Strojno učenje, umetna inteligenca, podatkovno rudarjenje ...

1. PREDAVANJE

izr. prof. Branko Kavšek

Vsebina predavanja

- Uvod: poplava podatkov
- Primeri aplikacij podatkovnega rudarjenja
- Podatkovno rudarjenje &
  odkrivanje znanja
- Tehnike podatkovnega rudarjenja

Trendi, ki vodijo v podatkovno poplavo

- Generiranje velikega števila podatkov:
  - Poslovne ustanove:
    banke, telekomunikacijske združbe,
    ostale poslovne združbe …
  - Znanstveni podatki:
    astronomija, biologija,
    kemija …
  - Splet:
    socialna omrežja,
    e-poslovanje …

Primeri ogromnega števila podatkov (pred 15 leti)

- Evropska VLBI (Very Long Baseline Interferometry) mreža 16-ih teleskopov generira **1 gigabit** podatkov **na sekundo** za vsakega od teleskopov v 25-ih dnevih opazovanj
  - že samo hranjenje teh podatkov predstavlja problem;

- AT&T posreduje **milijarde** klicev **na dan**
  - takšno količino podatkov ne morejo shraniti, zato je potrebna sprotna (“on the fly”) obdelava/analiza podatkovnega toka;

Največje zbirke podatkov leta 2003

- Komercialne baze podatkov (Winter Corp.-ova anketa 2003):
  - France Telecom ima največjo PB = ~30TB;
  - AT&T ima bazo = ~26 TB;

- Splet:
  - Alexa internetni arhiv: 7 let delovanja, 500 TB;
  - Poizvedbe v Google-u: 4+ milijarde strani, na stotine TB;
  - IBM WebFountain: 160 TB;
  - Internet Archive (www.archive.org): ~300 TB;

Od terabajtov do eksabajtov …

- UC Berkeley - ocena: **5 eksabajtov**
  (5 milijonov terabajtov) ustvarjenih podatkov v letu 2002.

  www.sims.berkeley.edu/research/projects/how-much-info-2003/

- **ZDA** ustvarijo **~40% vseh podatkov**,
  ki se hranijo v različnih oblikah

- Ocena za 2006: **161 eksabajtov** (IDC študija)

