Előfeldolgozás

Nagyméretű adathalmazok kezelése – Molnár András

Tartalom

- Történeti áttekintés, adatok rendelkezésre állása
- Attribútumok típusai, tulajdonságai
- Távolsági függvények
- Előfeldolgozás
  - Hiányzó értékek kezelése
  - Attribútum transzformációk
  - Adatok torzítása
  - Diszkretizálás
  - Normalizálás
  - Mintavételezés
  - Dimenzió csökkentés
  - Duplikátumok kiszűrése
  - Aggregáció
  - Monotonizáció

Történeti áttekintés

- 90-es évektől
- Döntések: megérzés alapján
- Cél: adat -> információ -> döntés
- Hagyományos adatbázisok nem elegendőek
- Megszületett: adatbányászat
- Eleinte káosz
- XXI. Századtól egyre népszerűbb
- Jövő: még rengeteg kihívás…

Adatok rendelkezésre állása

- Zajosak
- Inkonzisztensek
- Hiányos
- Több forrásból származóak
- Óriás méretűek

SZÜKSÉG VAN ELŐFELDOLGOZÁSRA!

Tartalom

- Történeti áttekintés, adatok rendelkezésre állása
- Attribútumok típusai, tulajdonságai
- Távolsági függvények
- Előfeldolgozás
  - Hiányzó értékek kezelése
  - Attribútum transzformációk
  - Adatok torzítása
  - Diszkretizálás
  - Normalizálás
  - Mintavételezés
  - Dimenzió csökkentés
  - Duplikátumok kiszűrése
  - Aggregáció
  - Monotonizáció

Attribútumok

- Kategorizálni kell
- Nagy adathalmazok -> táblaként
- Sorok: objektumok, példányok, rekordok
- Oszlopok: attribútumok

- Pl.: EMBER adathalmaz, tábla

  Példányok: személyek

  Attribútumok: neme, szül. helye, magassága stb.

Attribútumok típusai

- Kategória típusú
- Csak azonosság vizsgálat
- a=a’ vagy a!=a’
- Nevezik még: felsorolás vagy diszkrét típusnak
- Például: vallás, nemzetiség

- **Bináris típus**: férfi vagy nő?

Attribútumok típusai

- Sorrend típusú
- Értékek sorba rendezése
- Teljes rendezés az attribútumok értékén
- Vagyis: ha a!=a, akkor tudjuk, hogy a < a’ vagy a > a’
- Például: versenyhelyezés

Attribútumok típusai

- Intervallum típusú
  - Bevezetjük a + műveletet
  - Csoportot alkot az eddigi tulajdonsággal
  - Például: ember súlya, magassága

Attribútumok típusai

- Arány skálájú
- Intervallum típus tulajdonságai PLUSZ
- Meg lehet adni zérus értékeket
- Definiált két attr. érték hányadosa
- Gyűrűt alkotnak
- Például: évszámok (ha definiálva van a nullpont)

Attribútumok tulajdonságai

- Arány és Intervallum együttesen: NUMERIKUS típus
- Kategorizálás nem mindig triviális
- Példa:

  napsütéses, borús, esős

  Kategória? Intervallum?

Attribútumok statisztikai tulajdonsága

- Középértékre vonatkozó adatok: mintaátlag, medián, módusz
- Szóródásra vonatkozó adatok: empirikus szórásnégyzet, minimum, maximum, terjedelem (max és min érték közötti különbség)
- Eloszlásra vonatkozó adatok: empirikus kvantilisek, ferdeség, lapultság

Ferdeség, lapultság

- Szimmetria számszerűsítése
- 0: szimmetrikus
- Negatív: balra dől eloszlás
- Pozitív: jobbra dől az eloszlás
- Például: Gauss eloszlás = 0

