Adatbányászati technikák - Bevezetés 2017 tavasz

Csima Judit

BME, VIK
Számítástudományi és Információelméleti Tanszék

2018. február 5.

Miért érdekes az adatbányászat?

- Egyre több és több adat keletkezik:
  - kereskedelem: vásárlói kosarak, bankkártya tranzakciók adatai, online vásárlások adatai
  - webes log file-ok
  - genome szekvenálás
  - (űr)távcsövek, fényképek adatai
  - műholdak időjárási adatai
  - orvosi adatok
  - social media infók
  - online eszközökkel gyűjtött adatok

Miért kellenek, miért jók az algoritmusok?

- óriási mennyiségű adat: pl. genome szekvenálás, GenBank (NCBI) növekedése
  - 1982: $6,8 \cdot 10^5$ bázispár
  - 2007: $8 \cdot 10^8$ bázispár (másfél évente duplázódott a méret)
  - 2016. december: $2,2 \cdot 10^{11}$ (kilenc év alatt 300-szorosa lett és ez egyre durvább lesz)
- teljesen reménytelen ebben kézzel megtalálni bármit
- hagyományos adatfeldolgozási technikák nem jók
- de gyors, nagy memóriájú számítógépek vannak

Milyen haszon származhat az adatbányászatból?

- gyors, automatikus osztályozás: emberi, szakértői munka kiváltása
- rejtett összefüggések kinyerése (emiatt pénz, tudományos eredmény, stb.)
- osztályozás, profilalkotás alapján egyedi elérés (vásárlókhoz, választókhoz)

Néhány alkalmazás

- osztályozás automatikusan ember helyett, előtt (csillagászat, diagnosztika)
- hírfigyelő portálok
- ajánló rendszerek: online vásárlások, webes böngészés alapján

Mi is ez?

Mi az adatbányászat?

- nem-triviális, rejtett, de hasznos(nak tűnő) információ kinyerése
- sokszor más célra, vagy direkt cél nélkül gyűjtött adatokból
- nagy adathalmazból
- automatikusan

Óvatosan!
Sok adatból sok olyan dolog is kinyerhető, ami csak a véletlennek köszönhetően van ott. (Data mining eredetileg pejoratív kifejezés volt).

Óvatosan!

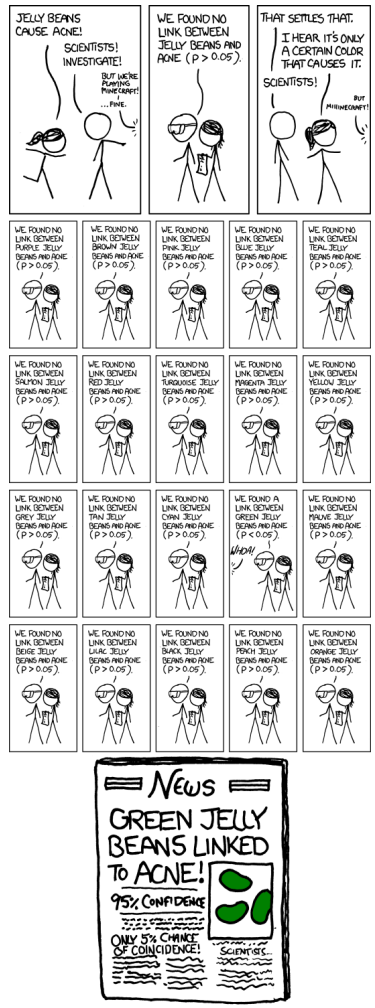
▸ xkcd

xkcd

A webcomic of romance, sarcasm, math, and language.

Significant

|< < Prev Random Next > >|



|< < Prev Random Next > >|

Permanent link to this comic: http://xkcd.com/882/
Image URL (for hotlinking/embedding): http://imgs.xkcd.com/comics/significant.png

RSS Feed - Atom Feed

Comics I enjoy:
Three Word Phrase, SMBC, Dinosaur Comics, Oglaf (nsfw), A Softer World, Buttersafe, Perry Bible Fellowship,

Óvatosan!

