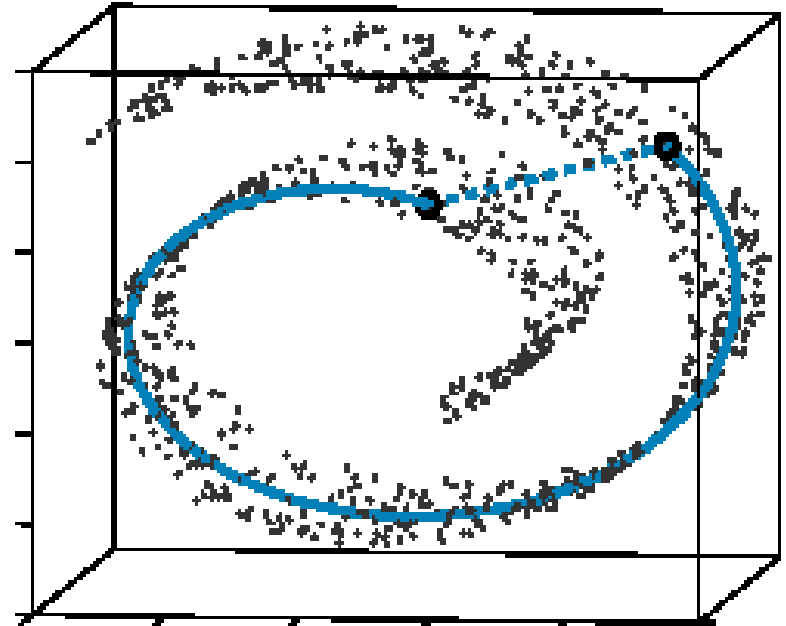
Dimenzió csökkentés: PCA

- Határozzuk meg a kovariancia mátrix sajátvektorait.
- Az új teret (koordinátatengelyeit) ezek a sajátvektorok határozzák meg.



Dimenzió csökkentés: ISOMAP

Tenenbaum, de Silva,
Langford (2000) Science



- Állítsuk elő a szomszédsági gráfot.
- A gráf minden pontpárára számoljuk ki a legrövidebb út hosszát – geodetikus távolság.
- Erre a távolság mátrixra alkalmazzuk az MDSt.

Dimenzió csökkentés: PCA

Dimensions = 206



Jellemzők részhalmazainak szelekciója

- A dimenzió csökkentés egy másik útja.
- Felesleges jellemzők
 - Egy vagy több attribútum által hordozott információt részben vagy teljesen megismétel.
 - Példa: egy termék vételára és az utána fizetendő adó.
- Lényegtelen jellemzők
 - Nem tartalmazznak az aktuális adatbányászati feladat számára hasznos információt.
 - Példa: a hallgató NEPTUN kódja többnyire nem befolyásolja a tanulmányi eredményt.

Jellemzők részhalmazainak szelekciója

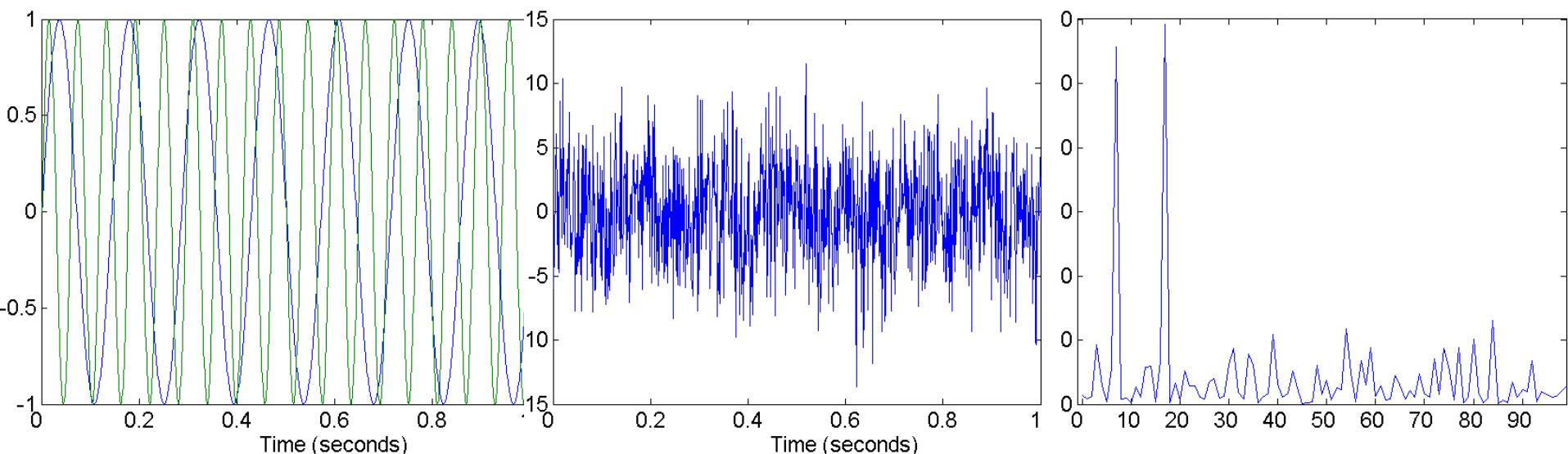
- Módszerek:
 - Nyers erő (brute force) megközelítés
 - ◆ Próbáljuk ki a jellemzők összes részhalmazát az adatbányászati algoritmus inputjaként.
 - Beágyazott megközelítés
 - ◆ A jellemzők szelekciója az adatbányászati feladat szerves részét alkotja.
 - Szűrő megközelítés
 - ◆ A jellemzőket az adatbányászati algoritmus futása előtt szelektáljuk.
 - Borító (wrapper) megközelítés
 - ◆ Az adatbányászati algoritmust fekete dobozként használjuk a legjobb attribútum részhalmaz megtalálására.