  www.usatoday.com/tech/news/2007-03-05-data_N.htm

- Projekcija za 2010: **988 eksabajtov**

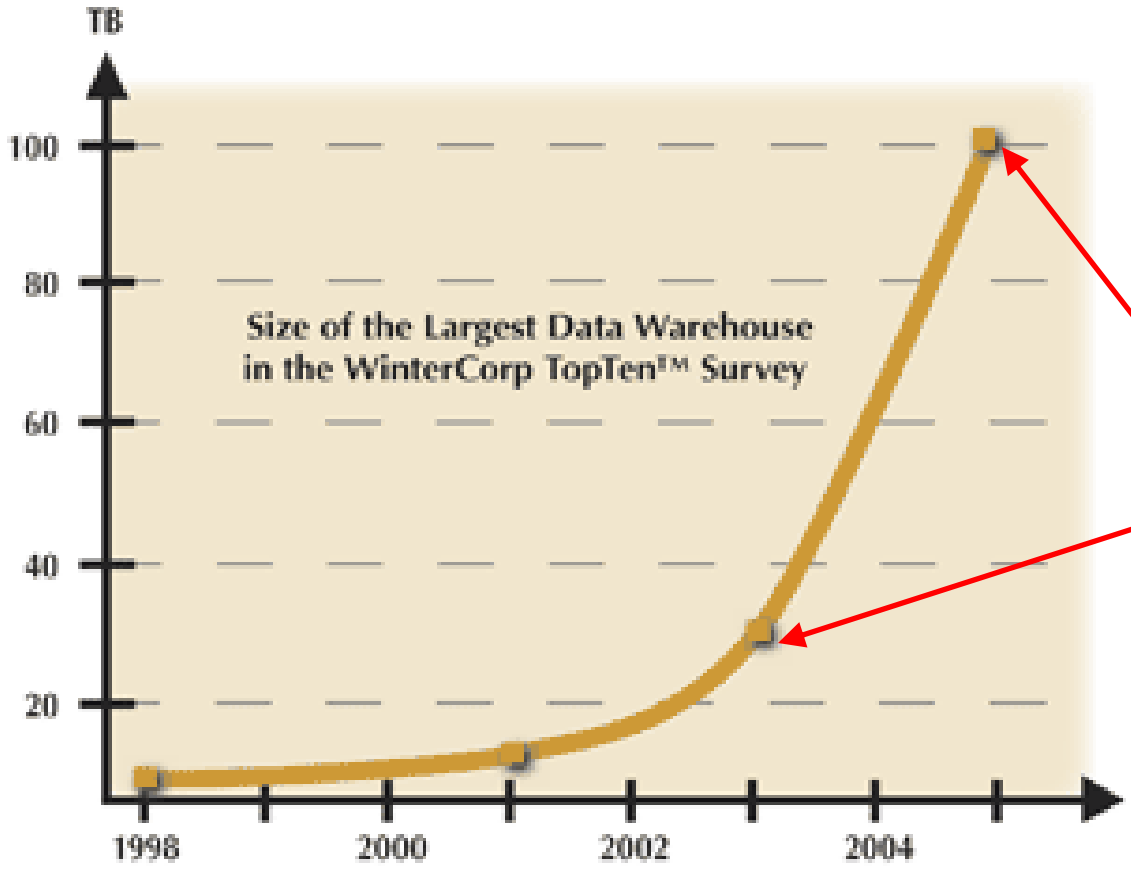
Največje zbirke podatkov leta 2005

Spet Winter Corp.-ova anketa komercialnih PB:

1. Max Planck Inst. for Meteorology: **222 TB**
2. Yahoo: **~100 TB** (največje podatkovno skladišče)
3. AT&T: **~94 TB**

http://dssresources.com/news/1010.php

Rast količine podatkov



V 2 letih se je velikost največje podatkovne baze **POTROJILA** !!!

Kaj pa danes?

Primer:
konec junija 2017 – CERN-ov podatkovni center shranjuje več kot **200 petabajtov** (200 milijonov gigabajtov)

Google-ovi podatkovni centri danes obdelajo **> 200 petabajtov** uporabniških podatkov vsak dan

Faktor rasti količine podatkov

- V letu 2002 je bilo ustvarjenih **2-krat več** podatkov kot leta 1999;
- V letu 2005 je bilo ustvarjenih **3-krat več** podatkov kot leta 2003;
- Zelo malo teh podatkov bo sploh kdaj pogledal človek

Podatkovno rudarjenje je zato **nujno potrebno**, če hočemo te podatke smiselno uporabiti !!!

Vsebina predavanja

- Uvod: poplava podatkov
- Primeri aplikacij podatkovnega rudarjenja
- Podatkovno rudarjenje & odkrivanje znanja
- Tehnike podatkovnega rudarjenja

Področja aplikacije strojnega učenja/podatkovnega rudarjenja

Znanost:

- astronomija, bioinformatika, odkrivanje novih zdravil …

Poslovni svet:

- CRM (Customer Relationship management), odkrivanje prevar (fraud detection), e-poslovanje, proizvodnja, šport/zabava, telekomunikacije, zdravstvo …

Splet:

- iskalniki, oglaševanje, web in text mining, socialna omrežja …

Državna uprava:

- nadzor, odkrivanje kaznivih dejanj, profiliranje davčnih utajevalcev …

Področja aplikacije

Kaj so najpomembnejše in najbolj razširjene poslovne aplikacije podatkovnega rudarjenja?

Podatkovno rudarjenje za profiliranje strank

- Napovedovanje “prebegov”,
- Usmerjeni marketing,
- Odplačevanje kreditov,
- Odkrivanje prevar,
- Bančništvo,
- Telekomunikacije,
- Prodaja na drobno …

“Prebegi”: študija primera

- Dejstvo:
  odstotek prebega od enega mobilnega operaterja k drugemu je cca. 25-30% letno (ZDA)!

- Kaj je tu naloga podatkovnega rudarjenja?
  - Predpostavka: imamo podatke o naročnikih za preteklih N mesecev.

“Prebegi”: študija primera (2)

Naloga:

- Napovedati kdo bodo potencialni “prebežniki” naslednji mesec.
- Oceniti vrednost le-teh in kakšna naj bo smiselna ponudba za njih, da bodo ostali.

