Adatbányászat

Szegedi Tudományegyetem

Követelmények – gyakorlat

- Gyakorlaton elérhető: max. 30 pont
- Szorgalmi időszak 3. hetétől kezdődően az utolsó előtti hétig minden héten tantermi Coospace teszt: 3 pont/teszt
- A 10 legjobb teszteredmény számít be
- Sikeres a gyakorlat, ha legalább 7 teszt pontszáma > 0, az összpontszám pedig ≥ 15
- Javító ZH az egész éves anyagból

Végső jegy	
[0 – 15)	elégtelen
[15 – 20)	elégséges
[20 – 24)	közepes
[24 – 27)	jó
[27 – 30)	jeles

Követelmények – kollokvium

- Kollokviumra a gyakorlat sikeres teljesítését követően lehet jelentkezni
- Szóbeli/írásbeli vizsga
- A szorgalmi időszak utolsó hetén elővizsga tehető

Egyebek

- Projektlehetőségek a mestint tanszéken

A félév felépítése

- Bevezetés, ismerkedés, leíró statisztika
- Adatok leírása, előfeldolgozása (adattisztítás,- és redukálás)
- Felügyelet nélküli módszerek az adatbányászatban
- Felügyelt módszerek az adatbányászatban (osztályozás)
- Kívülálló egyedek tanulása, eltéréselemzés (outlier detection)
- Gyakori mintázatok, asszociációs szabályok tanulása
- Gráf alapú módszerek az adatbányászatban
- Web méretű adatbányászat, szövegbányászat

Ajánlott irodalom

- A kurzus (angol nyelvű) jegyzete
- H. Witten, E. Frank, M. A. Hall: Data Mining: Practical Machine Learning Tools and Techniques
- P. Tan, M. Steinbach, V. Kumar: Introduction to Data Mining
- **J. Leskovec, A. Rajaraman, J. Ullman: Mining Massive Datasets**
- B. Liu: Web Data Mining
- C. Manning, P. Raghavan, H. Schütze: Introduction to Information Retrieval
- R. Duda, P. Hart, D. Stork: Pattern Recognition
- C. Bishop: Pattern Recognition and Machine Learning
- ...
- Bodon Ferenc, Buza Krisztián: Adatbányászat jegyzete

Motiváció

- 1 millió dolláros NetFlix Prize
- 20 ezer dolláros Kaggle StackOverflow Challange
- Facebook miatt bukhatjuk a hitelt (Index)
- Warren Buffet March Madness-szel kapcsolatos ajánlata (10^9\$)
- NSA's SKYNET program may be killing thousands of innocent people
- Kedvcsináló: $\|M\|_F^2 = \sum_i \sum_j \sum_k \sum_l \sum_m \sum_n p_{ik} q_{kl} r_{lj} p_{in} q_{nm} r_{mj}$

Motiváció képben

miközben tudásra szomjazunk

About 51,800 results (0.58 seconds)

Vásárlás: Az emberiség megoldatlan rejtélyei (2004)
konyv.arukereso.hu/.../termek-leiras.html - Cached - Translate this page
\'Belefulladunk az információkba, **miközben tudásra szomjazunk**\' - állapította meg az amerikai jövőkutató, John Naisbitt, arra utalva, hogy a tudás sokkal több, ...

"Belefulladunk a temérdek információba, **miközben** tudásr... - Yamm!
yamm.hu/.../11243096077-belefulladunk-a... - Cached - Translate this page
"Belefulladunk a temérdek információba, **miközben tudásra szomjazunk**." (Rutherford D. Rogers). több mint 2 éve Válasz RT Lájk ...

Flammich Mariann: Hackerek (Médiakutató)
www.mediakutato.hu/cikk/...03.../06.html - Cached - Translate this page
Az iskolában bébiételt tömnek a szánkba, **miközben** pedig steakre éhezünk [...] a hús, amit ... Mi **tudásra szomjazunk** és te bűnözőnek hívsz minket. Köztünk ...

Adatbányászat területei

- Üzleti alkalmazások
  - Hitelelemzés
  - Vásárlói csoportok szegmentálása
  - Ajánlórendszerek
  - Elvándorlás/lemorzsolódás (churn) analízis
- Tudományos alkalmazások
  - Csillagászat
  - Gyógyszerkutatás
  - Orvosi diagnosztika

Mi tehát az adatbányászat?