Új jellemzők (attribútumok) létrehozása

- Olyan új attribútumok létrehozása, amelyek az adatállományban lévő lényeges információkat használhatóbb formában tartalmazzák mint az eredeti attribútumok.
- Három általános módszer
 - Jellemző kinyerés (feature extraction)
 - ◆ terület függő (pl. képfeldolgozás, földrajz)
 - Új térre való leképezés
 - Jellemző szerkesztés
 - ◆ jellemzők kombinálása

Új térre való leképezés

- Fourier transzformáció
- Wavelet (hullám) transzformáció



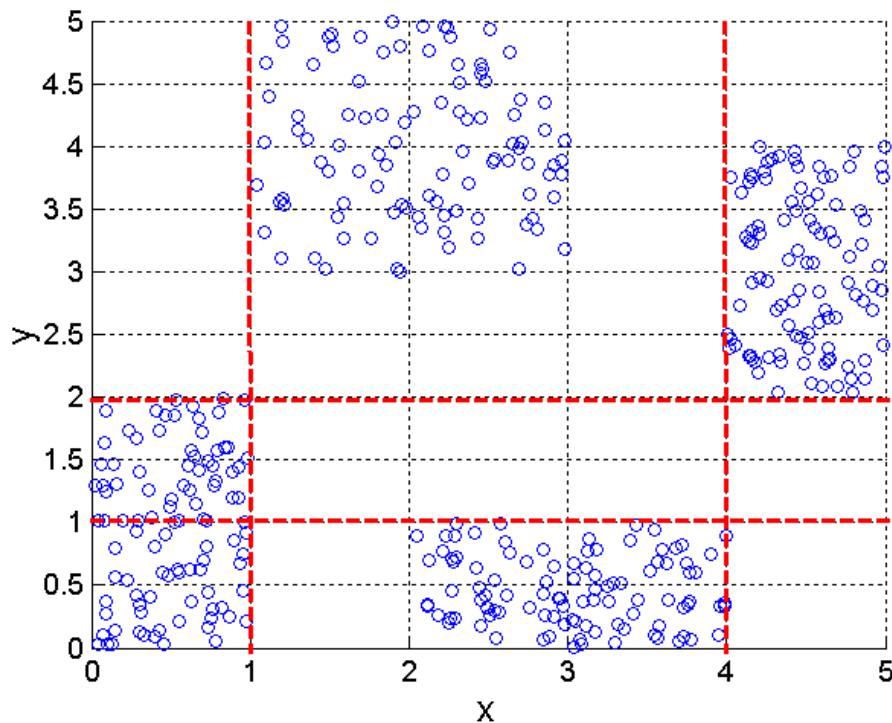
Két szinusz hullám

Két szinusz hullám + Zaj

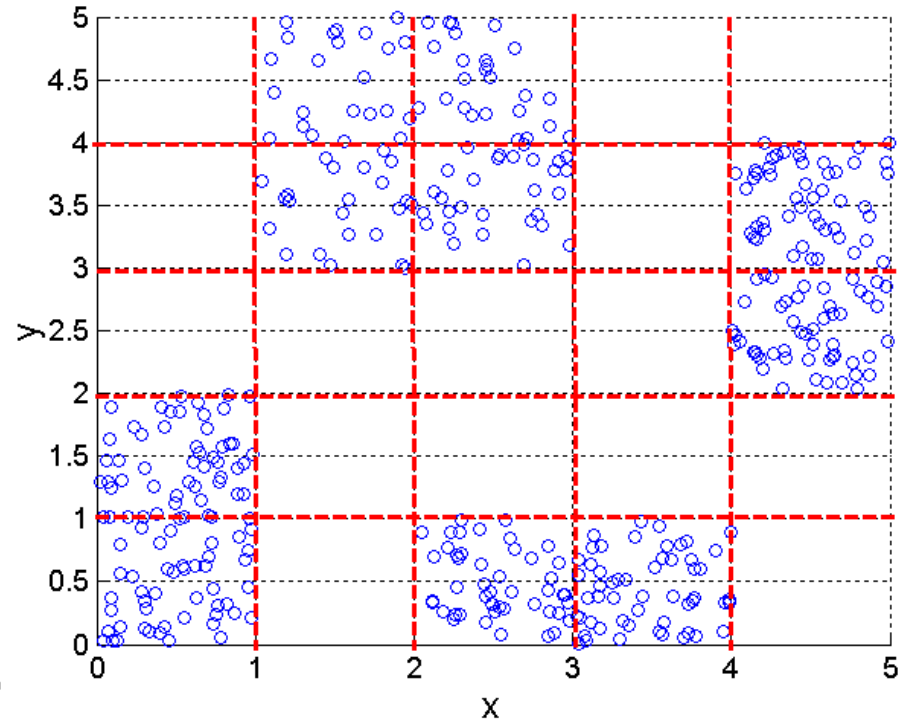
Frekvencia

Felügyelt diszkretizálás

- Entrópia alapú megközelítés

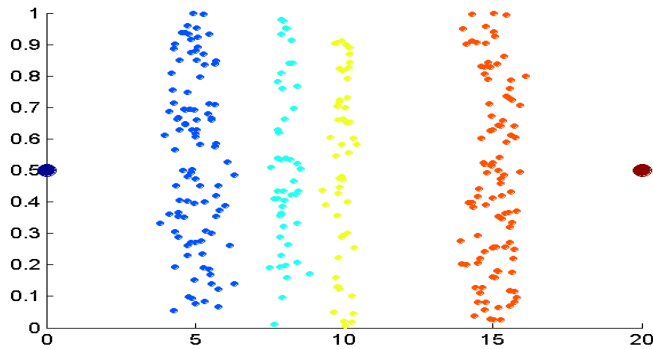


3 osztály x és y mentén

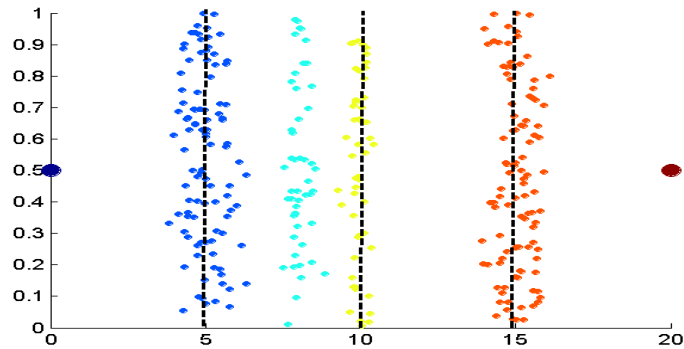


5 osztály x és y mentén

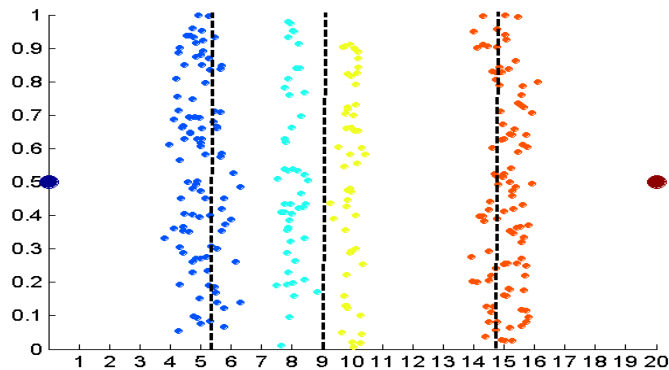
Nem-felügyelt diszkretizálás