“Prebegi”: rezultat

- Verizon Wireless – baza uporabnikov mobilnih storitev,
- Identifikacija potencialnih prebežnikov,
- Izgradnja (več) regionalnih modelov,
- Usmerjena reklama uporabnikom, pri katerih je največja verjetnost, da bodo sprejeli novo ponudbo,
- Zmanjšanje prebegov iz več kot 2%/mesec na manj kot 1.5%/mesec (ogromen efekt, > 30 M naročnikov).

(Poročilo iz leta 2003)

Ocena kreditne sposobnosti: študija primera

Situacija: Oseba zaprosi za posojilo.

Naloga: Naj banka posojilo odobri?

Opomba:
Ljudje z zelo visoko kreditno sposobnostjo po navadi ne potrebujejo posojila, ljudje z zelo nizko kreditno sposobnostjo posojila ne bodo mogli vrniti
➔ Za banko so najboljši/zanimivi tisti, ki so “nekje vmes”

Ocena kreditne sposobnosti: rezultati

- Banke oblikujejo kreditne modele s pomočjo najrazličnejših tehnik strojnega učenja,
- Hipoteke in razširjenost kreditnih kartic so rezultat uspešnega napovedovanja kreditne sposobnosti,
- Zelo razširjeno v veliko državah.

e-poslovanje

Oseba kupi knjigo (izdelek) na Amazon.com

Kaj je naloga?

Uspešno e-poslovanje: študija primera

Naloga:
Priporočilo drugih knjig (izdelkov), ki jih bo ta oseba zelo verjetno želela imeti.

Amazon.com uporablja metode razvrščanja v skupine glede na kupljene knjige:

- Kupci, ki so kupili "**Advances in Knowledge Discovery and Data Mining**", so kupili tudi "**Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations**".

Program za priporočila je dokaj uspešen.

Neuspešno e-poslovanje: študija primera (KDD-Cup 2000)

Podatki:
Podatki o klikih in nakupih s spletne strani Gazelle.com, trgovine z obutvijo ++

Naloga:
Opisati obiskovalce spletne strani,
ki so v povprečju kupili za več kot $12.

Podatki = 3.465 nakupov, 1.831 obiskovalcev,

Zelo zanimive analize s strani tekmovalcev

- Na tisoče porabljenih ur - $X.000.000 (milijoni) za svetovanja,

Skupna prodaja: -$Y.000,

Zaključek: Gazelle.com gre v stečaj, Avg 2000.

Mikromreže genoma: študija primera

Če imamo podane mikromreže za določeno število pacientov, ali lahko:

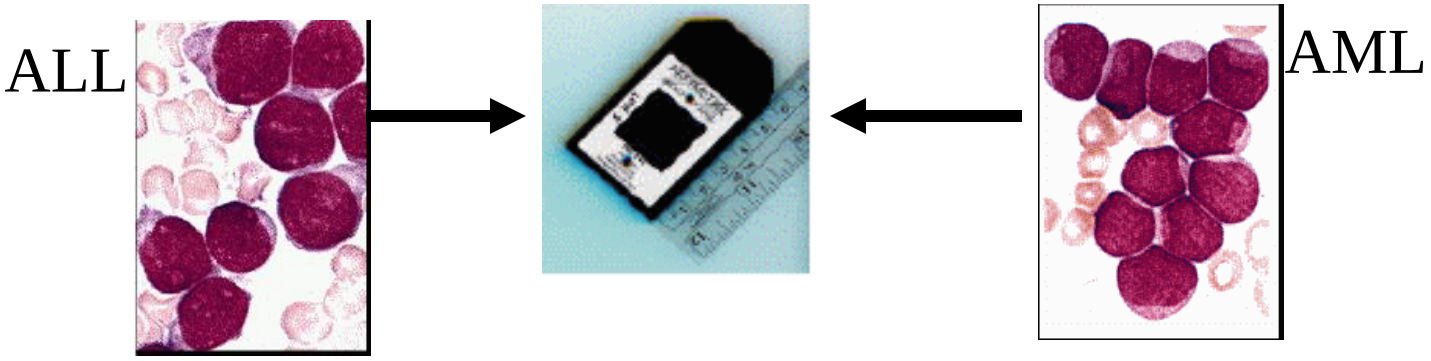
- z zanesljivostjo diagnosticiramo bolezen?
- napovemo rezultat zdravljenja?
- priporočimo najboljše možno zdravljenje?