- Hasznos, (gyakran) nem várt mintázatok felismerése nagy mennyiségű adathalmazból (pl. Web)
- Előretörését a technológia (háttértárak, memória, processzor, GPGPU, MapReduce) fejlődés tette lehetővé
- A segítségével megszerzett tudás legyen érthető, érvényes, hasznos és újszerű
- $\approx$ tudásfeltárás (Knowledge Discovery)

Mi nem adatbányászat?

- Adatbázis lekérdezések
- "Egyszerű" statisztika (eszköz ugyanakkor lehet)
- Bonferroni-elv: sok adatot figyelembe vétele mellett véletlen események is látszólag érdekessé válhatnak

Total Information Avareness program

- Nagy Testvér bevezetése Amerikában
- Gondolatkísérlet: gonosztevők keresése szállodai bejelentkezések alapján
- 10^9 ember 0,01 valószínűséggel száll meg egy nap valamely 100 férőhelyes szállodában $\rightarrow$ 100000 ($10^9 \cdot 0,01/100$) szálloda van
- 1000 nap alatt hány gyanús eset (2 ember 2 éjszaka ugyanabban a hotelben szállt meg) kerülne kiszűrésre?
- P(x és y ugyanott száll meg egy éjszaka)$= 0,01 \cdot 0,01 \cdot 0,00001 = 10^{-9}$
- P(x és y 2 alkalommal is ugyanott száll meg egy éjszaka) $= (0,01 \cdot 0,01 \cdot 0,00001)^2 = 10^{-18}$
- Lehetséges éjszaka-ember párosok száma = $\binom{10^9}{2} \cdot \binom{10^3}{2} \approx 2,5 \cdot 10^{23}$
- Gyanús párosok $\approx 2,5 \cdot 10^{23} \cdot 10^{-18} = 250000$

Rhine-paradoxon

- David Rhine parapszichológiai vizsgálatai
- Diákoknak 10 kártyalap színét (piros/kék) kellett "megjósolniuk"
- Eredmény: a megvizsgáltak közel 0,1 %-a parapszichológiai képességekkel bír (v.ö. 2^{-10})
- A visszahívott "parafenoménok" legközelebb már átlagos eredményt nyújtottak

Rhine konklúziója?