- Eloszlás csúcsosságát adja meg

Tartalom

- Történeti áttekintés, adatok rendelkezésre állása
- Attribútumok típusai, tulajdonságai
- Távolsági függvények
- Előfeldolgozás
  - Hiányzó értékek kezelése
  - Attribútum transzformációk
  - Adatok torzítása
  - Diszkretizálás
  - Normalizálás
  - Mintavételezés
  - Dimenzió csökkentés
  - Duplikátumok kiszűrése
  - Aggregáció
  - Monotonizáció

Távolsági függvények

- Objektumok hasonlóságának vizsgálatára
- Távolság (különbözőség) függvény: hasonlóság inverze
- d(x,y) : x és y közötti különbség számszerűsítve
- Objektum önmagától nem különbözhet: d(x,x)=0
- Szimmetrikusság: d(x,y)=d(y,x)
- Háromszög egyenlőtlenség: d(x,y)<d(x,z) + d(y,z)
- Hasonlósági függvényé alakítható

Bináris attribútumok távolsága

x \ y	1	0	$\sum$
1	q	r	q+r
0	s	t	s+t
$\sum$	q+s	r+t	m

- Invariáns: bináris attribútum mindkét értéke ugyanolyan fontos
- pl: férfi vagy nő

  eltérő attribútumok relatív száma: $$d(x,y) = \frac{r+s}{m}.$$

Bináris attribútumok távolsága

- Variáns távolság: aszimmetria van az értékek között
- Például: orvosi jelentés kimenetele
- Jaccard koeffciens:

$$d(x, y) = 1 - \frac{q}{m - t} = \frac{r + s}{m - t},$$

- Vegyes attribútum: ha szimmetrikus és aszimmetrikus is van a bináris attribútumok között

Kategória típusú attr. távolsága

- Véges sok értéket vehet fel
- Például: ember szeme színe, vallása, családi állapota
- Nem egyezések száma (szimmetrikus):

  ahol u megadja, hogy x és y közül

  mennyi nem egyezett meg

$$d(x, y) = \frac{u}{m},$$

- Jaccard-koeffciens értelmezhető itt is -> aszimmetrikusság

Kategória típusú attribútum távolsága: példa

$$x_1 = (\text{életkor} = 20\text{–}25,\ \text{autó} = \text{Opel},\ \text{végzettség} = \text{egyetem},\ \text{nem} = \text{férfi},\ \text{családi állapot} = \text{nőtlen})$$

és

$$x_2 = (\text{életkor} = 20\text{–}25,\ \text{végzettség} = \text{egyetem},\ \text{nem} = \text{nő},\ \text{családi állapot} = \text{házas},\ \text{hobbi} = \text{könyvek olvasása},\ \text{vallás} = \text{buddhista})$$

$$d(x_1, x_2) = 1 - \frac{2}{7} = 0.714.$$

Sorrend típusú attribútum távolsága

- Például: 8 általános, befejezett középiskola, érettségi, főiskolai diploma, egyetemi diploma, doktori cím
- Arány ~ Sorrend: Forma-1-es verseny
- Egész számokkal helyettesítjük: 1 – M –ig
- Célszerű leképezni a 0-tól 1-ig tartó intervallumba (normalizálás)

Intervallum attribútumok távolsága

- Például: ember súlya, magassága, vagy egy ország éves átlaghőmérséklete
- Általában: valós számok
- Tekinthető: m dimenziós tér egy-egy pontjainak
- Elemek különbsége: a vektorok távolsága

Euklideszi-norma: $L_2(\vec{z}) = \sqrt{|z_1|^2 + |z_2|^2 + \cdots + |z_m|^2}$

Manhattan-norma: $L_1(\vec{z}) = |z_1| + |z_2| + \cdots + |z_m|$

Minkowski-norma: $L_p(\vec{z}) = (|z_1|^p + |z_2|^p + \cdots + |z_m|^p)^{1/p}$

Feladat

Példa: Számítsuk ki a következő két objektum L_1 (Manhattan-norma), L_2 (Euklideszi-norma), L_4 (Minkowski-norma $p = 4$ mellett), és L_∞ távolságát: $\vec{x} = (5, 8, 3, 5)$ és $\vec{y} = (6, 13, 4, 2)$