Primer: ALL/AML podatki

38 učnih primerov, 34 testnih, ~7.000 genov

2 razreda: Acute Lymphoblastic Leukemia (ALL) vs
Acute Myeloid Leukemia (AML)

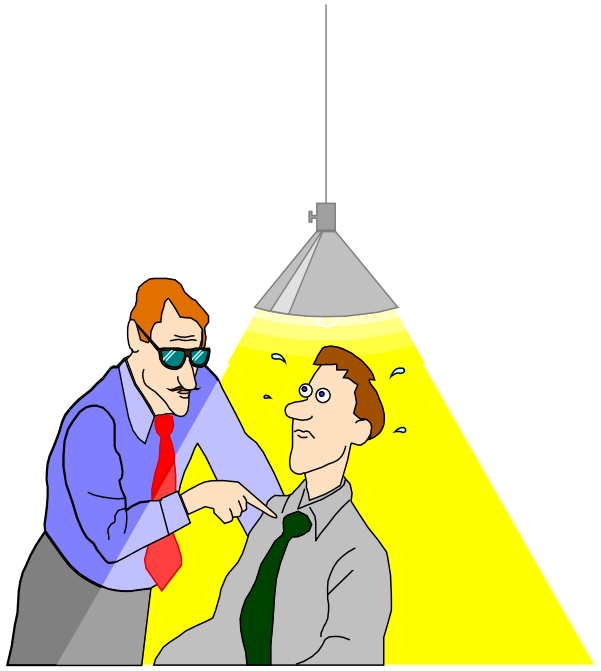
Uporabimo učne primere za izgradnjo diagnostičnega modela



Rezultat na testnih primerih:
33/34 pravilnih napovedi, 1 napaka (možna napačna označitev)

Varnost in ugotavljanje prevar: študija primera

- Prevare s kreditnimi karticami
- Pranje denarja
  - FAIS (US Treasury)
- Telefonske prevare
  - AT&T, Bell Atlantic, British Telecom/MCI
- Bio-terorizem na olimpijadi v Salt Lake-u 2002



Podatkovno rudarjenje in zasebnost

- Leta 2006 NSA (National Security Agency) objavi, da so leta analizirali klice z namenom odkrivanja terorističnih mrež
- Analize podatkov iz socialnih omrežij
- Vdor v zasebnost – ti je vseeno ali se vsebina tvojih klicev shranjuje v vladnih bazah?
- Kaj če NSA program odkrije enega pravega terorista na 1.000 “napačnih”? 1.000.000 napačnih?

Problemi prikladni za podatkovno rudarjenje

- Zahtevajo odločitve na podlagi znanja,
- Imajo spreminjajoče okolje,
- Imajo ne-optimalne trenutne rešitve,
- Imajo dostopne, zadostne in relevantne podatke,
- Imajo veliko poplačilo za prave odločitve!

Če imamo opravka z osebnimi podatki,
je zelo pomemben vidik zasebnosti !!!

Vsebina predavanja

- Uvod: poplava podatkov
- Primeri aplikacij podatkovnega rudarjenja
- Podatkovno rudarjenje & odkrivanje znanja
- Tehnike podatkovnega rudarjenja