A parafenomének elvesztik különleges képességeiket, ha tudtukra juttatják azokat

Simpson-paradoxon

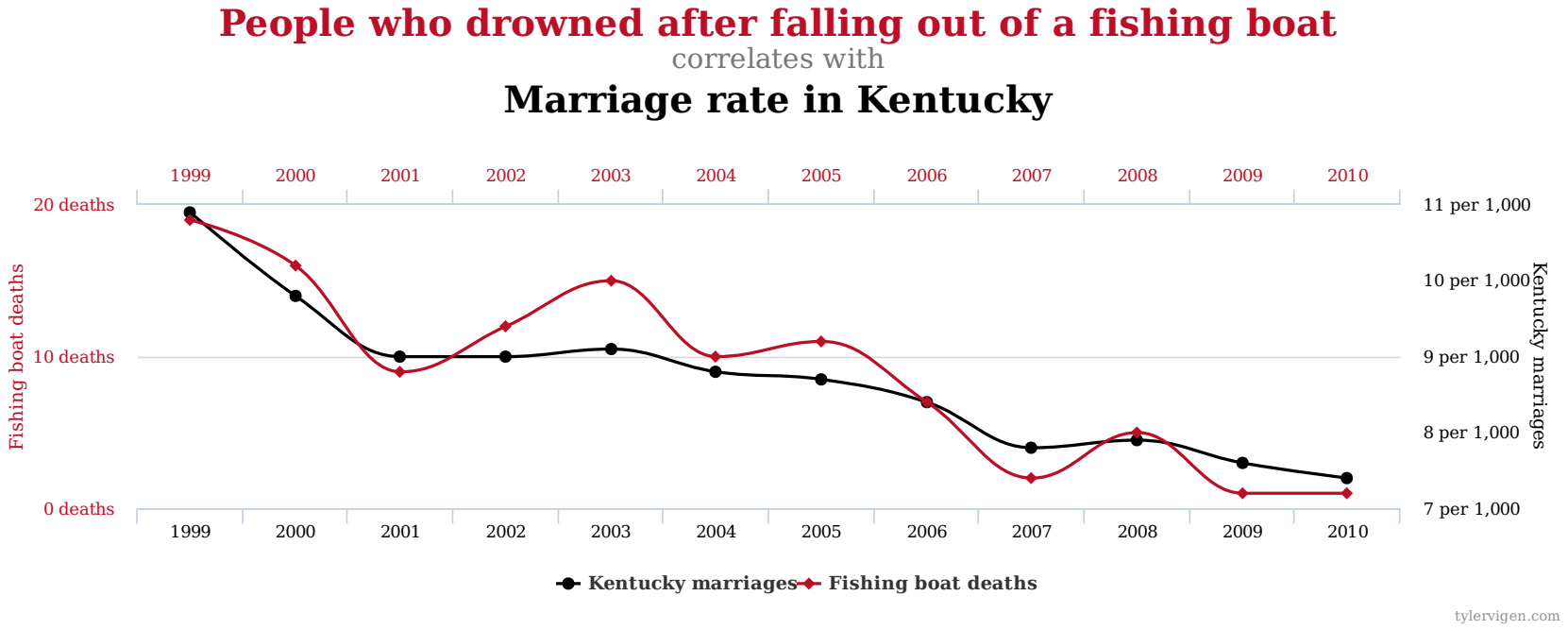
- Avagy az adatokból könnyebb hibás következtetésre jutni, mint azt gondolnánk

	Felvett/Jelentkezett	
	Lány	Fiú
A szak	7/100	3/50
B szak	91/100	172/200
Összesen	98/200	175/250

- Az aggregált adatok alapján úgy tűnik, a lányokat kisebb arányban vették föl a fiúknál (vö. 49% vs. 70%)
- A szakonkénti mutatók viszont épp ellenkező trendet mutatnak

Korreláció vs. kauzalitás

- Kentucky államban kötött házasságok és halászhajóból vízbefullók száma >0.95 korrelációt mutat [1]



[1] Forrás

Mindezek után, adatbányászat-e?

- Üzletlánc vásárlói adatbázisában szereplők átlagéletkorának meghatározása
- Honvédségi adatbázisban átlagos cipőméret meghatározása a hibás adatok kiszűrése mellet
- Tomográfiai leletekben elváltozásokra utaló jegyek keresése
- Választásra jogosultak nemek szerinti megoszlásának vizsgálata
- Választási hajlandóság/részvételi arány előrejelzése

Rokonterületek

- Matematika: valószínűségszámítás, statisztika, gráfelmélet, algebra, analízis
- Algoritmus-, bonyolultságelmélet
- Adatbáziskezelés (SQL és/vagy Solr, Elasticsearch, Kibana, ...)
- Gépi tanulás, mesterséges intelligencia

Eszközök, szoftverek

- Üzleti irányultságú célszoftverek (pl. SAS)
- Számos gépi tanuló, regressziós, transzformációs stb. módszert tartalmazó szoftvercsomagok (nem üzleti jellegű feladatokhoz)
  - Weka, MALLET
  - Clementine (SPSS Inc.), Intelligent Miner (IBM), DBMiner (Simon Fraser Univ.)
  - Octave, Matlab, Maple, R
  - Python (numpy, scipy, scikit-learn, pandas)
  - ...

Adatbányászat tárgya

- (Nagy méretű) **adathalmaz**ok, melyek (általában) sok **jellemző**vel leírt **adatobjektum**okból épülnek fel

Adatobjektum	**Jellemző**
rekord	mező
adatpont	dimenzió
minta/mérés	változó
egyed/eset/példány	jellemző, tulajdonság

- Dimenzionalitás átka (kissé pongyolán): a dimenziószám nagyságának növekedésével exponenciálisan növő adatpontra van szükség (a megfelelő pontosság eléréséhez)
- Nagydimenziós térben nehezen értelmezhető a távolság fogalma $\rightarrow$ diemnziócsökkentő eljárások (lásd később)

