Minden az adatról

Csima Judit

BME, VIK,
Számítástudományi és Információelméleti Tanszék

2014. február 13.

Adathalmaz

- elvileg bármi, ami információt hordoz és amiből valamilyen összefüggéseket akarunk kinyerni
- leggyakrabban úgy gondolunk az adathalmazra, mint egy tábázatra (data frame)
- sorok (rekordok): az egyes megfigyelések, emberek, esetek
- oszlopok az attribútumok, ezek azok a jellemzők, amik valamilyen értéket felvesznek minden egyes sorban
- egy esetet jellemeznek a sorának az attribútum-értékei
- lehet az adat eredendően másféle is, de arra törekszünk, hogy ilyen alakra hozzuk

Nyers adat (raw data)

- ahogy az adatot kapjuk, eredeti állapotában
- így nem lehet vele dolgozni, előfeldolgozás (preprocessing) szükséges
- data munging: az adatok elfogadható, feldolgozható formára hozása, nincs mindig bevált recept, sok idő
- de csak egyszer kell megcsinálni

Feldolgozott adat (processed data)

- feldolgozásra alkalmas állapotba hozott adat
- sok lépésből állhat az előfeldolgozás (erről később részletesen)
- nagyon fontos, hogy az előfeldolgozás is dokumentáltan történjen (honnan töltöttem le az adatot, mit csináltam vele, használt kódok is)

Tidy data (szép, tiszta adat)

Ez a minimumkövetelmény:

- egy táblában (egy sorhalmazban) azonos típusú sorok legyenek csak: pl. csak kórházak statisztikái vagy csak egyes emberekre vonatkozó sorok
- egy sor egy esetnek feleljen meg (pl. egy kórház vagy egy ember, egy eset)
- egy oszlop egy változónak feleljen meg, konzisztensen

További elvárások

Jó lenne tudni, hogy

- melyik oszlop milyen típusú adatot tartalmaz: attribútum fajtája, jelentése
- vannak-e hiányzó értékek
- vannak-e kilógó értékek (outlier)
- attribútum-értékek eloszlása milyen az egyes oszlopokon belül: át kell-e skálázni valamit
- van-e redundancia, azaz vannak-e azonos információt hordozó oszlopok

Ennek eléréséhez mindenféle technikák, erről majd az adatelemzés felépítésénél beszélünk részletesebben (data munging)

Attribútumok típusai: egy lehetséges felosztás

- folytonos:
  - valós értékeket vesz fel (de néha azt is folytonosnak hívjuk, amikor megszámlálhatóan végtelen lehetséges érték van)
  - pl. hőmérséklet, magasság, testsúly
- diszkrét:
  - véges sok (vagy megszámlálhatóan végtelen sok érték)
  - pl. irányítószám, életkor, nem, darabszám
  - gyakran egész számokkal reprezentált, néha címkékkel (label)
- bináris:
  - speciális diszkrét attribútum: 0 és 1 a lehetséges értékek
  - gyakran asszimmetrikus jelentésű: a 0 azt jelenti, hogy valami nincs, nem igaz
  - gyakran ritka adatmátrixokban szerepel: nagyon sok a 0 (például dokumnetum-szó mátrixok)
  - speciális kezelés lehet néha szükséges

Attribútumok típusai: egy (hasonló) felosztás

- kvalitatív attribútumok (categorical attribute)
  - címkék, például személy neme, családi állapota, kapott terápia, túlsúlyos-e?
  - értelmes műveletek: gyakoriságok (hisztogramon ábrázolva)
  - jó, hasznos, ha az attribútumok értékei kifejezőek (pl. férfi-nő és nem 1-2)
  - R-ben ennek a factor típusú változók felelnek meg
- kvantitatív attribútumok
  - életkor, testsúly, testmagasság, BMI index
  - értelmes műveletek: medián, percentilisek, esetleg átlag, szórás
  - kérdés, hogy csak a sorrend számít vagy a különbség illetve az arány is értelmes, pl. 20 °C az nem kétszer olyan meleg, mint 10 °C