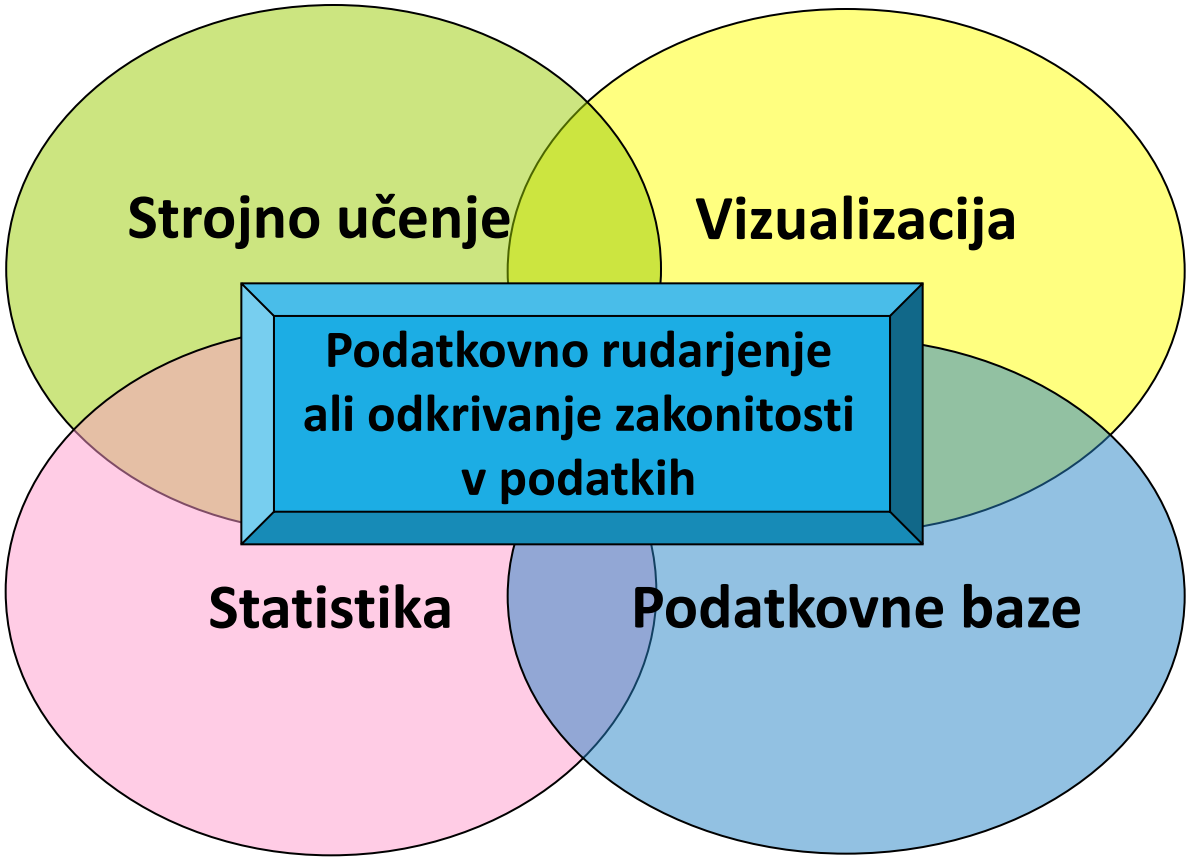
Definicija “odkrivanja zakonitosti”

Odkrivanje zakonitosti v podatkih je:

- *Netrivialni proces* identifikacije
  - *veljavnih*
  - *novih*
  - *potencialno uporabnih*
  - in nazadnje *razumljivih vzorcev* v podatkih.

Povzeto iz:
Advances in Knowledge Discovery and Data Mining, Fayyad, Piatetsky-Shapiro, Smyth, and Uthurusamy, (Chapter 1), AAAI/MIT Press 1996