Adathalmazok formátumai

- Adatbázistáblák/sémák
  - Tranzakciós listák (kosarak)
- Adatmátrix
  - Előfordulási (pl. dokument-term) mátrix

Változók típusai mérési skálájuk szerint

Attribútumtípus		Leírás	Példák	Statisztikái
Kategória	Nominális	Két változó egyezését tudjuk csupán vizsgálni	városnév, hajszín	módusz, entrópia, korreláció, χ^2-teszt
	Ordinális	Két változó között értelmezhető a $>$ reláció	jegyek, {rossz, tűrhető, jó}	medián, percentilisek
Numerikus	Intervallum	Két változó különbsége értelmezhető	cipőméret, dátumok, °C	átlag, szórás, stat. tesztek
	Arányskála	Képezhető két érték hányadosa is	kor, hossz, Kelvin fok	százalék, harmonikus közép

Változók további csoportosítási lehetősége - Diszkrét és folytonos változók

- Diszkrét változó: véges vagy megszámlálhatóan végtelen értékkészlettel rendelkeznek
- Folytonos változó: értékkészlete a teljes valós számhalmaz
- Mérési skálák vs. értékkészlet
  - Nominális és ordinális skálájúak jellemzően diszkrétek
  - Intervallum és arányskálájúak jellemzően folytonosak
  - Bináris folytonos változó?
  - Diszkrét gyakoriségi értékek?

Változók további csoportosítási lehetősége - Szimmetria vs. aszimmetria

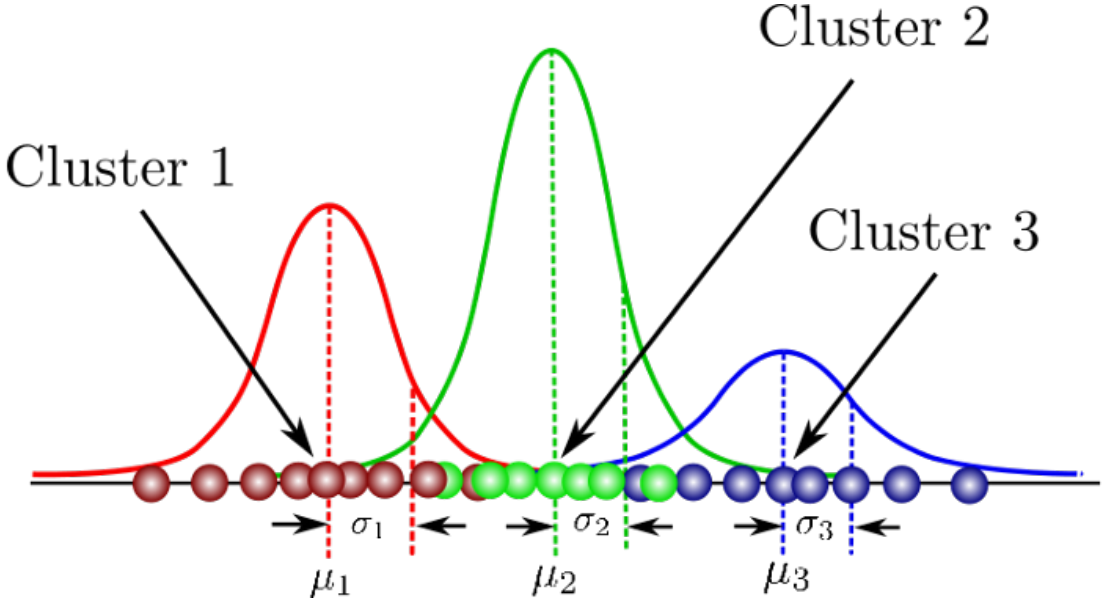
- Egy jellemző meg nem léte nem feltétlen azonos jelentőségű egy jellemző meglétével két egyed hasonlóságának vizsgálata kapcsán
- pl. ritka dokumentumvektorok, ahol jellemzően a teljes szótár elenyésző hányada szerepel a dokumentumokat reprezentáló vektorokban