Megoldás

$$\vec{z} = \vec{x} - \vec{y} = (-1, -5, -1, 3).$$
$$L_1(\vec{x} - \vec{y}) = L_1(\vec{z}) = |-1| + |-5| + |-1| + |3| = 10.$$
$$L_2(\vec{x} - \vec{y}) = L_2(\vec{z}) = \sqrt{(-1)^2 + (-5)^2 + (-1)^2 + (3)^2} = 6.$$
$$L_4(\vec{x} - \vec{y}) = L_4(\vec{z}) = (|-1|^4 + |-5|^4 + |-1|^4 + |3|^4)^{1/4} = \sqrt[4]{708}.$$
$$L_\infty(\vec{x} - \vec{y}) = L_\infty(\vec{z}) = \max\{|-1|, |-5|, |-1|, |3|\} = 5.$$

Intervallum attr. folytatása

- Fontos a mértékegység!
- Nem mindegy, hogy méter vagy milliméter…
- Normalizálás -> [0,1] intervallumba
- Súlyozás is fontos: két ember összehasonlításánál fontosabb a hajszín, mint a lábfej nagysága

$$d(x,y) = \sqrt{w_1|x_1 - y_1|^2 + w_2|x_2 - y_2|^2 + \cdots + w_m|x_m - y_m|^2},$$

- Nem lineáris lépték: algoritmusok futási ideje
- Más megközelítések szükségesek: intervallum hasonlóság, sorrendi megközelítés
- Példaul: két algoritmus hasonlósága: 2-szerese a másiknak

Vegyes attribútumok távolsága

- Két objektum más-más attribútumokkal
- Megoldás:
- Csoportosítsuk az attribútumokat típusuk szerint (n darab csoport)
- Határozzuk meg típusok szerint a távolságokat!
- Képezzük [0,1] intervallumba a távolságokat!
- Minden csoportnak feleltessünk meg egy-egy dimenziót a térben!
- Így egy vektort kapunk! (n dimenziós)
- Távolság = vektor hossza
- Célszerű súlyozni

Tartalom

- Történeti áttekintés, adatok rendelkezésre állása
- Attribútumok típusai, tulajdonságai
- Távolsági függvények
- Előfeldolgozás
  - Hiányzó értékek kezelése
  - Attribútum transzformációk
  - Adatok torzítása
  - Diszkretizálás
  - Normalizálás
  - Mintavételezés
  - Dimenzió csökkentés
  - Duplikátumok kiszűrése
  - Aggregáció
  - Monotonizáció

Előfeldolgozás - Hiányzó értékek kezelése

- Attribútumok hiányoznak -> pl. orvosi lap: dohányzásról való leszokás ideje
- **Törlés?** Lecsökkenhet az adatbázis mérete…
- Értékes adatok veszhetnek el
- **Hiányzó értékek feltöltése?**
- Alapértelmezett értékekkel, módusszal (kategória attr.)
- Új objektum súlyozással
- Medián, Átlag (intervallum attribútum)

Hiányzó értékek kezelése

- Osztályozó és regressziós algoritmusok
- Hasonló objektumok keresése
- x objektum hasonló y-hoz (donor)
- x-nek hiányzik A értéke, y-nak ismert
- x-nek A értéke y A attribútuma
- Milyen gyakran lehet donor y? Ha túl sokszor, akkor az probléma lehet

Tartalom

- Történeti áttekintés, adatok rendelkezésre állása
- Attribútumok típusai, tulajdonságai
- Távolsági függvények
- Előfeldolgozás
  - Hiányzó értékek kezelése
  - Attribútum transzformációk
  - Adatok torzítása
  - Diszkretizálás
  - Normalizálás
  - Mintavételezés
  - Dimenzió csökkentés
  - Duplikátumok kiszűrése
  - Aggregáció
  - Monotonizáció

Attribútum transzformációk

- Létrehozása
- Törlése
- Lényegtelenek felismerése

Attribútum transzformációk:
Új attribútum beszúrása

- Példa: testtömeg és magasság adott
- Betegség vizsgálatánál a testtömeg index a fontos
- Apriori tudás: testtömegindex (ezt az osztályozó algoritmus előtt adjuk meg!)
- Osztályozónak így javul a teljesítménye
- Segítsünk az osztályozónak!