- adatok feldolgozása nem mindig adatbányászat
- leíró adatelemzés (pl. népszámlálási adatok) összegzése még nem adatbányászat
- az összefüggések megtalálása adatbányászat, de: összefüggés (korreláció) nem jelent ok-okozati összefüggést
- ok-okozati összefüggések feltárása általában már nem adatbányászat: ahhoz randomizált vizsgálatok kellenek kontrollcsoportok felhasználásával, ellenőrzött körülmények között

Óvatosan!

- testmagasság és hajhossz: minél magasabb valaki, annál rövidebb a haja
- lelkészek és alkoholisták száma egy városban
- fagylalteladás és vízbefulladás: minél több fagylaltot adnak el egy időszakban, annál többen fulladnak vízbe
- minél több tűzoltó megy ki egy tűzesethez annál nagyobb lesz a kár

Óvatosan!

▸ xkcd

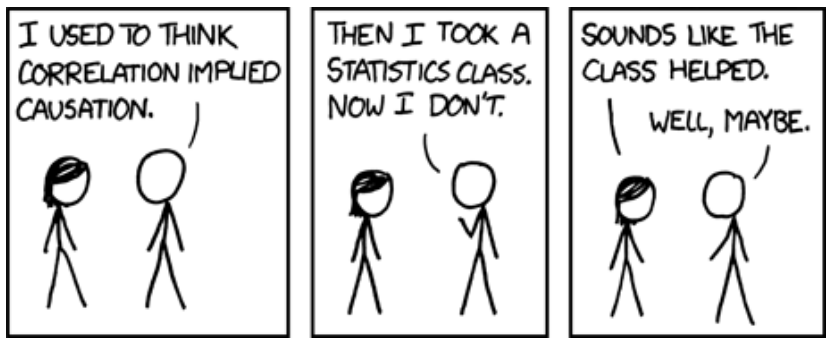
Archive
What If?
Blag
Store
About

xkcd

A webcomic of romance, sarcasm, math, and language.

Correlation

|< < Prev Random Next > >|



|< < Prev Random Next > >|

Permanent link to this comic: http://xkcd.com/552/
Image URL (for hotlinking/embedding): http://imgs.xkcd.com/comics/correlation.png

Eredete, hasonló tudományterületek

- machine learning (gépi tanulás), statisztika, adatbáziskezelés
- ezekből mind átvett elemeket, de ezek egyike sem jó önmagában a nagy adathalmazok kezelésére (big data): rengeteg adat (sok sor), nagy dimenziószám (sok oszlop), heterogén, gyakran elosztott adathalmazok, nem tiszta adatok

Alapfeladatok

- Előrejelzés, predikció (prediction): néhány változó értékei alapján egy másik (cél)változó értékének előrejelzése
  - Regresszió (regression): folytonos értékű célváltozó, supervised learning
  - Osztályozás (classification): diszkrét értékű célváltozó, supervised learning
  - Anomália detektálás, supervised learning
- Leíró módszerek: ember számára értelmezhető mintázatok keresése az adathalmazban
  - Változók közötti kapcsolat leírása, adott modell paramétereinek beállítása, modellválasztás
  - Klaszterezés (clustering): önmagában is érdekes, illetve az exploratory elemzés részeként is
  - Asszociációs szabályok keresése

Regresszió

- folytonos változó értékét meghatározni a többi változó értékéből
- pl.: lakások adatai: szobaszám, négyzetméter, van-e ablak a fürdőn, stb. alapján lakás ára
- blogbejegyzés jellemzői (kulcsszavak, hossz, stb.) alapján kommentszám előrejelzés
- új termék reklámjára fordított összeg, termék bizonyos jellemzői alapján várható eladási adatok
- film adatai (romantikus, kardozós, gyilkolászós, gyártási költség, szereplők) alapján várható bevétel vagy várható tetszési index

Regresszió

- lehet lineáris és nem-lineáris modellt alkotni
- kérdés, hogy mik a fontos változók, kellenek-e származtatott változók, lineáris modell kell-e, adott modellben mik a paraméterek (ez egy egész nagy terület önmagában)
- machine learning-ben jól kidolgozott elmélet
- mi nem (nagyon) tanulunk erről