Jellemzőmanipuláció

- Jellemződiszkretizáció
  - Bizonyos algoritmusoknak kedvez(het) sebesség és eredményesség szempontjából, ha numerikus helyett kategória típusú értékekkel dogozhatnak
- Jellemzőszűrés
  - Elméletben több jellemző segítségével mélyebb összefüggések tanulhatók, a gyakorlat azonban erre rá tud cáfolni
  - Próbáljuk meg a rendelkezésünkre álló jellemzők leghasznosabb részhalmazát kiválasztani (részhalmazokból azonban exponenciálisan sok van! $\rightarrow$ heurisztikák alkalmazása)

Irányítatlan (felügyeletlen) diszkretizáció

- Egyedek osztálycímkéjének ismerete nélkül próbáljuk meg jellemzőiket diszkretizálni
  - Fix szélességű intervallumok kialakítása
    - Gyér intervallumok alakulhatnak ki
    - Jellemzőértékek sűrű régiói kerülhetnek szeparálásra
  - Fix (egyed)számossággal rendelkező intervallum kialakítása: minden intervallumba ugyanannyi egyed essen ("sík" hisztogram)
  - Sűrűség alapú módszerek (pl. Gaussian Mixture Model)



Irányított (felügyelt) diszkretizáció

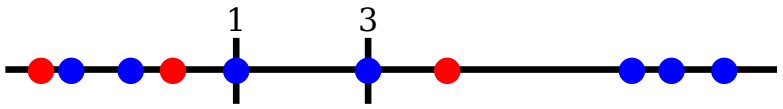
- A jellemzők értékkészleteinek particionálása során figyelembe veszi a jellemzőértékek osztálycímkékkel való együttállását
- Történhet pl. kölcsönös információ, információnyereség (information gain), χ^2, ... mutató alapján

Emlékeztető

X és Y véletlen változók kölcsönös információtartalma:
$MI(X;Y) = H(X) - H(X|Y) = \sum_{x \in X} \sum_{y \in Y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)}$
Mennyivel csökken X véletlen változó értékére vonatkozó bizonytalanságunk, ha Y értékét megismerjük?

Példa kölcsönös információtartalom alapú diszkretizálásra

Tegyük föl, hogy 10 megfigyelésünk van, amelyek mindegyike a pozitív (P) vagy negatív (N) osztályok valamelyikébe tartozik. Az X véletlen változó diszkretizálását az $X = 1$ vagy $X = 3$ érték mellett éri-e meg jobban elvégezni?



	P	N
$X \leq 1$	2	3
$X > 1$	1	4

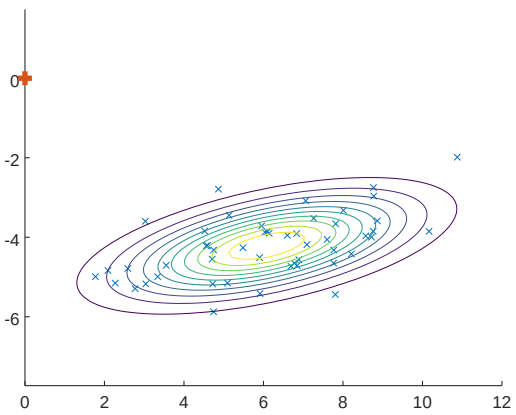
	P	N
$X \leq 3$	2	4
$X > 3$	1	3

$$\frac{2}{10}\log_2\frac{4}{3} + \frac{3}{10}\log_2\frac{6}{7} + \frac{1}{10}\log_2\frac{2}{3} + \frac{4}{10}\log_2\frac{8}{7} \approx 0.035 > \frac{2}{10}\log_2\frac{10}{9} + \ldots \approx 0.006$$

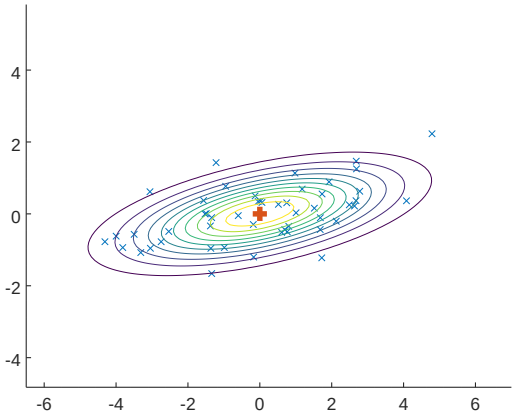
Azaz az $X = 1$ mentén történő diszkretizáció preferálandó az $X = 3$ mentén történő diszkretizációval szemben