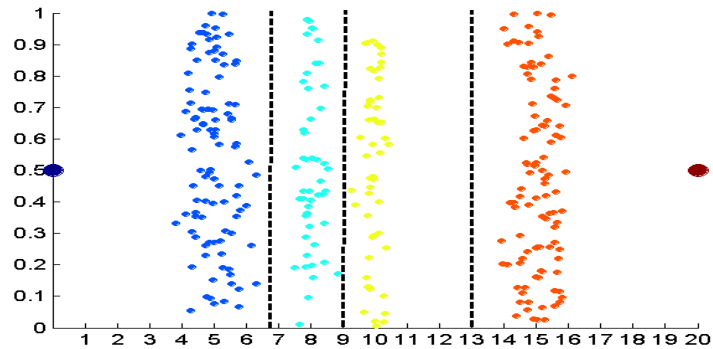
Adatok



Egyenlő szélességű intervallumok



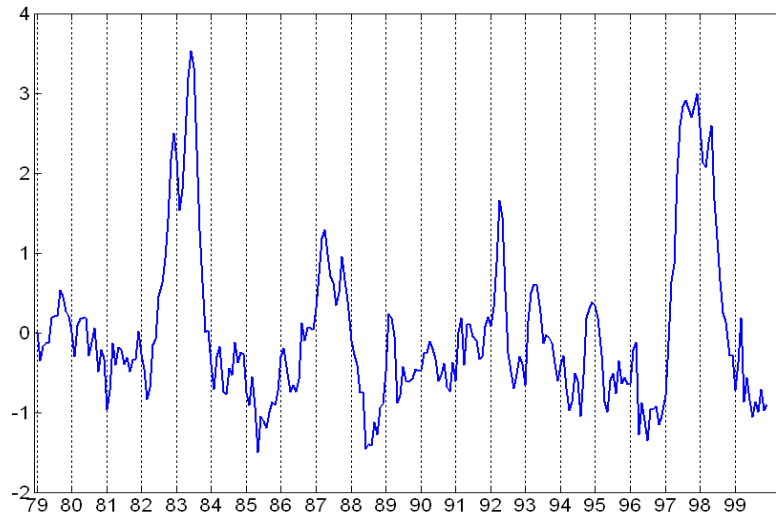
Egyenlő gyakoriságok



K-közép módszer

Attribútumok transzformációja

- Olyan függvény, amely adott attribútum értékeinek halmazát képezi le helyettesítő értékek egy új halmazára úgy, hogy minden régi érték egy új értékkel azonosítható.
 - Elemi függvények: x^k , $\log(x)$, e^x , $|x|$
 - Standardizálás és normalizálás



Hasonlóság és távolság

- Hasonlóság
 - Két objektum (rekord) hasonlóságát méri.
 - Minél nagyobb az értéke annál nagyobb a hasonlóság.
 - Általában a $[0,1]$ intervallumban veszi fel az értékeit.
- Távolság
 - Két objektum (rekord) különbözőségét méri.
 - Minél kisebb annál nagyobb a hasonlóság.
 - A minimális távolság általában 0.
 - A felső korlát változó.
- A szomszédság fogalma egyaránt utalhat hasonlóságra és távolságra.