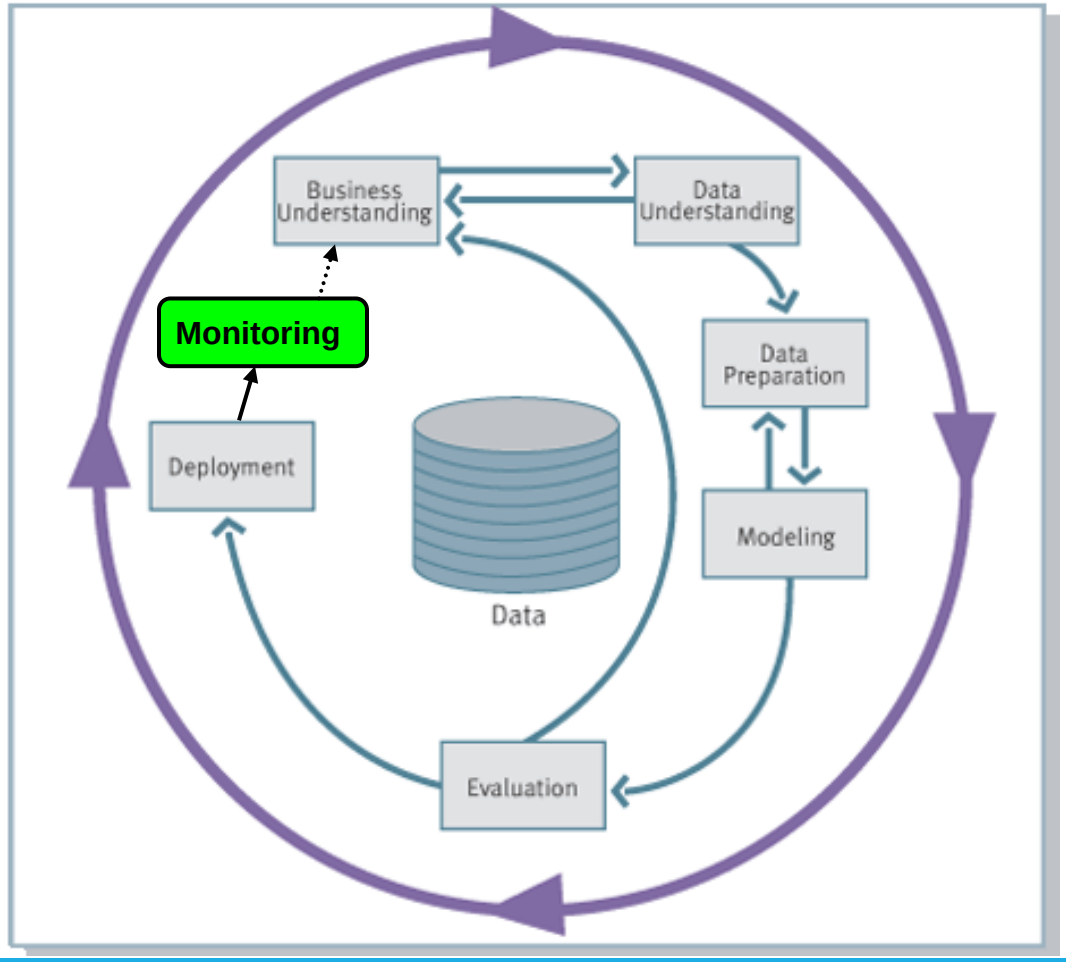
Povezana področja



Statistika, strojno učenje in podatkovno rudarjenje

- Statistika:
  - Bolj usmerjena v teorijo
  - Večji fokus na testiranju hipotez
- Strojno učenje:
  - Uporaba hevristik
  - Usmerjeno v izboljševanje performans učnih agentov
  - Se ukvarja tudi z učenjem v realnem času in robotiko (kar se podatkovno rudarjenje ne toliko)
- Podatkovno rudarjenje ali odkrivanje zakonitosti v podatkih:
  - Združuje teorijo in hevristike
  - Usmerjeno v celoten proces odkrivanja zakonitosti, kar vključuje "čiščenje" podatkov, učenje, integracijo in vizualizacijo rezultatov
- **Razmejitve so "mehke"**.

Proces odkrivanja zakonitosti v podatkih – CRISP-DM metodologija



Glej tudi:
https://en.wikipedia.org/wiki/Cross_Industry_Standard_Process_for_Data_Mining
za več informacij

Iz zgodovine:
Več imen za podatkovno rudarjenje

Data Fishing, Data Dredging: 1960 –

- Uporabljano s strani statistikov (velja za slabo ime);

Data Mining: 1990 –

- Uporabljano s strani strokovnjakov za BP;
- V letu 2003 dobi slab prizvok zaradi TIA;

Knowledge Discovery in Databases: 1989 –

- Uporabljano s strani AI in ljudi s področja strojnega učenja;

tudi:
Data Archaeology, Information Harvesting, Information Discovery, Knowledge Extraction ...

Trenutno:
Data Mining (podatkovno rudarjenje) in
Knowledge Discovery (odkrivanje zakonitosti) se uporablja kot sinonima.