Attribútumok felosztása: még egy felosztás

Rekord típusú adatokból álló táblázat, mátrix

- számokból álló m soros, n oszlopos táblázat
- gyakran az n dimenziós tér pontjainak tekintjük a sorokat
- speciális eset: dokumentum-szó mátrix
  - sorok a dokumentumok, oszlopok a kulcsszavak
  - bináris attribútum mutatja, hogy szerepel-e az adott szó vagy diszkrét attribútum mutatja az előfordulás darabszámát
  - általában rengeteg oszlop van, nagy a dimenzió
- speciális eset még: tranzakciós adatokból származtatott adatmátrix: eredetileg halmazok, de könnyen átalakítható a dokumentum-szó mátrixhoz hasonlóan

Attribútumok felosztása: még egy felosztás

Nem rekord típusú adathalmaz, ilyeneket általában addig alakítjuk, amíg rekord típusúak lesznek

- grafikus adatok: molekulák közötti kapcsolatok: ki kivel kapcsolódik, kötések szögei
- képek: pixelsorozatra fordítható le vagy valami származtatott feature-lista alapján kap számszerűsíthető attribútumokat minden kép
- térbeli és/vagy időbeli kapcsolat is van a sorok között: pl. adott pillanatban meteorológiai mérések több helyen (ábrázolásnál jó ennek tudatában lenni)

Adatminőséggel kapcsolatos kérdések

- Mik a lehetséges problémák az adattal?
- Hogy vesszük észre ezeket?
- Hogyan kezeljük a megtalált hibákat?

Mik a lehetséges problémák az adattal?

- mérési hibák
- inkonzisztencia, pl. az adathalmaz egyik felében km, a másikban m-ben vannak az adatok
- hiányzó adatok
- duplikátumok: feleslegesen ismétlődő sorok, nem mindig teljesen egyformák, pl. adatbázisban ugyanaz az ember több hasonló lakcímmel
- furcsa, nehezen hihető adatok (mindenki túlsúlyos az adatbázis szerint vagy minden lakásban 100-nál több szoba van)
- outlier-ek: kilógó, furcsa. másmilyen sorok vagy attribútumértékek (lehet, hogy baj, lehet, hogy nem)

Hogy vesszük észre ezeket?

- ez az előfeldolgozás és az exploratory elemzés része
- grafikus ábrázolás: eloszlások, hisztogramok
- összegző függvények futtatása az adatokra (mean, median, percentilisek, R-ben summary)

Hogyan kezeljük a megtalált hibákat?

- az mindig jó, ha legalább tudjuk, hogy mivel állunk szemben
- van amivel nem lehet sokat tenni (pl. mérési hiba), de legalább tudatában vagyunk annak, hogy volt ilyen
- amúgy meg adattisztítás, erről később részletesen
- hiányzó értékek:
  - lehet, hogy nem baj (nem minden sorban értelmes az adott attribútum)
  - megoldás lehet az adott érték pótlása vagy a sor törlése
  - az is lehet, hogy elég, ha tudunk a jelenségről
- duplikátumok: észrevenni őket és azonosítani a közel azonosakat (néha csak ezt a részt hívjuk adattisztításnak)
- outlier: lehet, hogy el kell hagyni, de lehet, hogy épp az ilyeneket akarom megtalálni

Hasonlóság, különbözőség

- Sokszor fontos lehet annak mérése, számszerűsítése, hogy két sor (két pont) mennyire hasonlít
- Legfontosabb ilyen helyzet a klaszterezés, amikor a hasonlóakat akarjuk egybe gyűjteni
- A hasonlóság illetve különbözőség mérésére többféle lehetséges függvény van
- A használt függvény mindenképpen függ attól, hogy milyen típusú attribútumokból áll a sor (folytonos vagy sem illetve kvalitatív vagy kvantitatív)
- Alapmegközelítés, hogy oszloponként (mezőnként) definiáljuk a távolságot és aztán a sorok távolsága ezekből adódik (erről később)
- Először azt kell tisztázni, hogy egy oszolopon belül mit jelent két érték távolsága

Hasonlóság jellemzői (similarity)