Hasonlóság/távolság egyszerű attribútumnál

p és q jelöli két objektum attribútum értékét.

Attribute Type	Dissimilarity	Similarity
Nominal	$d = \begin{cases} 0 & \text{if } p = q \\ 1 & \text{if } p \neq q \end{cases}$	$s = \begin{cases} 1 & \text{if } p = q \\ 0 & \text{if } p \neq q \end{cases}$
Ordinal	$d = \frac{ p-q }{n-1}$ (values mapped to integers 0 to $n-1$, where n is the number of values)	$s = 1 - \frac{ p-q }{n-1}$
Interval or Ratio	$d = p - q $	$s = -d, s = \frac{1}{1+d} \text{ or } s = 1 - \frac{d - \min_d}{\max_d - \min_d}$

Table 5.1. Similarity and dissimilarity for simple attributes

Euklideszi távolság

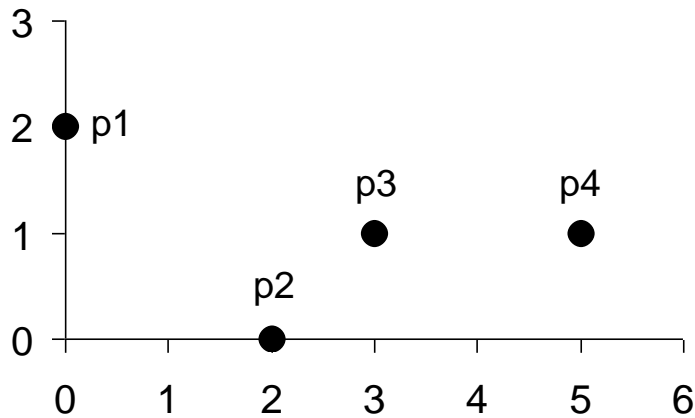
- Euklideszi távolság:

$$\text{dist} (p , q) = \sqrt{\sum_{k=1}^n (p_k - q_k)^2}$$

A képletben n jelöli a dimenziót (attribútumok száma), p_k és q_k pedig a k -adik attribútum értéke (koordinátája) a p és q objektumoknak (rekordoknak).

- Ha a skálák különbözőek, akkor előbb standardizálni kell.

Euklideszi távolság



pont	x	y
p1	0	2
p2	2	0
p3	3	1
p4	5	1

	p1	p2	p3	p4
p1	0	2.828	3.162	5.099
p2	2.828	0	1.414	3.162
p3	3.162	1.414	0	2
p4	5.099	3.162	2	0

Távolság mátrix

Minkowski távolság

- Az euklideszi távolság általánosítása

$$\text{dist} (p , q) = \left(\sum_{k=1}^n | p_k - q_k |^r \right)^{\frac{1}{r}}$$

A képletben r paraméter, n a dimenzió (attribútumok száma) p_k és q_k pedig a k -adik attribútum értéke (koordinátája) a p és q objektumoknak (rekordoknak).

Példák Minkowski távolságra

- $r = 1$: háztömb (Manhattan, taxi, L_1 norma) távolság.
 - Egy ismert példa az ún. Hamming távolság, amely éppen a különböző bitek száma két bináris vektorban.
- $r = 2$: euklideszi távolság
- $r \rightarrow \infty$: „szupremum” (L_{\max} norma, L_{∞} norma) távolság.
 - Két vektor koordinátái közötti különbségek abszolút értékének maximuma.
- Ne tévesszük össze r és n szerepét, ezek a távolságok minden dimenzió, azaz n mellett értelmezhetőek.