Folytonos változók előfeldolgozása – centralizálás

- Minden megfigyelésből vonjuk ki az átlagos megfigyelés értékét
- Az eredeti változók így az átlagtól vett eltérésüket fogják mutatni (a jellemzők várhatóértéke 0 lesz)



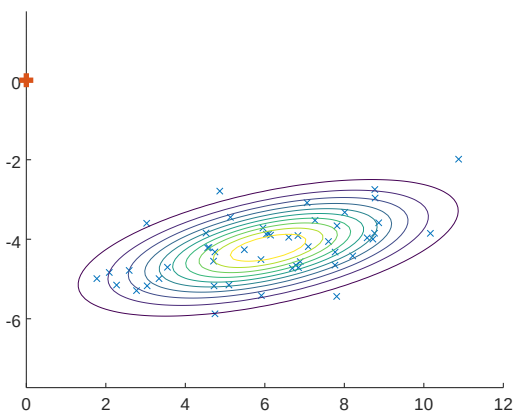
(a) „Nyers” adat



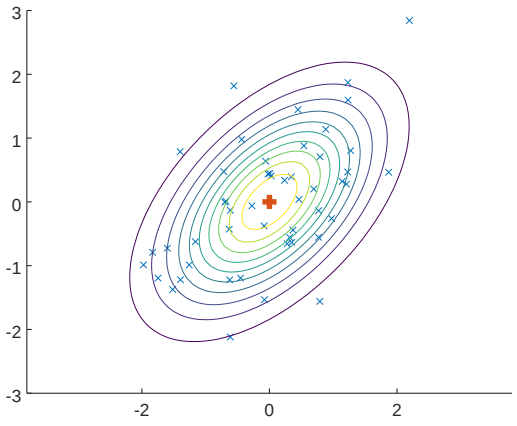
(b) Centralizált adat

Folytonos változók előfeldolgozása – standardizálás

- Fejezzük ki a változókat a statisztikából ismert z értékükkel
  - A megfigyelt érték a szórás hányszorosával (és milyen irányban) tér el az átlagos megfigyeléstől



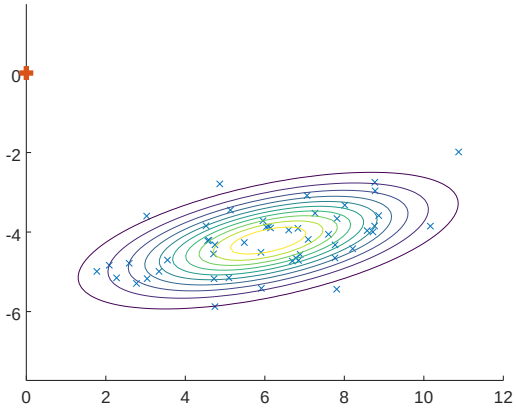
(a) „Nyers" adat



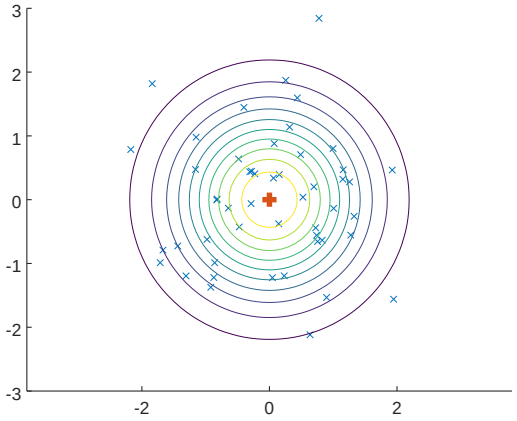
(c) Standardizált adat

Folytonos változók előfeldolgozása – fehérítés

- Tüntessük el a jellemzők között fennálló korrelációt
  - A Σ kovarianciamátrixszal bíró (centralizált) X megfigyelésmátrix korrelálatlanná tehető XL transzformációval
    - L legyen olyan, hogy arra $\Sigma^{-1} = LL^{\mathsf{T}}$ teljesüljön