Attribútum transzformációk:
Attribútumok törlése

- Sok felesleges attribútum -> ZAJ
- Döntési fa: kihagyhatja a felesleges attribútumokat (de sok zajjal nem hatákony)
- Döntési fa: nem csodamódszer!
- Szükség van a törlésre! -> segítünk a további műveleteket

Attribútum transzformációk:
Lényegtelen attr. törlése

- Két csoport: „filter” és a wrapper
- FILTER: külső kritérium alapján (osztálycímke, más attr.-kal való korreláció stb.) értékeli az egyes attribútumot -> csak a relevánsokat tartjuk meg
- WRAPPER: osztályozó algoritmus, különböző attribútum halmazokkal. Bevonjuk így az osztályozó algoritmusokat! Legjobb eredményt tartjuk meg!

Tartalom

- Történeti áttekintés, adatok rendelkezésre állása
- Attribútumok típusai, tulajdonságai
- Távolsági függvények
- Előfeldolgozás
  - Hiányzó értékek kezelése
  - Attribútum transzformációk
  - Adatok torzítása
  - Diszkretizálás
  - Normalizálás
  - Mintavételezés
  - Dimenzió csökkentés
  - Duplikátumok kiszűrése
  - Aggregáció
  - Monotonizáció

Adatok torzítása

- Motiváció: titkosítás, konkurencia elleni elrejtés, adott módszer mennyire érzékeny a zajra, magánszemélyek védelme
- AOL, 2006: botrány
- Probléma: keresőknek személyes adatokat adunk meg
- Ad-hoc ötletek nem jók
- Új kutatási témakor: **bizonyítható biztonság**

Tartalom

- Történeti áttekintés, adatok rendelkezésre állása
- Attribútumok típusai, tulajdonságai
- Távolsági függvények
- Előfeldolgozás
  - Hiányzó értékek kezelése
  - Attribútum transzformációk
  - Adatok torzítása
  - Diszkretizálás
  - Normalizálás
  - Mintavételezés
  - Dimenzió csökkentés
  - Duplikátumok kiszűrése
  - Aggregáció
  - Monotonizáció

Diszkretizálás

- Numerikus attribútum kategória típusúvá alakítása
- Értékkészletet intervallumokra osztjuk, ezekhez kategóriát rendelünk
- Minden intervallumhoz kategóriát rendelünk
- Információvesztés, de segíti az algoritmusokat
- PKI: egyenlő hosszú intervallumok -> adatpontok négyzetgyöke a kategória

1R módszer

- Egyszerű osztályozó módszer diszkretizálással
- Pl.: tanítóminta, hőmérsékletek Fahranheitban mérve, adott osztályokkal (rendezve)

64	65	68	69	70	71	72	72	75	75	80	81	83	85
1	0	1	1	1	0	0	1	1	1	0	1	1	0

64	65	68	69	70	71	72	72	75	75	80	81	83	85
1	0	1	1	1	0	0	1	1	1	0	1	1	0
1	0		1		0			1		0		1	0

- Határok: $64.5,\ 66.5,\ 70.5,\ 72,\ 77.5,\ 80.5,\ 84.$

Tartalom

- Történeti áttekintés, adatok rendelkezésre állása
- Attribútumok típusai, tulajdonságai
- Távolsági függvények
- Előfeldolgozás
  - Hiányzó értékek kezelése
  - Attribútum transzformációk
  - Adatok torzítása
  - Diszkretizálás
  - Normalizálás
  - Mintavételezés
  - Dimenzió csökkentés
  - Duplikátumok kiszűrése
  - Aggregáció
  - Monotonizáció

Normalizálás

- Attribútum elemeit másik intervallum elemeivel helyettesítjük
- Eloszlás ugyanaz marad!
- Min-Max normalizálás:

$$a'_j = \frac{a_j - min_A}{max_A - min_A},$$

- Standard normalizálás:

$$a'_j = \frac{a_j - \bar{A}}{\sigma_A},$$

ahol $\bar{A}$ az A attribútum átlaga, σ_A pedig a szórása.

