Podatkovno rudarjenje

predavatelj: Tomaž Curk
(tomaz.curk@fri.uni-lj.si)

asistent: Martin Stražar
(martin.strazar@fri.uni-lj.si)

2015-2016

Govorilne ure

Takoj po predavanjih, sreda ob 12:05 ali ponedeljek ob 13:00, kabinet R3.15

Z vajami začnemo naslednji teden, 29.2.2016.

Pravila igre

- 5 domačih nalog (4 redne + 1 dodatna)
  - vsaka 100 točk, zbrati je potrebno 201 točk (> 50% rednih točk)
  - največ 7 dni zamude, vsak dan izgubite 10% točk,
    po sedmih dneh 0 točk
  - priznane točke = (ocenjene točke)×0.9^{dni_zamude}
    n.pr.: dobili 90 točk, a 2 dneva in 1 min zamude: $90 \times 0.9^3 = 65.61$
- 1 projekt ("pisni izpit")
  - **do 14.3.2016**: izberite problem/podatke in oddajte predlog
  - **do 16.4.2016**: vmesno poročilo in predstavitev pred vrstniki
  - **do 30.5.2016**: analizirajte (prikažite, modelirajte) in napišite končno poročilo o zanimivih, nepričakovanih ugotovitvah
- izpit
  - pogoj: opravljene domače naloge
  - pisni: poročilo o projektu (vmesno in zaključno)
  - ustni: zagovor rezultatov projekta in poznavanje uporabljenih metod

Prisotnost - test

http://goo.gl/forms/h24XfnFBwQ

Motivacija in predznanje

http://goo.gl/forms/quPXr48Whn

Uvodni primer: analiza podatkov o izboru predmetov

Podobnost predmetov

- S_x – množica vseh študentov, ki so izbrali predmet x
- podobnost med predmetoma i in j:

$$s(i,j) = \frac{|S_i \cap S_j|}{|S_i \cup S_j|}$$

[delež študentov, ki so izbrali oba predmeta, med vsemi študenti, ki so izbrali vsaj enega od dveh predmetov]

Podobnosti predmetov – mera razdalje

- različnost med predmeti je obrnjena podobnost:

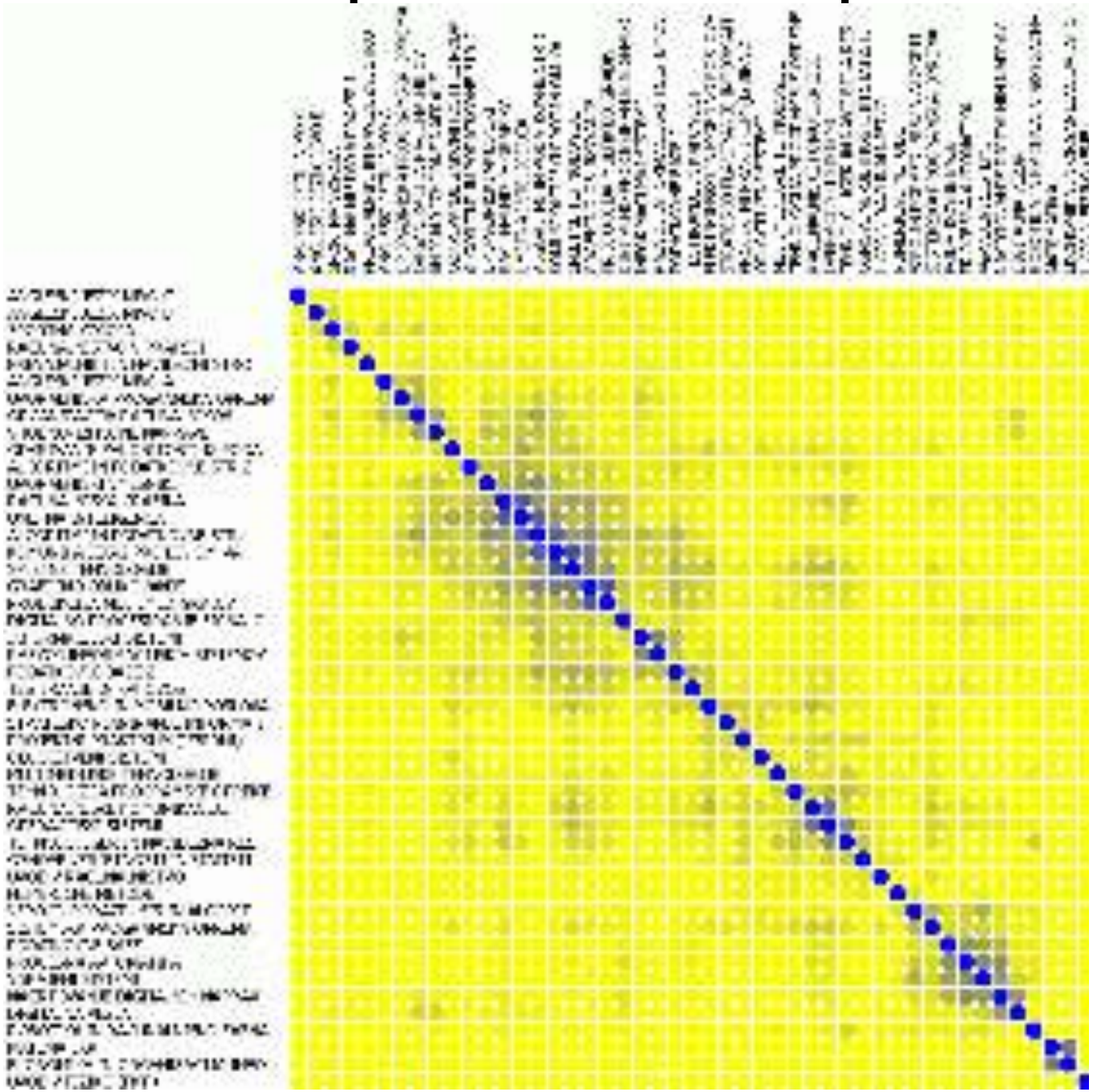
$$d(i,j) = 1 - \frac{|S_i \cap S_j|}{|S_i \cup S_j|}$$