- Azt méri, hogy ennyire hasonlóak, egyformák
- Minél nagyobb a szám, annál hasonlóbbak
- Szimmetrikus, azaz p és q hasonlósága ugyanaz, mint q és p hasonlósága
- Általában $[0, 1]$ közötti értékek (ritkábban $[0, \infty]$ közötti értékeket vesz fel

Különbözőség (dissimilarity)

- Azt méri, hogy mennyire különböznek
- Minél kisebb az érték, annál egyformábbak
- Általában a 0 jelentése az, hogy egyformák
- Szimmetrikus, azaz p és q különbözősége ugyanaz, mint q és p különbözősége

Mikor mit használunk?
Kategorikus attribútumoknál

- hasonlóság: 1, ha egyformák és 0, ha nem egyformák
- különbözőség pont fordítva: 0, ha egyformák és 1, ha nem egyformák
- ha a címkék által kódolt dolgok között van valami csoportosítás, akkor lehet nem bináris is a függvény: aminosav szekveciák összevetésénél nem csak az számít, hogy egyformák-e, mert vannak nem egyforma, de hasonló aminosavak (hidrofób versus hidrofil, alakjuk, stb.)
- bioinformatikában rengeteg féle pontozómátrix van: egyforma aminosavakra az érték 0, különben meg minél különbözőbbek, annál nagyobb

Mikor mit használunk?
Ha az értékek egy adott intervallumból kerülhetnek ki

Ha a lehetséges értékek $1, 2, \ldots, n$:

- különbözőség:
  - p és q különbözősége $\frac{|p-q|}{n-1}$
  - ez 0 és 1 közé lövi be a különbözőséget
  - 0, ha megegyeznek
- hasonlóság:
  - p és q hasonlósága $1 - \frac{|p-q|}{n-1}$
  - ez 0 és 1 közé lövi be a hasonlóságot
  - 1, ha megegyeznek

Mikor mit használunk?
Ha az értékek nem egy véges intervallumból valók

- különbözőség:
  - p és q különbözősége $d(p, q) = |p - q|$
  - ez 0 és ∞ közé lövi be a különbözőséget
  - 0, ha megegyeznek
- hasonlóság:
  - sokféleképp származtatható a fenti különbözőségből hasonlóság
  - ellentett, azaz -d(p,q): $-\infty$ és 0 közötti értékeket vesz fel
  - $\frac{1}{1+d}$: 0 és 1 közötti értékek

Több azonos típusú attribútummal rendelkező sor összehasonlítása

- Oszloponként képezzük a távolságot
- Aztán:
  - vagy összegezzük az oszloponkénti távolságokat
  - vagy az összeget elosztjuk az oszlopszámmal
  - vagy súlyozott összeget számolunk (és utána osztunk az oszlopszámmal)
  - oszloponkénti távolságképzés előtte szükség lehet átskálázásra (standardizálás): azonos nagyságrendűek legyenek az attribútumok értékei (szobaszám versus négyzetméter)

Távolság fogalma

Leggyakrabban egy speciális alakú különbözőség-fogalommal dolgozunk, ennek neve távolság.
Jellemzői:

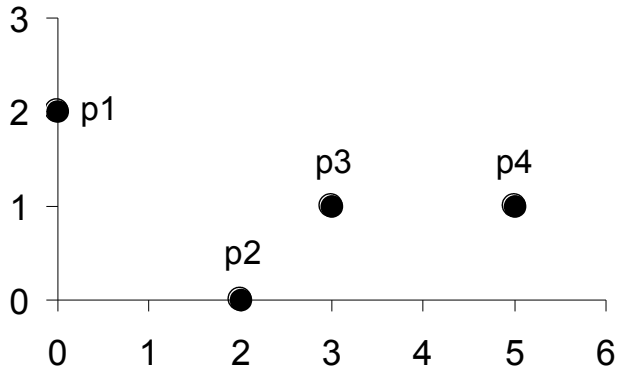
- $d(p, q) \geq 0$ mindig igaz és $d(p, q) = 0$ csak akkor, ha $p = q$ (reflexivitás)
- $d(p, q) = d(q, p)$ (szimmetria)
- $d(p, q) \leq d(p, r) + d(r, q)$ minden p, q, r esetén (háromszög egyenlőtlenség)

Más néven: metrika.