Minkowski távolság

point	x	y
p1	0	2
p2	2	0
p3	3	1
p4	5	1

L1	p1	p2	p3	p4
p1	0	4	4	6
p2	4	0	2	4
p3	4	2	0	2
p4	6	4	2	0

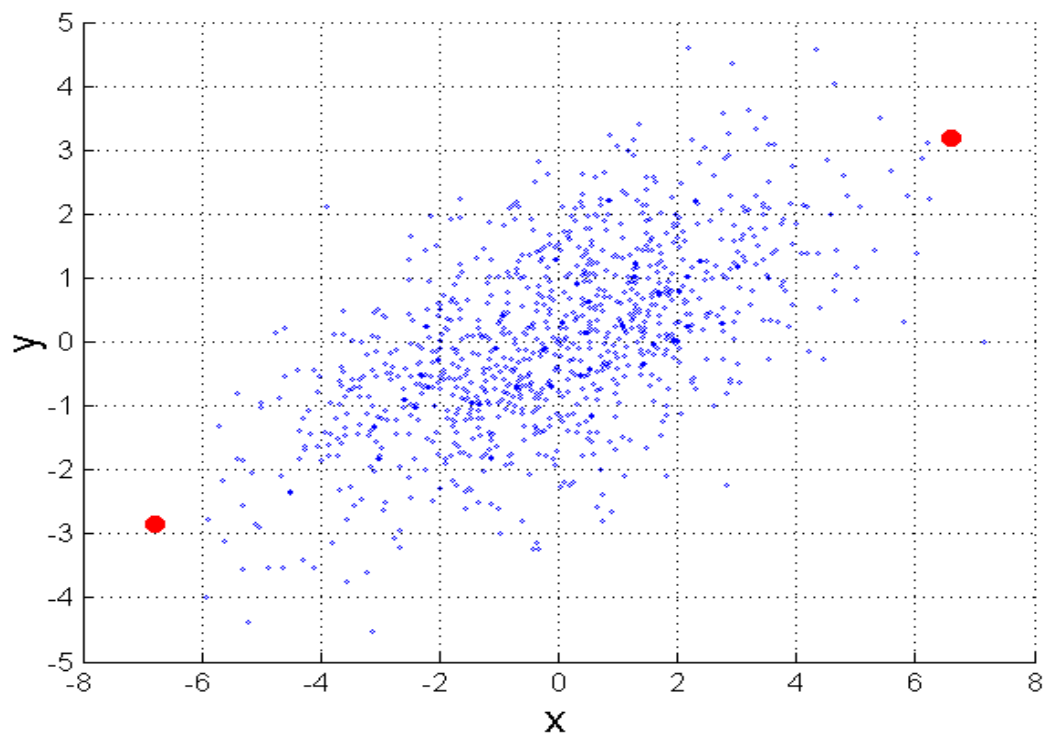
L2	p1	p2	p3	p4
p1	0	2.828	3.162	5.099
p2	2.828	0	1.414	3.162
p3	3.162	1.414	0	2
p4	5.099	3.162	2	0