Tartalom

- Történeti áttekintés, adatok rendelkezésre állása
- Attribútumok típusai, tulajdonságai
- Távolsági függvények
- Előfeldolgozás
  - Hiányzó értékek kezelése
  - Attribútum transzformációk
  - Adatok torzítása
  - Diszkretizálás
  - Normalizálás
  - Mintavételezés
  - Dimenzió csökkentés
  - Duplikátumok kiszűrése
  - Aggregáció
  - Monotonizáció

Mintavételezés

- Rengeteg adat -> lassú feldolgozás
- Mintavételezés az adatokból -> gyorsabb feldolgozás
- Hátrány: nem elég pontos az eredmény

Mintavételezés: Statisztika vs Adatbányászat

- STATISZTIKA
- Teljes adathalmaz *megfigyelése* túl drága lenne vagy nem kivitelezhető

- ADATBÁNYÁSZAT
- Rendelkezésre állnak az adatok
- Óriás méretű adathalmazok -> túl költséges, túl sok idő

Mintavételezés: Csernov-korláttal

- **Elemek előfordulásának valószínűségének közelítése a relatív gyakorisággal**
- Előfordulása: gyakori minták, asszociációs szabályoknál
- Adott egy minta
- X elem előfordulásának valószínűsége p
- M méretű megfigyelés(minta) áll rendelkezésre
- Mintavételezés hibázik:

$$\text{hiba}(m) = \mathbb{P}\Big(\big|\text{rel. gyakoriság}(x) - p\big| \geq \epsilon\Big)$$

Mintavételezés: Csernov-korláttal 2.

- Xi val.változó értéke 1, ha x-et választottuk i. húzásnál, különben 0

  $Y = \sum_{i=1}^{m} X_i$.

- Húzások egymástól függetlenek -> Y egy m,p paraméterű binomiális eloszlás

$$\begin{aligned}\text{hiba}(m) &= \mathbb{P}\left(\left|\frac{Y}{m} - p\right| \geq \epsilon\right) = \mathbb{P}\left(\left|Y - m \cdot p\right| \geq m \cdot \epsilon\right) \\ &= \mathbb{P}\left(\left|Y - \mathbb{E}[Y]\right| \geq m \cdot \epsilon\right) \\ &= \mathbb{P}\left(Y \geq m \cdot (\mathbb{E}[X] + \epsilon)\right) + \mathbb{P}\left(Y \leq m \cdot (\mathbb{E}[X] - \epsilon)\right)\end{aligned}$$

- AH OL: m*p a bin. eloszolás várható értéke

Mintavételezés: Csernov-korláttal 3.

- Csernov korlát (Hoeffding speciális esete):

$$\mathbb{P}\Big(Y \geq m \cdot (\mathbb{E}[X] + \epsilon)\Big) \leq e^{-2\epsilon^2 m}$$

$$\mathbb{P}\Big(Y \leq m \cdot (\mathbb{E}[X] - \epsilon)\Big) \leq e^{-2\epsilon^2 m}$$

- EBBŐL megkapjuk: $\text{hiba}(m) \leq 2 \cdot e^{-2\epsilon^2 m}$

- Ha hibakorlát DELTA, akkor: $m \geq \frac{1}{2\epsilon^2} \ln \frac{2}{\delta}.$

Mintavételezés: Csernov-korláttal 4.