Euklideszi távolság

- Leggyakrabban ezt használjuk, ha a sorok értelmezhetők n-dimenziós térben levő pontokként
- $p = (p_1, \ldots, p_n)$ és $q = (q_1, \ldots, q_n)$ két pont a térben
- $d(p, q) = \sqrt{\sum_{k=1}^{n} (p_k - q_k)^2}$
- itt is kellhet előbb a standardizálás:
  - $\frac{p - mean(p)}{sd(p)}$, azaz kivonjuk az átlagot és osztunk a szórással
  - vagy $\frac{p - min(p)}{max(p) - min(p)}$

Euclidean Distance



point	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

Distance Matrix

Minkowski távolság, L_r távolság

- Euklideszi távolság álatalánosítása
- $p = (p_1, \ldots, p_n)$ és $q = (q_1, \ldots, q_n)$ most is két pont a térben
- van egy paramétere, r, ez valami $1, 2, \ldots$ egész szám
- $d(p, q) = \sqrt[r]{\sum_{k=1}^{n} |p_k - q_k|^r}$
- $r = 2$ az Euklideszi távolság
- itt is kellhet előbb a standardizálás (minél nagyobb az r, annál inkább)
- ez minden r egész szám esetén metrika

Minkowski távolság, speciális esetek

- $r = 1$: Manhattan távolság
  - L_1 távolsága $(1, 2)$ és $(7, 0)$-nak 8, ennyi blokkra/sarokra vannak egymástól
- $r = 2$ az Euklideszi távolság
- van olyan is, hogy $r = \infty$, ez az L_∞, néha hívják L_{max}-nak is
  - egyik definíció: $d(p, q) = \lim_{r \to \infty} \sqrt[r]{\sum_{k=1}^{n} |p_k - q_k|^r}$
  - ami ugyanaz, mint $d(p, q) = \max_{k \in \{1,2,\ldots,n\}} |p_k - q_k|$
  - ez is metrika

Minkowski Distance

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

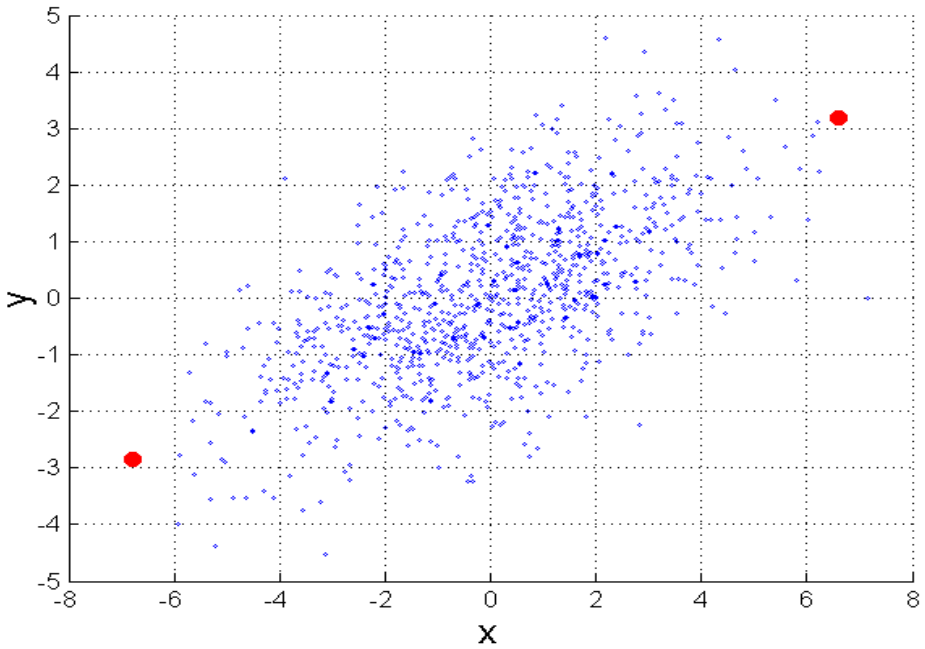
Distance Matrix

Mahalanobis távolság

- az L_r távolságok nem veszik figyelembe, hogy az adatmátrix oszlopai nem feltétlenül függetlenek
- szélsőséges esetben lehet két azonos oszlop, ennek eltérése így duplán számít
- erre megoldás lehet az, ha a mátrixot átalakítjuk az elemzés lőtt, új változók bevezetésével vagy a régiek közül néhány elhagyásával (erről később részletesen lesz szó)
- vagy megoldás az, ha olyan távolságfogalmat használunk, ami ellensúlyozza az oszlopok korreláltságából adódó torzítást