L_{∞}	p1	p2	p3	p4
p1	0	2	3	5
p2	2	0	1	3
p3	3	1	0	2
p4	5	3	2	0

Távolság mátrix

Mahalanobis távolság

mahalanobi $s(p, q) = (p - q) \Sigma^{-1} (p - q)^T$

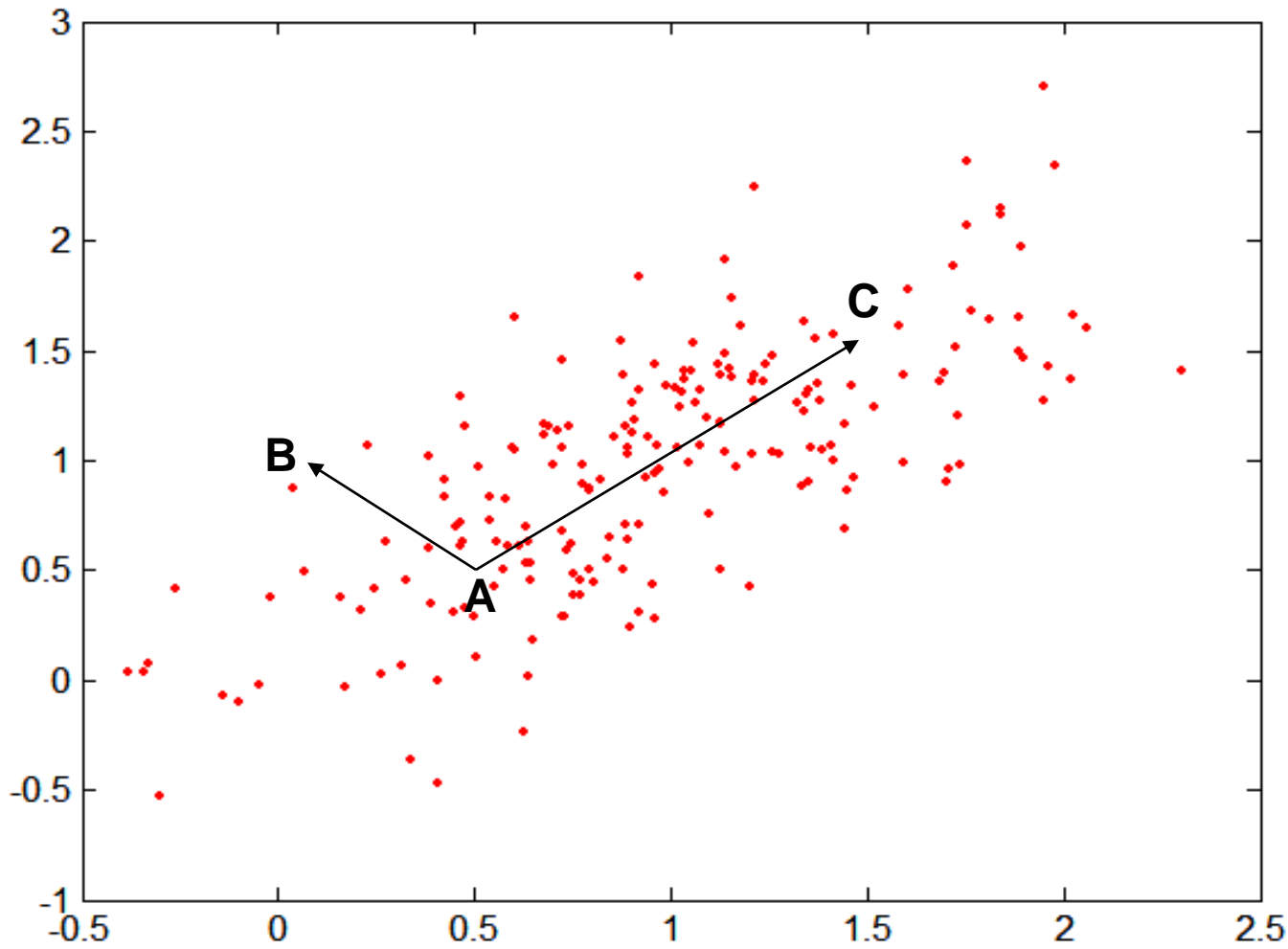


Σ az X input adatok kovariancia mátrixa

$$\Sigma_{j,k} = \frac{1}{n-1} \sum_{i=1}^n (X_{ij} - \overline{X}_j)(X_{ik} - \overline{X}_k)$$

A vörös pontok euklideszi távolsága 14.7, míg a Mahalanobis távolságuk 6.

Mahalanobis távolság



Kovariancia mátrix:

$$\Sigma = \begin{bmatrix} 0.3 & 0.2 \\ 0.2 & 0.3 \end{bmatrix}$$

A: (0.5, 0.5)

B: (0, 1)

C: (1.5, 1.5)

Mahal(A,B) = 5

Mahal(A,C) = 4

A távolság általános jellemzői

- A különböző távolság fogalmak, pl. euklideszi, néhány jól ismert jellemzővel bír.

1. $d(p, q) \geq 0$ minden p és q esetén, továbbá $d(p, q) = 0$ akkor és csak akkor ha $p = q$ (nemnegativitás),
2. $d(p, q) = d(q, p)$ minden p és q esetén (szimmetria),
3. $d(p, r) \leq d(p, q) + d(q, r)$ minden $p, q,$ és r pontra (háromszög egyenlőtlenség),

ahol $d(p, q)$ a p és q pontok (objektumok) közötti távolságot jelöli.

- Az olyan távolságot, amely eleget tesz a fenti tulajdonságoknak **metrikának** nevezzük.

A hasonlóság általános jellemzői

- A hasonlóságoknak szintén van néhány jól ismert tulajdonsága.
 1. $s(p, q) = 1$ (vagy a maximális hasonlóság) akkor és csak akkor ha $p = q$,
 2. $s(p, q) = s(q, p)$ minden p és q esetén (szimmetria),ahol $s(p, q)$ jelöli a p és q pontok (objektumok) közötti hasonlóságot.

Bináris vektorok hasonlósága

- Gyakran előfordul, hogy objektumoknak, p és q , csak bináris attribútumai vannak.
- Hasonlóságokat a következő mennyiségek révén definiálhatunk:

M_{01} = azon attribútumok száma, ahol $p=0$ és $q=1$,

M_{10} = azon attribútumok száma, ahol $p=1$ és $q=0$,

M_{00} = azon attribútumok száma, ahol $p=0$ és $q=0$,

M_{11} = azon attribútumok száma, ahol $p=1$ és $q=1$.

- Egyszerű egyezés és Jaccard együttható:

SMC = egyezők száma / attribútumok száma

$$= (M_{11} + M_{00}) / (M_{01} + M_{10} + M_{11} + M_{00})$$

J = az 11 egyezések száma / a nem mindkettő 0 attribútumok száma

$$= (M_{11}) / (M_{01} + M_{10} + M_{11})$$

Példa SMC és Jaccard hasonlóságra

$$p = 1000000000$$

$$q = 0000001001$$

$M_{01} = 2$ (azon attribútumok száma, ahol $p=0$ és $q=1$)

$M_{10} = 1$ (azon attribútumok száma, ahol $p=1$ és $q=0$)

$M_{00} = 7$ (azon attribútumok száma, ahol $p=0$ és $q=0$)

$M_{11} = 0$ (azon attribútumok száma, ahol $p=1$ és $q=1$)

$$\text{SMC} = (M_{11} + M_{00}) / (M_{01} + M_{10} + M_{11} + M_{00}) = (0+7) / (2+1+0+7) = 0.7$$

$$J = (M_{11}) / (M_{01} + M_{10} + M_{11}) = 0 / (2 + 1 + 0) = 0$$

Koszinusz hasonlóság

- Ha d_1 és d_2 két dokumentumot leíró vektor (nemnegatív egész koordinátájúak), akkor

$$\cos(d_1, d_2) = (d_1 \bullet d_2) / \|d_1\| \|d_2\| ,$$

ahol \bullet jelöli a skaláris szorzatot $\|d\|$ pedig a d vektor hossza.