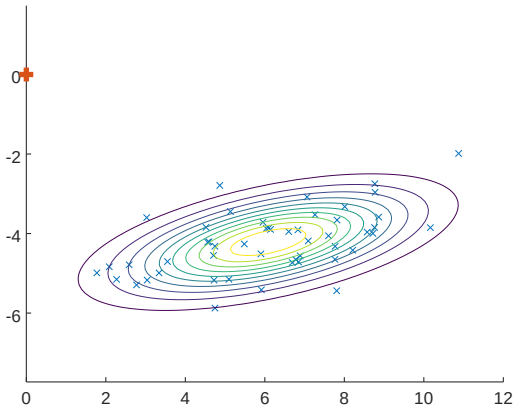
(a) „Nyers” adat



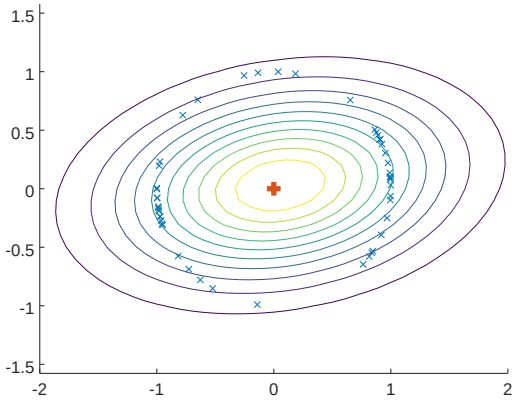
(d) Fehérített adat

Adatpontok normalizálása – egységnyi hosszúvá tevés

- A normalizálás a jellemzőkön túl a megfigyelések tekintetében is történhet
  - A többdimenziós $x \neq 0$ megfigyelésvektor egységnyi hosszúvá tehető $x_u = \frac{1}{\sqrt{x^T x}} x$ módon