- Jelentése: 27000 méretű minta, 1% az esélye, hogy a relatív gyakoriság és a valószínűség közti különbség 0.01-nél nagyobb

ϵ	δ	m
0.05	0.01	1060
0.01	0.01	27000
0.01	0.001	38000
0.01	0.0001	50000
0.001	0.01	2700000
0.001	0.001	3800000
0.001	0.0001	5000000

Mintavételezés: mintaméret becslése

- Túl pesszimista eljárás a Csernov-féle
- Kevesebb méretű minta is elég
- Keressünk más eljárásokat!

- Csernovnál jobb, ha a hibát a valószínűség és a relatív előfordulás hányadosából származtatjuk és binomiális eloszlást használunk.
- Binomiális sem a legjobb -> hipergeometrikus
- Részletek: Bodon-Buza könyvben

Arányos mintavételezés

- Előzőleg: véletlenül választottuk az elemeket
- Nem kell véletlennek lenni! Reprezentatív legyen!
- Reprezentatív, ha a mintán végzett elemzés ugyanazt adja, mintha az egész mintával dolgoztunk volna!
- Például: eredeti adatbázisban osztályokra vannak osztva az objektumok. A mintában is ugyanannak az aránynak kell megmaradni!

Tartalom

- Történeti áttekintés, adatok rendelkezésre állása
- Attribútumok típusai, tulajdonságai
- Távolsági függvények
- Előfeldolgozás
  - Hiányzó értékek kezelése
  - Attribútum transzformációk
  - Adatok torzítása
  - Diszkretizálás
  - Normalizálás
  - Mintavételezés
  - Dimenzió csökkentés
  - Duplikátumok kiszűrése
  - Aggregáció
  - Monotonizáció

Sokdimenziós adatok, dimenziócsökkentés

- Sok attribútum = sok dimenzió
- PL: 1 objektum= 1 szöveg
- Minden attribútum 1-1 szó előfordulásának száma
- Ez így több ezer attribútum…

Dimenzióátok

- Rengeteg attribútum
- Könnyebb dolga van az algoritmusoknak? NEM!
- Korreláció, sok irreleváns attribútum…
- Dimenzióátok: sok dimenziós adatok bányászatának problémája
- Dimenziónövekedés -> sűrűség csökkenés
- PL: 1000 db kétdimenziós objektum. Átlagosan: 1000/10/10=10 db egy egységbe

  Ugyanez száz dimenzióban 10^{-97} …

Ez baj, mert a klaszterező algoritmusok a sűrűsé alapján végzik a becslésüket.

Dimenzióátok: távolságok koncentrációja

- Adott d dimenziós tér
- Generáljuk véletlenszerűen pontokat
- Legyen l_d legtávolabbi pontok különbsége
- $l_d' = l_d$ / (legközelebbi pontok távolsága)
- l_d' 0-hoz tart d növekedésével
- Következtetés: távolságfogalom d növelésével egyre inkább nem használható

Dimenziócsökkentő eljárások

- Akkor jó, ha hasonlóságtartó az eljárás, vagyis jó becslése a csökkentett dimenziójú objektum az eredetinek
- Eredeti halmaz: m x n –es M mátrix
- Új halmaz: m x k –ás M’ mátrix
- n>>k, ahol n és k az attribútumok számát jelöli
- Elfér a memóriába, könnyebb vizualizálni: két, háromdimenziós adatokat sokkal könnyebb ábrázolni
- Csak a legfontosabb attribútumokat tartjuk meg, vagyis zajszűrésnek tekinthető az eljárás

Szinguláris eljárás

- Klasszikus lineáris algebrai alapú
- M’ mátrix soraiból jól közelíthető az Euklédeszi távolság
- Illetve az attribútumok értékeiből számított skaláris szorzattal mért hasonlóság
- Részletek: Bodon-Buza 88. oldal