- Példa:

$$d_1 = 3 \ 2 \ 0 \ 5 \ 0 \ 0 \ 0 \ 2 \ 0 \ 0$$

$$d_2 = 1 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 1 \ 0 \ 2$$

$$d_1 \bullet d_2 = 3 \cdot 1 + 2 \cdot 0 + 0 \cdot 0 + 5 \cdot 0 + 0 \cdot 0 + 0 \cdot 0 + 0 \cdot 0 + 2 \cdot 1 + 0 \cdot 0 + 0 \cdot 2 = 5$$

$$\|d_1\| = (3^2 + 2^2 + 0^2 + 5^2 + 0^2 + 0^2 + 0^2 + 2^2 + 0^2 + 0^2)^{0.5} = (42)^{0.5} = 6.481$$

$$\|d_2\| = (1^2 + 0^2 + 0^2 + 0^2 + 0^2 + 0^2 + 0^2 + 1^2 + 0^2 + 2^2)^{0.5} = (6)^{0.5} = 2.245$$

$$\cos(d_1, d_2) = .3150$$

A Jaccard együttható általánosítása

- Tanimoto együttható
- A Jaccard együttható módosítása azért, hogy alkalmazható legyen folytonos illetve egész értékű attribútumokra.
 - Bináris attribútumok esetén a Jaccard együtthatót kapjuk vissza

$$T(p, q) = \frac{p \bullet q}{\|p\|^2 + \|q\|^2 - p \bullet q}$$

Korreláció

- Az objektumok vagy attribútumok közötti lineáris kapcsolat erősségét méri.
- Két objektum (attribútum), p és q , közötti korreláció kiszámításához először standardizáljuk őket, majd skaláris szorzatot veszünk

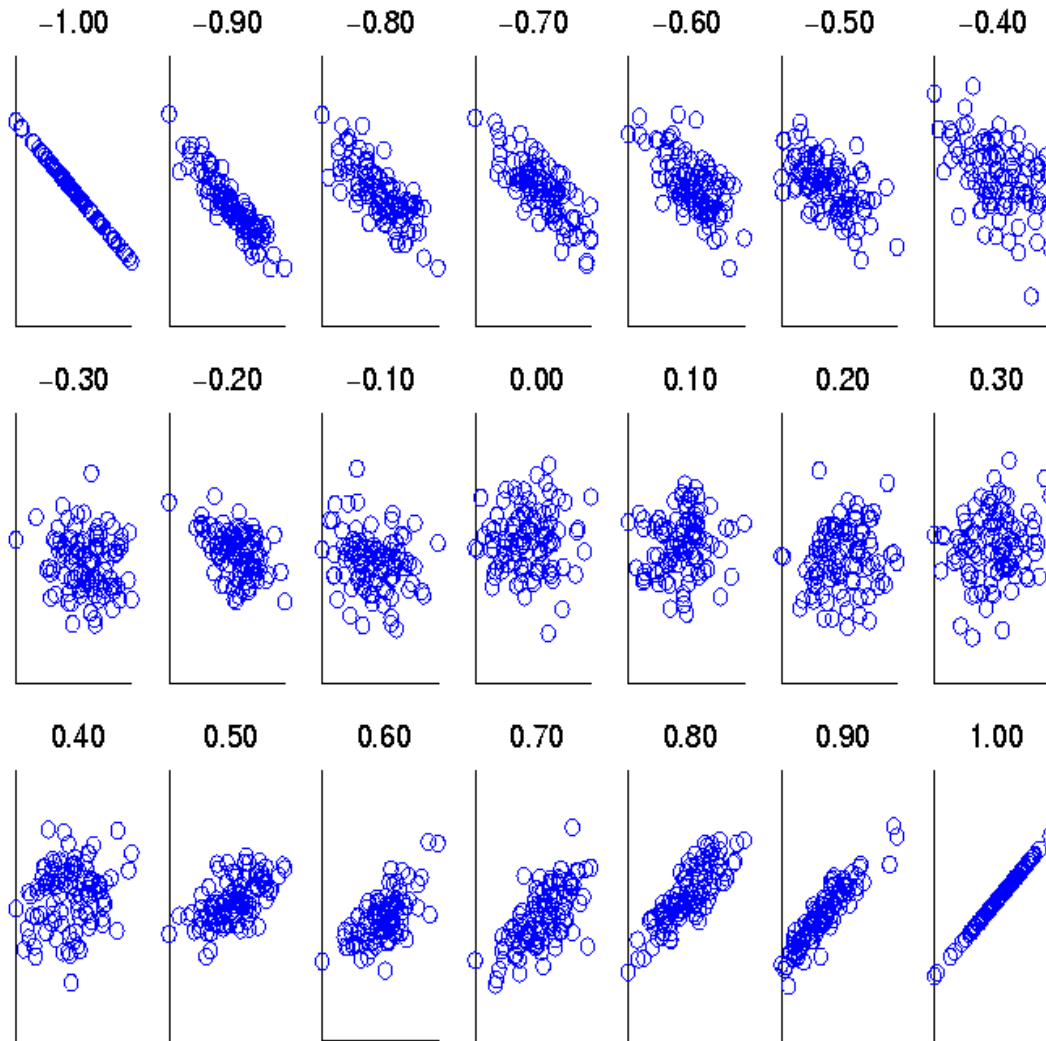
$$p'_k = (p_k - \bar{p}) / s(p)$$

$$q'_k = (q_k - \bar{q}) / s(q)$$

$$\text{korreláció } (p, q) = p' \cdot q'$$

ahol \bar{p} az átlag, $s(p)$ pedig a szórás.