Mahalanobis Distance

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

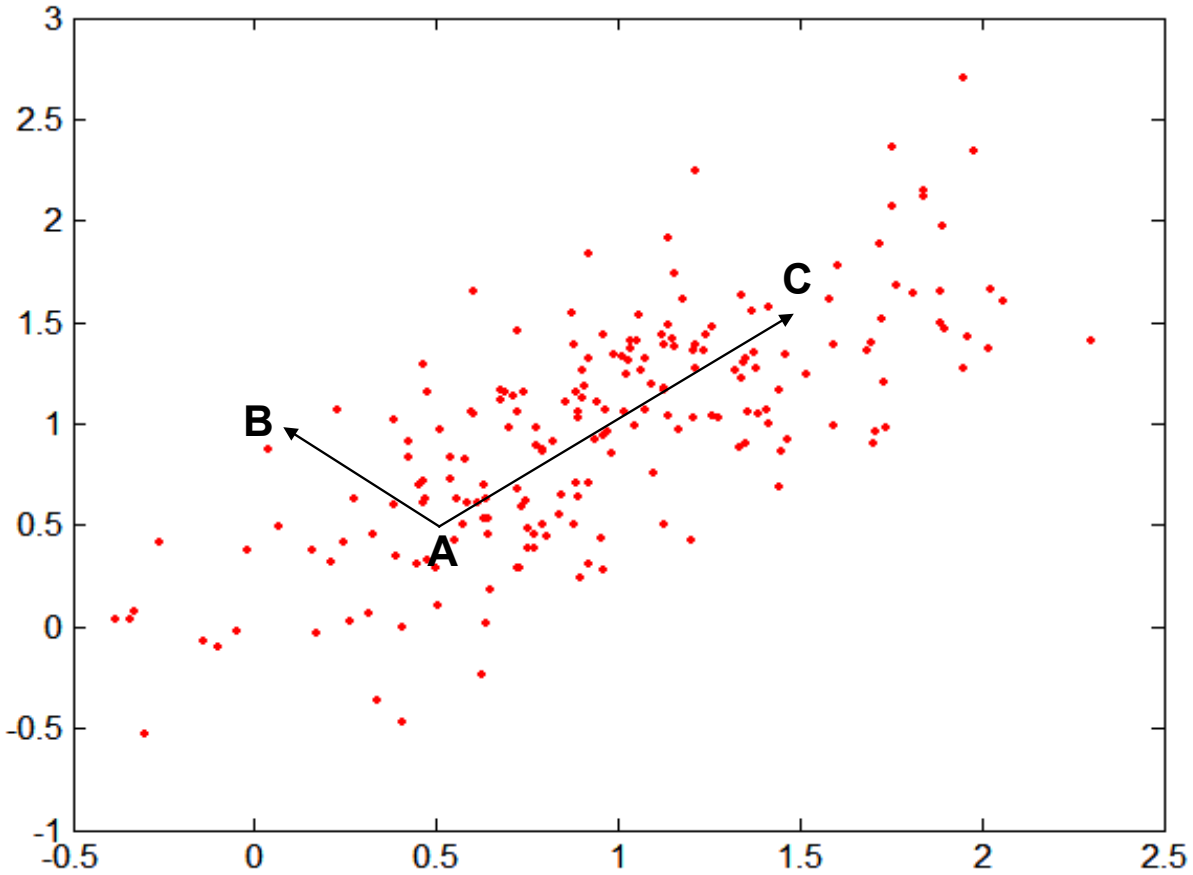


Σ is the covariance matrix of the input data *X*

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

For red points, the Euclidean distance is 14.7, Mahalanobis distance is 6.