Osztályozás: definíció

- angolul classification
- supervised learning
- rekordok adottak, több attribútummal (változóval)
- egyik változó a célváltozó (target vagy class): bináris vagy legalábbis diszkrét értékű
- ennek értékét kell megjósolni a többi változó értékéből: modellt kell választani és belőni a paramétereit
- lehetséges modellek: döntési fák, Bayes-osztályozók, neurális hálózatok
- cél: olyan modellt felépíteni, ami jól jelzi előre a célváltozó értékét

Osztályozás: módszertana

- adathalmaz két (esetleg három) részre osztása:
  - training set: ezekből az adatokból tanítjuk be a modellt (paraméterek beállítása úgy, hogy a training set-en minél kisebb legyen a hiba)
  - test set: itt nézzük meg, hogy egy (a modell felépítésekor nem látott) adathalmazon mennyire jól jelez előre
  - néha validation set: ha több modellt is kipróbálunk, akkor mindegyikhez belőjük a legjobb paramétert, a belőtt modelleket versenyeztetjük a validation set-en és a legjobbat kiértékeljük a test set-en
- modell értékeléséhez: kell valami számszerű, mérhető "jóság", ez sok minden lehet (erről később, egész nagy elmélet ez is)

Példa osztályozásra: direkt marketing online áruházban

- adott vásárlót érdemes-e megkeresni egy új termék direkt reklámjával (várhatóan megveszi-e?)
- célváltozó itt bináris: igen/nem
- az alapján, hogy: miket vett eddig, családi állapot, életkor, stb.

Példa osztályozásra még: csillagászat

- rengeteg kép készül, szakértő ember ehhez képest minimális
- osztályozzuk, hogy ami a képen van az csillag, galaxis vagy valami más esetleg különösen érdekes
- a különösen érdekeset majd megnézi az a kevés szakértő
- módszer: képszegmentálás, képekhez attribútumok rendelése, ezek alapján modellépítés

Példa osztályozásra még: karakterfelismerés

- kézzel írott számjegy micsoda?
- ez nem bináris, hanem 10 lehetséges érték van
- tipikus megközelítés: 10 külön osztályozó a 10 külön értékre, mindegyikre egy valószínűség jön ki: p_i valószínűséggel az i-es számjeggyel van dolgunk, válasszuk azt, amire p_i a legnagyobb

Példa osztályozásra még: hezitáló ügyfelek felismerése

- adott ügyfél készül-e lelépni, szolgáltatót váltani?
- célváltozó itt is bináris: igen/nem
- az alapján, hogy milyenek az általa igénybe vett szolgáltatások jellemzői: hívásidők, hívások darabszáma, mikor telefonál, stb.
- training set: a szerződést felmondó ügyfelek felmondás előtti viselkedését leíró adatok

Példa osztályozásra még: spamszűrő

- adott email spam-e?
- célváltozó itt is bináris: igen/nem
- az alapján, hogy milyen szavak fordulnak elő a levélben (esetleg hogy honnan jön), van-e valami jellegzetessége a levélnek
- training set: egy csomó levél, amikről tudjuk ezeket a jellemzőket és azt, hogy spamek-e a felhasználók szerint

Pszichológiai profilalkotás

- emberek csoportosítása profilokba
- demográfiai adatok, online megnyilvánulások alapján
- a profil alapján célzott üzenetek

Anomália detektálás

- speciális osztályozás:
  - bináris célváltozó: fura/nem fura
  - a fura az ritka, emiatt más módszerek kellenek, mint egy sima osztályozásnál
- példa: bankkártya tranzakciók adatai (összeg, vásárlás helye, korábbi vásárlások adatai stb.) alapján eldönteni, hogy ez most egy lopott kártyás tranzakció-e
- példa még: data centerben számítógép adatai működés közben (memóriahasználat, processzoridő telítettsége, disc elérések száma, stb.) alapján eldönteni, hogy el van-e romolva
- repülőgép motorok tesztelése: vibráció, melegedés, stb. alapján eldönteni, hogy oké-e vagy esetleg alaposabban meg kell nézni