A korreláció szemléltetése



A pontdiagramok szemléltetik a -1 -től 1 -ig terjedő hasonlóságot.

Hasonlóságok összekapcsolása

- Előfordul, hogy az attribútumok nagyon különböző típusúak viszont egy átfogó hasonlóságra van szükségünk.

1. For the k^{th} attribute, compute a similarity, s_k , in the range $[0, 1]$.

2. Define an indicator variable, δ_k , for the k^{th} attribute as follows:

$$\delta_k = \begin{cases} 0 & \text{if the } k^{th} \text{ attribute is a binary asymmetric attribute and both objects have} \\ & \text{a value of 0, or if one of the objects has a missing values for the } k^{th} \text{ attribute} \\ 1 & \text{otherwise} \end{cases}$$

3. Compute the overall similarity between the two objects using the following formula:

$$\text{similarity}(p, q) = \frac{\sum_{k=1}^n \delta_k s_k}{\sum_{k=1}^n \delta_k}$$

Hasonlóságok összekapcsolása súlyokkal

- Nem mindig akarjuk az összes attribútumot ugyanúgy kezelni.
 - Használjunk w_k súlyokat, melyek 0 és 1 közé esnek úgy, hogy az összegük 1.

$$\text{similarity}(p, q) = \frac{\sum_{k=1}^n w_k \delta_k s_k}{\sum_{k=1}^n \delta_k}$$

$$\text{distance}(p, q) = \left(\sum_{k=1}^n w_k |p_k - q_k|^r \right)^{1/r}$$

Sűrűség

- A sűrűség alapú csoportosításhoz szükséges a sűrűség fogalmának tisztázása.
- Példák:
 - Euklideszi sűrűség
 - ◆ Euklideszi sűrűség = egységnyi térfogatba eső pontok száma
 - Valószínűségi sűrűség
 - Gráf alapú sűrűség

Cella alapú euklideszi sűrűség

- Osszuk egyenlő térfogatú téglalap alakú cellákra a tartományt és definiáljuk a sűrűséget úgy, mint amely arányos a cellákba eső pontok számával.

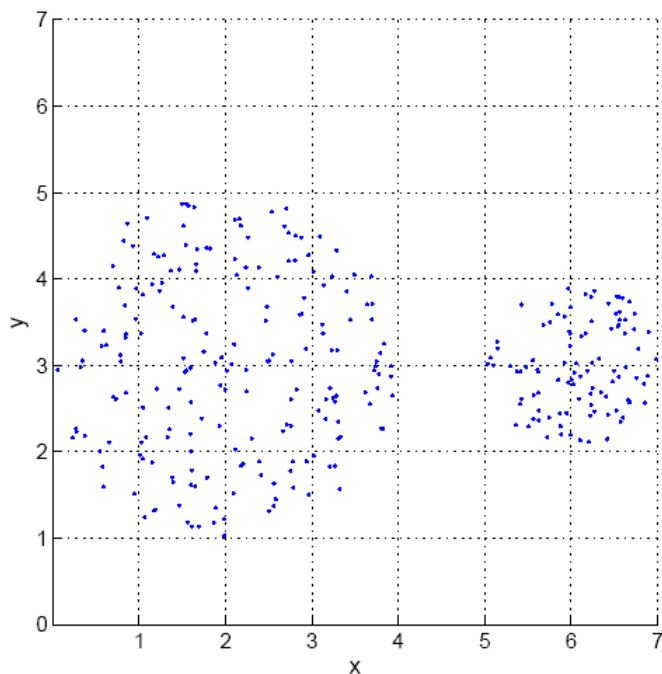


Figure 7.13. Cell-based density.

0	0	0	0	0	0	0
0	0	0	0	0	0	0
4	17	18	6	0	0	0
14	14	13	13	0	18	27
11	18	10	21	0	24	31
3	20	14	4	0	0	0
0	0	0	0	0	0	0

Table 7.6. Point counts for each grid cell.

Középpont alapú euklideszi sűrűség

- A sűrűség egy pontban arányos a pont körüli adott sugarú környezetbe eső pontok számával.

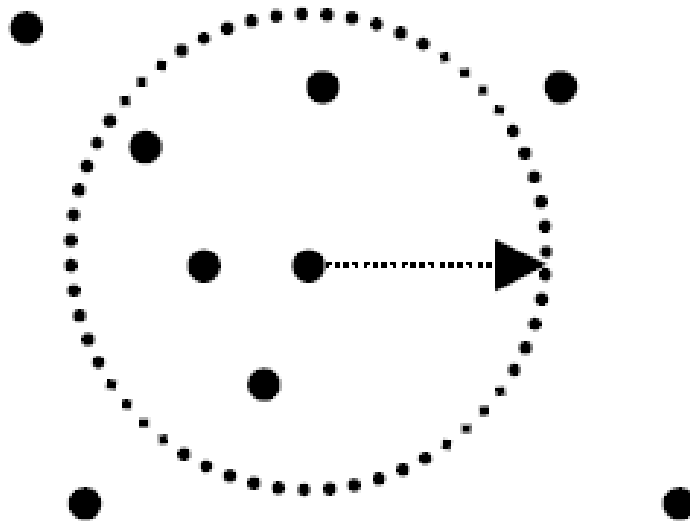


Figure 7.14. Illustration of center-based density.