Mahalanobis Distance



Covariance Matrix:

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

A: (0.5, 0.5)

B: (0, 1)

C: (1.5, 1.5)

Mahal(A,B) = 5

Mahal(A,C) = 4

Bináris vektorok hasonlósága

- ha binárisak az adatok, akkor nagyon gyakran ritka adatmátrixról van szó: szinte minden bejegyzés 0 (dokumentum-szó mátrix, tranzakciós mátrix)
- ebben az esetben az eddigi távolságfogalmak nem informatívak: szinte mindenki egyformának látszik
- kéne valami speciálisabb távolság ezekre az esetekre
- p és q most is n hosszú vektorok, de minden komponens értéke 0 vagy 1
- itt hasonlóságok vannak (azaz minél nagyobb az érték, annál egyformábbak)

Simple matching coefficient (SMC)

- M_{01} = hány helyen van p-ben 0 és q-ban 1
- M_{10} = hány helyen van p-ben 1 és q-ban 0
- M_{00} = hány helyen van p-ben és q-ban is 0
- M_{11} = hány helyen van p-ben és q-ban is 1
- $\text{SMC} = \frac{(M_{00} + M_{11})}{(M_{00} + M_{11} + M_{01} + M_{10})}$
- SMC tehát = ahol egyeznek osztva az attribútumok számával
- SMC tehát = ahol egyeznek osztva az attribútumok számával
- ez lényegében az L_1 távolságnak megfelelő hasonlóság

Jaccard együttható

- SMC nem jól mér, ha ritka az adatmátrix
- mert nagyon befolyásolja a SMC szerinti hasonlóságot ha sok közös nulla van (pl. sok olyan szó, ami egyik dokumentumban sincs benne)
- megoldás: a közös nullák ne számítsanak: Jaccard együttható
- $\text{Jaccard} = \frac{(M_{11})}{(M_{11} + M_{01} + M_{10})}$
- hány közös előfordulás van a valahol előforduló szavak számához képest

SMC versus Jaccard: Example

p = 1 0 0 0 0 0 0 0 0 0

q = 0 0 0 0 0 0 1 0 0 1

$M_{01} = 2$ (the number of attributes where $p$was 0 and $q$was 1)

$M_{10} = 1$ (the number of attributes where $p$was 1 and $q$was 0)

$M_{00} = 7$ (the number of attributes where $p$was 0 and $q$was 0)

$M_{11} = 0$ (the number of attributes where $p$was 1 and $q$was 1)

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

Cosine hasonlóság

- dokumentum-szó mátrix esetén hasznos, ha a mátrix gyakoriságokat tartalmaz (nem bináris, hanem azt mutatja, hogy hányszor szerepelt egy kulcsszó)
- p és q két azonos hosszúságú, egész számokból álló vektor (továbbra is igaz, hogy sok bennük a nulla)
- $cos(p, q) = \frac{p \cdot q}{\|p\| \cdot \|q\|}$
- azaz skalárisan összeszorozzuk a két vektort és osztunk a hosszuk szorzatával
- ismert közeépiskolából, hogy ez a síkon a két vektor szögének a cosinus-a
- ez igaz három dimenzióban is

Cosine Similarity

- If d_1 and d_2 are two document vectors, then

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

where • indicates vector dot product and $\| d \|$ is the length of vector d.

- Example:

d_1 = **3 2 0 5 0 0 0 2 0 0**

d_2 = **1 0 0 0 0 0 0 1 0 2**

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

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

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

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

Különböző fajta attribútumokat sorok összehasonlítása

- az eddigi módszerek akkor jók, ha az összehasonlítandó vektorok azonos típusú értékeket tartalmaznak minden oszlopban
- ha nem így van:
  - csoportosítsuk össze az egyformákat: binárisak, kategorikusak, folytonosak, stb.
  - számoljuk ki az egyes csoportokra a hasonlóságot vagy távolságot
  - arra figyeljünk, hogy azonos típusú dolgot számoljunk mindenhol (vagy távolság vagy hasonlóság)
  - valahogyan (esetleg súlyozva az egyes részek nagysága vagy értéke szerint) eredő távolságot vagy hasonlóságot definiálunk