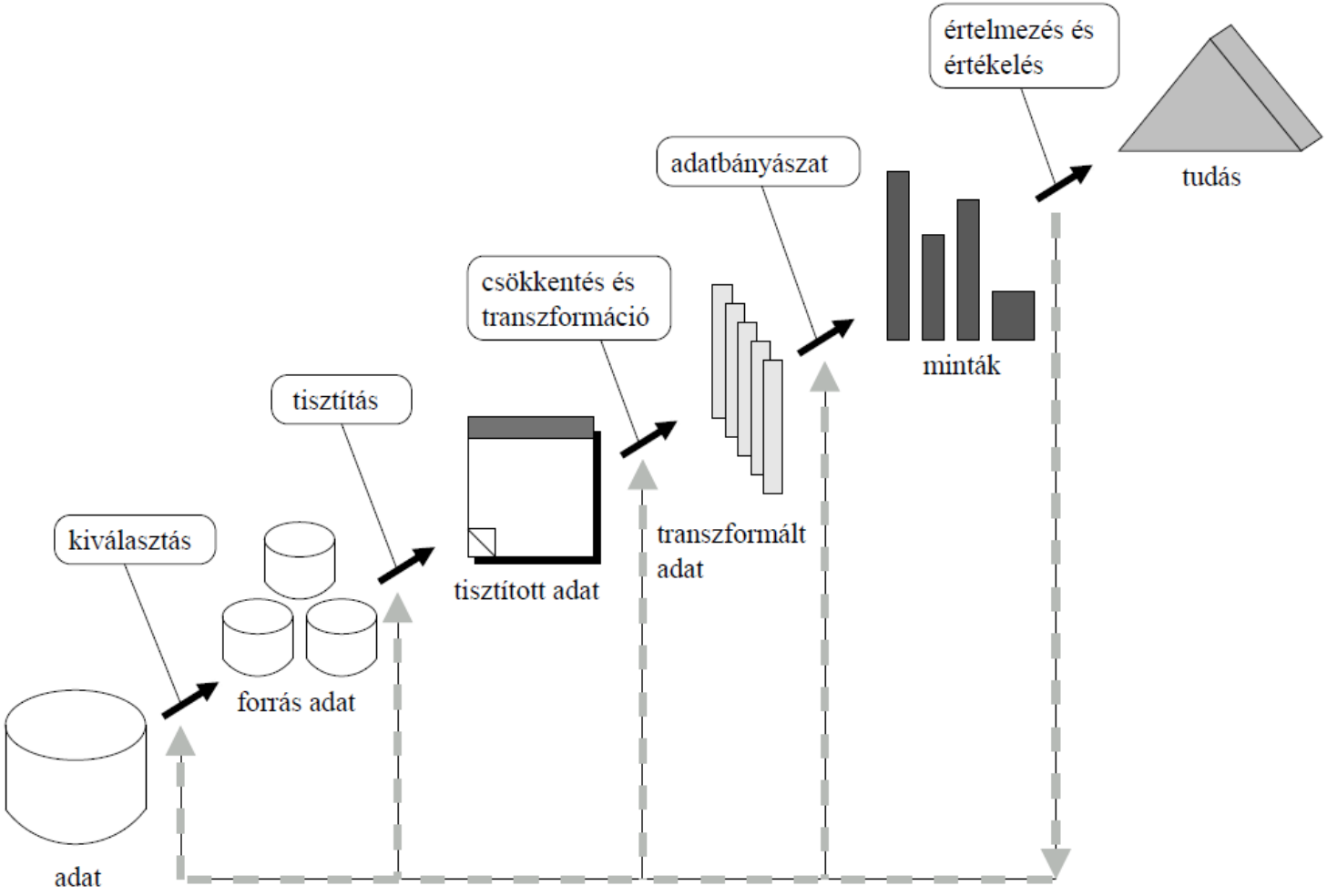


(a) „Nyers" adat



(e) Egységnyi hosszú (és centralizált) adat

A tudásfeltárás folyamata



Adatminőség

- Minden adatbázisban találhatók hibák, inkonzisztenciák
  - adatgyűjtés során
  - adatbevitel során
  - mérési hiba

Adattisztítás

- Hiányzó adatok kezelése pl. $x_i = (0, 0, 4, 2, ?, ?, 1, 6, ' True')$
  - hiányzó adat becslése (pl. k-"leghasonlóbb"/teljes sokaság móduszával/átlagával/mediánjával, esetleg Expectation Maximization)
  - adatsor elhagyása
  - jellemző elhagyása a modellépítésből
- Zajos adatok kezelése (pl. `életkor`=280)
  - Hiányzó adatokhoz hasonlóan
- Duplikátum egyedek kiszűrése

Véletlen változók

- Kísérlet leírására szolgálnak
- n kísérlet/mérés eredménye: $x_1, x_2, \ldots, x_n$
- Várható érték: $\mu_X = \mathbb{E}[X] = \sum_{x \in X} P(X = x) * x$
- Variancia: Átlagtól való átlagos eltérés
- $Var(X) = \sum_{x \in X} P(X = x) * (x - \mu_X)^2 = \mathbb{E}[(x - \mu_X)^2] = \mathbb{E}[X^2] - \mathbb{E}^2[X]$ (biz.: táblán)

Példa

X=[1,4,7]
$\mathbb{E}[X] = (1 + 4 + 7)/3 = 4$
$\mathbb{E}[X^2] = (1 + 16 + 49)/3 = 22$
$Var(X) = 22 - 4^2 = 6$

Egy kis algebra (ami még jól fog jönni)

- Euklideszi távolság: $\|\mathbf{a}\|_2 = \sqrt{\sum_{i=1}^{d} a_i^2}$
- Belső szorzat, skalárszorzat: $\mathbf{a}^\intercal \mathbf{a} = \sum_{i=1}^{d} a_i^2$
- Sajátérték, sajátvektor
  Jobboldali sajátérték: $A\mathbf{x} = \lambda\mathbf{x} \Rightarrow x \neq 0 \Leftrightarrow \det(A - \lambda I) = 0$
  (Miért?)
  pl. A=$\begin{bmatrix} 3 & \sqrt{20} \\ \sqrt{20} & 4 \end{bmatrix}$ –> $\lambda^2 - 7\lambda - 8 = 0$
  Baloldali sajátérték: $\mathbf{y}A = \lambda\mathbf{y}$