MDS

- Objektumok közti távolságok távolság mátrix-szal reprezentálva
- Cél: csökkenteni a dimenziót őgy, hogy az objektumok páronkénti távolsága minél jobban közelítsen az eredeti értékhez
- MDS célfüggvényt definiál, melyet optimalizál
- Sok esetben használható (ahol a távolság definiálható): sztringek, idősorok stb.
- Célfüggvény:

$$\text{stressz} = \sqrt{\frac{\sum_{i=1}^{n}\sum_{j=1}^{n}(d'_{i,j} - d_{i,j})^2}{\sum_{i=1}^{n}\sum_{j=1}^{n} d_{i,j}^2}}$$

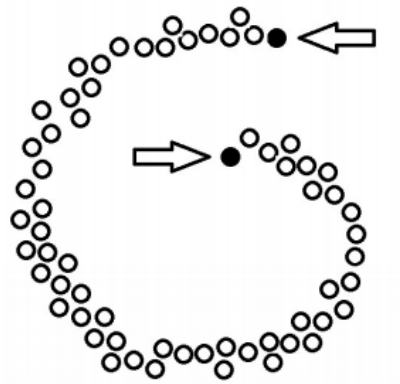
Ahol $d_{i,j}$ i és j objektum eredeti távolsága, $d'_{i,j}$ leképzés utáni távolság és n az adatbázisbeli objektumok száma

Algoritmus: stressz értékét csökkenti, a kisdimenziós térben elhelyezett objektumok mozgatásával

Részletek: http://en.wikipedia.org/wiki/Multidimensional_scaling

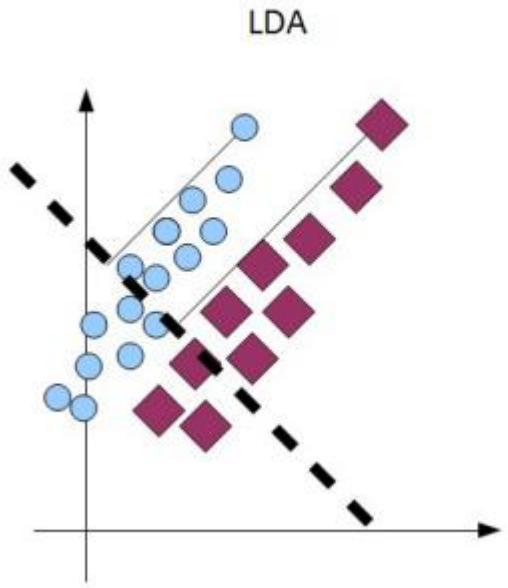
ISOMAP

- Annyiban különbözik az MDS-től, hogy mit tekint $d_{i,j}$ távolságnak a stressz függvény számításakor
- Távolságok valamely távolsági függvény alapján (pl. Euklédeszi)
- Távolságok alapján egy szomszédossági mátrix: minden objektumot összeköt a k darab legközelebbi szomszédjával
- Ezt követően kiszámolja az objektumok közti legközelebbi szomszéd gráfbeli legrövidebb utak hosszát: a $d_{i,j}$ távolság tehát az i és j objektumok közti legközelebbi szomszéd gráfbeli legrövidebb út hossza lesz.
- Így a két pont távolsága a csigavonal mentén definiálódik.



LDA

- Osztálycímkékkel rendelkezünk
- Osztályozási feladatokhoz
- 1 dimenziósra is csökkenthetjük az objektumot
- Figyelembe veszi az osztálycímkéket és olyan irányt keres, amelyre vetítve az osztályok minél jobban elkülönülnek



Minimash

- Sorok = attribútumok
- Oszlopok = objetumok
- Cél: attribútumok (jelen esetben sorok) csökkentése
- Használják: weboldalak kiszűrésénél, koppintások, kalózmásolatok felderítésénél, hasonló tulajdonságú felhasználók keresésénél
- Részletek: Bodon-Buza 92. oldal