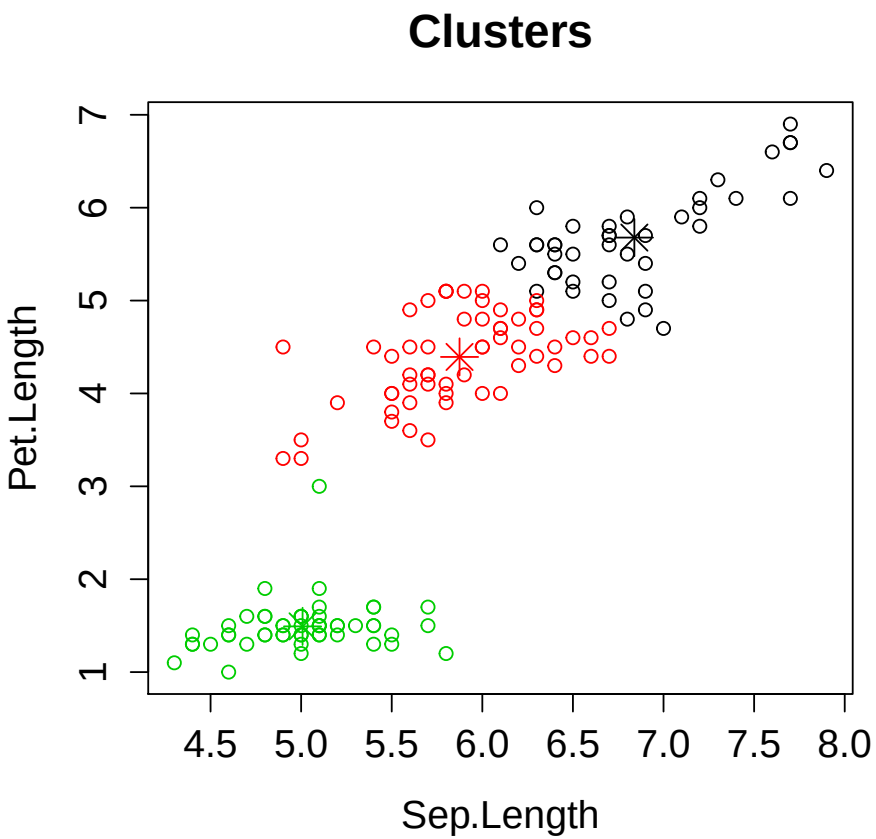
Klaszterezés: definíció

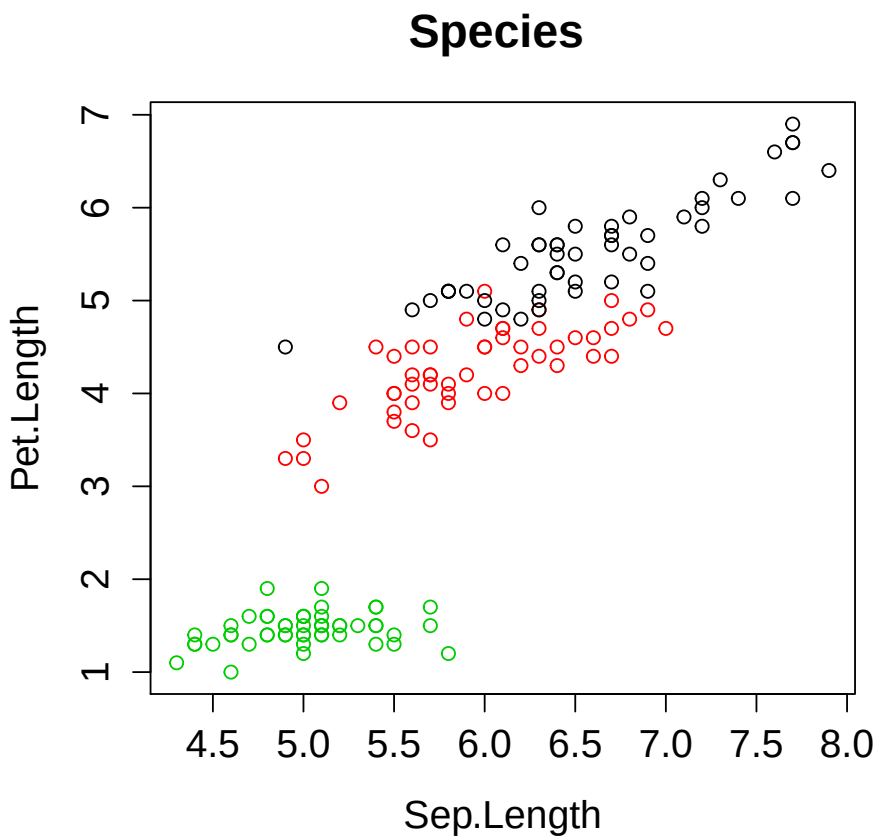
- unsupervised learning: nincs címkézés a pontokon
- hasonlósági mérték alapján (erről később) csoportokba sorolja a hasonló pontokat
- cél, hogy az egy csoportba kerülők hasonlóbbak legyenek egymáshoz, mint azok, akik külön kerülnek
- klaszterek számát gyakran előre kell tudni
- gyakran az adathalmazzal való ismerkedés része
- külön kérdés, hogy milyen változók alapján klaszterezünk és hogyan jelenítjük meg a klasztereket