Súlyozás általában

- akkor is akarhatunk súlyozni, ha egyszerűen csak vannak attribútumok, amik kevésbé fontosak
- például L_r normát is lehet súlyozni: $\sqrt[r]{\sum_{k=1}^{n} w_k \cdot |p_k - q_k|^r}$

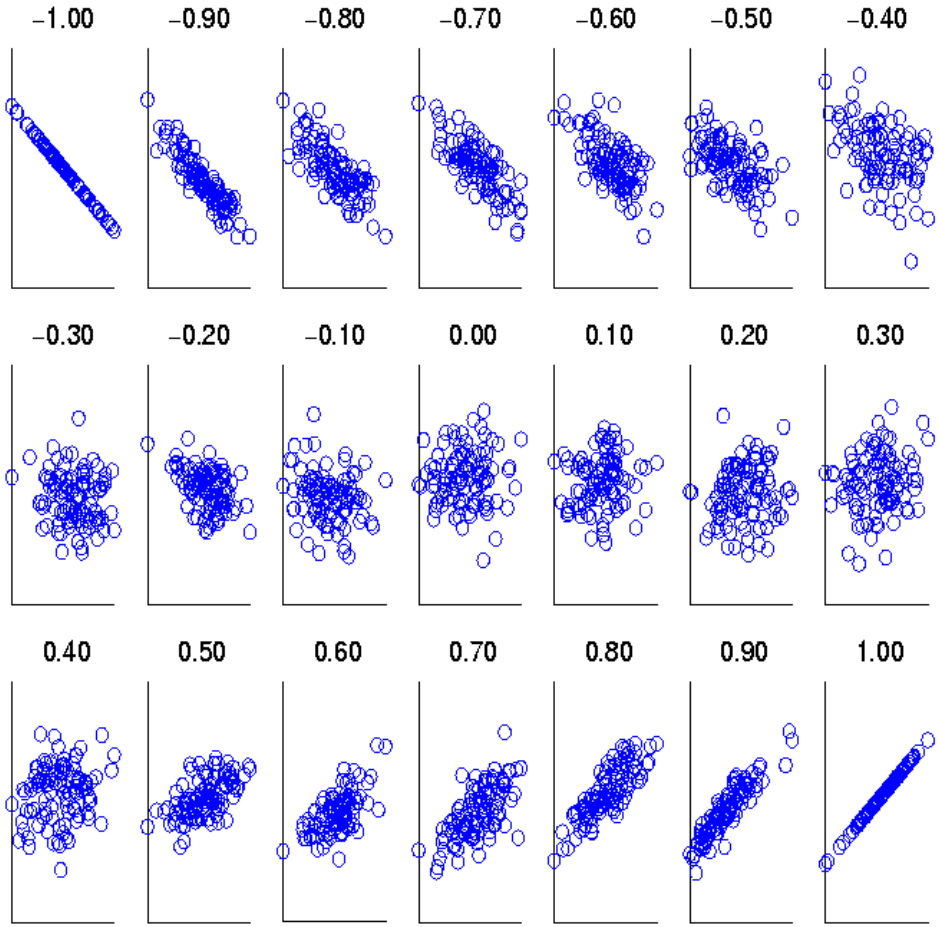
Korreláció

- ezzel általában oszlopokat hasonlítunk össze
- nem az algoritmusokban használjuk, hanem az előfeldolgozásnál, amikor az algoritmusokban használt attribútumokat határozzuk meg
- két oszlop, azaz két attribútum közötti lineáris kapcsolatot méri
- arra lehet jó, hogy ha nagy a korreláció két oszlop között, akkor esetleg elég egyiket bevenni az elemzésbe
- vigyázat! nem minden kapcsolatot derít fel, csak a lineárisat!

Korreláció: definíció

- előbb standardizáljuk az oszlopokat: p_k helyett $p'_k = \frac{p_k - mean(p)}{sd(p)}$, hasonlóan q'
- correlation(p,q) = $\frac{p' \cdot q'}{n}$ (skalárszorzat, osztva a hosszal)
- ez ugyanaz, mint a szokásos definíció
- beépített függvénnyel számoljuk R-ben: cor

Visually Evaluating Correlation



Scatter plots showing the similarity from –1 to 1.