Tartalom

- Történeti áttekintés, adatok rendelkezésre állása
- Attribútumok típusai, tulajdonságai
- Távolsági függvények
- Előfeldolgozás
  - Hiányzó értékek kezelése
  - Attribútum transzformációk
  - Adatok torzítása
  - Diszkretizálás
  - Normalizálás
  - Mintavételezés
  - Dimenzió csökkentés
  - Duplikátumok kiszűrése
  - Aggregáció
  - Monotonizáció

Duplikátumok kiszűrése

- Egy adat kétszer lett felvéve, két adatforrás is tartalmazza stb.
- Pl: tévedésből egy ügyfelet többször vettünk fel
- Cél: egy objektum csak egyszer szerepeljen

- Legegyszerűbb eset: minden attr. Megegyezik, akkor szűrünk -> erőforrás igényes: $O(n^2)$
- Rendezzük sorba! Pl. gyorsredenzés: O(n*logn)+ 1 végigolvasás

- Közelítőleg egyezők keresése: nagy kihívás
- Pl. két embert kétszer vettek fel, de másodjára elfelejtették a szül. helyét felvenni vagy több helyről integrált leírások pl: két webáruházból integrált termékleírás

Sorted neighborhood technika

- Feltételezünk a duplikátumok szűrésére létezik egy szabályrendszert -> ez természetesen alkalmazásfüggő
- Minden objektumhoz egy kulcsértéket rendelünk
- Karakterlánc = kulcsérték <- attribútumok alapján
- Rendezzük kulcsérték szerint
- Jó kulcsérték választás -> duplikátumok közel kerülnek
- Végigolvasás, s legfeljebb W távolságra lévő objektumokra alkalmazzuk a szabályrendszert
- Lelke: jó kulcsképzési szabály
- Célszerű több kulcsképzési szabályt alkalmazni

Regressziós és osztályozó modellek a duplikátumok szűrésére

- Nem mindig létezik szabályrendszer a duplikátumok szűrésére…
- Néhány száz objektumra definiáljuk, hogy duplikátum-e
- Ezek alapján osztályozó, regressziós eljárás -> később

Tartalom

- Történeti áttekintés, adatok rendelkezésre állása
- Attribútumok típusai, tulajdonságai
- Távolsági függvények
- Előfeldolgozás
  - Hiányzó értékek kezelése
  - Attribútum transzformációk
  - Adatok torzítása
  - Diszkretizálás
  - Normalizálás
  - Mintavételezés
  - Dimenzió csökkentés
  - Duplikátumok kiszűrése
  - Aggregáció
  - Monotonizáció

Aggregáció

- Adatok csökkentésére
- Mintázatok felismerésére
- Például: bolthálózat -> naponkénti, hetenkénti, ügyfelenkénti, körzeti vetítés, csoportosítás stb. vagy négyzetkilométerre csapadék adatok, a szomszédos cellákat összevonjuk
- Túl részletes eredmények -> vmely szempont alapján összevonjunk
- Rosszul megválasztott aggregáció -> mintázatok, összefüggések elvesznek

Tartalom

- Történeti áttekintés, adatok rendelkezésre állása
- Attribútumok típusai, tulajdonságai
- Távolsági függvények
- Előfeldolgozás
  - Hiányzó értékek kezelése
  - Attribútum transzformációk
  - Adatok torzítása
  - Diszkretizálás
  - Normalizálás
  - Mintavételezés
  - Dimenzió csökkentés
  - Duplikátumok kiszűrése
  - Aggregáció
  - Monotonizáció

Monotonizáció

- Osztály attribútum gyakran ordinális
- Például: ügyfelek kockázata
- Monoton klasszifikációs algoritmusok
- Teljesülni kell:

$$a_x^1 \leq a_y^1 \wedge a_x^2 \leq a_y^2 \wedge \ldots \Rightarrow c_x \leq c_y.$$

- Attribútumokból következtet
- Például: emberekről tároljuk: magasság, kor, tömeg, alkohol és cigi adatokat
- Cigi, alkohol monoton, de magasság, tömeg nem!!!
- Magasság, tömeg → testtömeg index, ez már monoton