Vsebina predavanja

- Uvod: poplava podatkov
- Primeri aplikacij podatkovnega rudarjenja
- Podatkovno rudarjenje & odkrivanje znanja
- Tehnike podatkovnega rudarjenja

Najbolj pogoste tehnike strojnega učenja

Klasifikacija: napovedovanje razreda primerov

Razvrščanje v skupine: iskanje skupin v podatkih

Asociacije: npr. A & B & C se pogosto pojavljajo

Vizualizacija: ljudem omogoča lažje odkrivanje zakonitosti

Povzemanje: opisovanje skupin podatkov

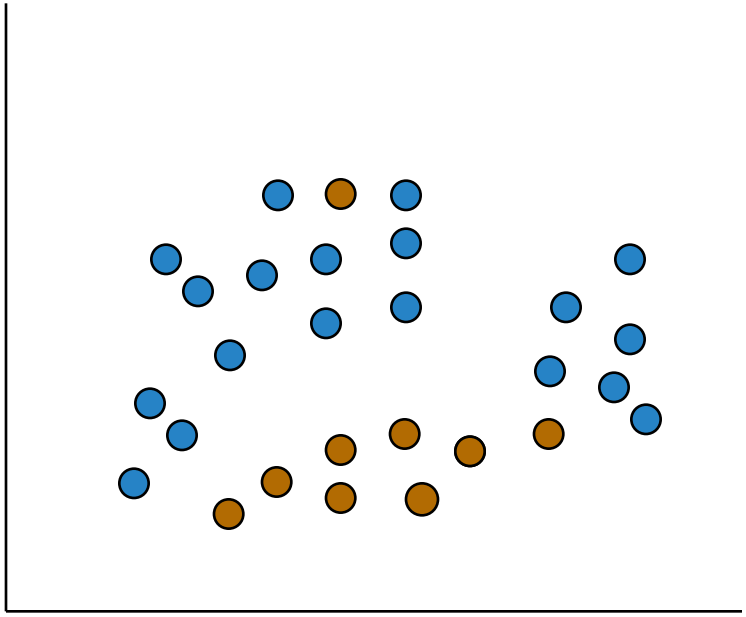
Odkrivanje odstopanj: iskanje znakov sprememb

Regresija/ocenjevanje: napovedovanje zveznih vrednosti

Analiza povezav: odkrivanje relacij med podatki

Tehnike podatkovnega rudarjenja: napovedovanje (klasifikacija)

Iz označenih primerov se nauči klasifikacijskega modela, ki napove razred primera

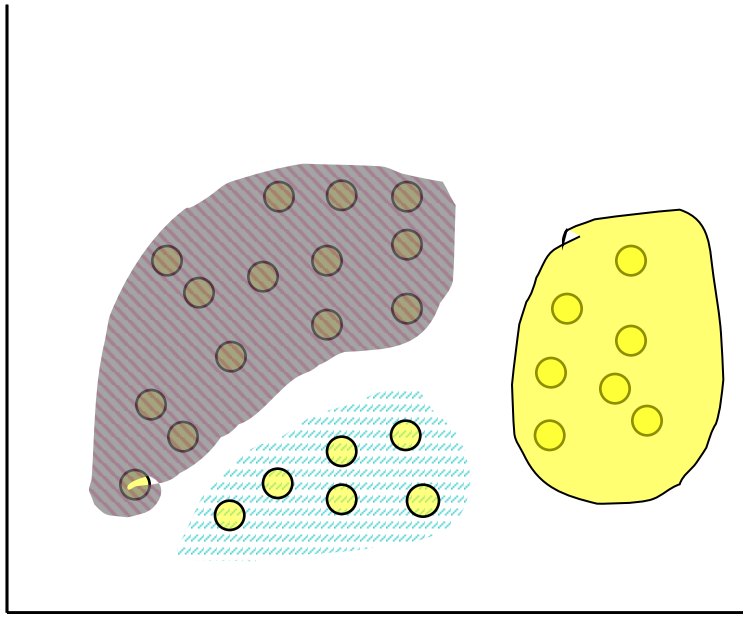


Več pristopov:

statistične metode,
odločitvena drevesa,
nevronske mreže ...

Tehnike podatkovnega rudarjenja: razvrščanje v skupine (clustering)

Poišče “naravne” skupine podatkov, če ima podane neoznačene primere



Povzetek

- Tehnološki trendi vodijo v podatkovno poplavo
  - Podatkovno rudarjenje je potrebno, da lahko podatkom damo smisel;
- Podatkovno rudarjenje ima veliko aplikacij, ki so lahko uspešne ali pa ne;
- Proces odkrivanja znanja v podatkih
- Tehnike podatkovnega rudarjenja
  - Napovedovanje (klasifikacija), razvrščanje v skupine (clustering) …