Iris data set R-ben

```
> head(iris)
  Sep.Length Sep.Width Pet.Length Pet.Width Species
1        5.1       3.5        1.4       0.2  setosa
2        4.9       3.0        1.4       0.2  setosa
3        4.7       3.2        1.3       0.2  setosa
4        4.6       3.1        1.5       0.2  setosa
5        5.0       3.6        1.4       0.2  setosa
6        5.4       3.9        1.7       0.4  setosa
```

Iris klaszterezés





Klaszterezés példa: market segmentation

- cél: a vásárlók felosztása csoportokra vásárlási szokások szerint (barkácsáruháznál: barkácsolós, kertészkedős, stb.; könyváruháznál: gyerekkönyvek, anime, romantikus regények, stb.)
- attribútumok: korábbi vásárlások adatai, vásárló adati (életkor, nem, lakóhely, stb.)
- távolság, klaszterzés jóságának mérése: mennyire hasonlóak az egy klaszterbe eső emberek vásárlásai

Klaszterezés példa még: dokumentum osztályozás

- Dokumentumok osztályozása úgy, hogy a hasonló témájúak egy csoportba jussanak
- attribútumok, amik alapján csoportosítok: kulcsszavak
- előfeldolgozás munkás: mi a kulcsszó, minek a gyakorisága számít
- speciális alakú (ritka) adatmátrix jön létre, speciális technikák kellenek
- például a Google news szolgáltatás

Asszociációs szabályok: vásárlói kosarak adatai

- egy rekord azt tartalmazza, hogy mit vett egyszerre egy ember (vásárlói kosár)
- pl. egy vásárlás az, hogy {kenyér, tej, kóla, pelenka, sör}
- cél: olyan szabályok felállítás, amik azt írják le, hogy ha valaki vesz A terméket (vagy $A_1, A_2, \ldots A_k$ termékeket együtt) , akkor (valószínűleg) vesz B-t is
- valószínűleg: beállítható paraméter, hogy mikor mondjuk ezt
- óvatosan: van sok triviális szabály, olyanok kellenek, amik meglepőek, érdekesek

Asszociációs szabályok mire jók?

- Ha van egy $\{A, ...\} \rightarrow B$ szabályunk, akkor esélyes, hogy A árát csökkentve nő B eladása
- boltban áruk elhelyése a polcokon: amiket gyakran vesznek együtt az legyen hasonló helyen
- ezzel kezdődött az adatbányászat, sokat vizsgált terület, de aztán kissé elfelejtődött
- mostanában újra, mert gyakori elemhalmazok keresése érdekes sok helyen

Mik a kihívások?

- scalability: skálázhatóság nagy halmazokra, azaz sok sorral is el kell tudni bánni
- sok dimenzió: rengeteg attribútum lehet
- heterogén adatok, sok adattípus, speciális adatok
- adatminőség (hibák, nem konzisztens adatbázisok)
- elosztott számítás jó lenne
- data streams: folyamatosan jövő adatok, mindenre csak egyszer nézhetünk rá