- vzamemo vse podatke in izračunamo razdalje med vsemi možnimi pari

Podobnosti predmetov – izračunane razdalje

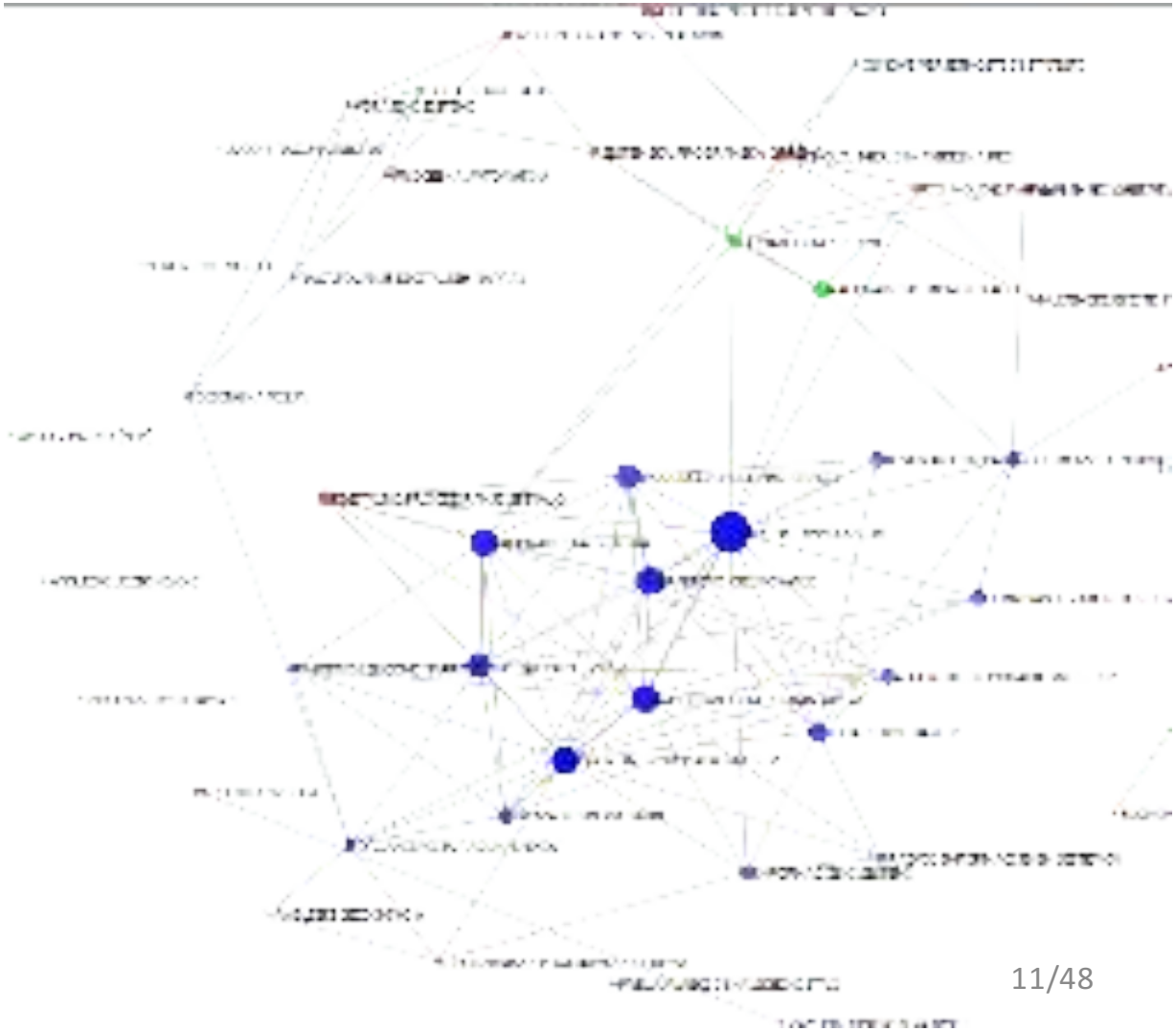
	A	B	C	D	E	F	G	H	I	J
1	47	labelled								
2	PODATKOVNE BAZE	0	1	1	1	0.981	0.645	1	0.793	0.556
3	ANGLESKI JEZIK NIVO B	1	0	0.917	0.952	1	0.968	0.933	1	
4	ANGLESKI JEZIK NIVO C	1	0.917	0	0.944	1	0.964	0.977	1	
5	ANGLESKI JEZIK NIVO A	1	0.952	0.944	0	1	0.973	0.941	1	
6	TEHNOLOGIJA PROGRAMSKE OPREME	0.981	1	1	1	0	0.983	0.915	0.92	
7	NACRTOVANJE DIGITALNIH NAPRAV	0.645	0.968	0.964	0.973	0.983	0	1	0.794	0.548
8	RAZVOJ INFORMACIJSKIH SISTEMOV	1	0.933	0.977	0.941	0.915	1	0	1	0.966
9	VZPO.IN PORAZD. SIS.IN ALGORIT	0.793	1	1	1	0.92	0.794	1	0	0.733
10	PROCESNA AVTOMATIKA	0.556	1	1	1	1	0.548	0.966	0.733	
11	TEHNOLO. IGER IN NAVIDEZNA RES	0.895	1	1	1	0.677	0.97	0.951	0.912	0.985
12	RACUNALNISKE KOMUNIKACIJE	0.844	1	1	1	0.743	0.857	0.897	0.894	0.815
13	UMETNA INTELIGENCA	0.96	0.943	0.966	0.912	0.92	0.91	0.913	0.959	0.971
14	TESTIRANJE IN KAKOVOST	0.873	0.983	0.982	0.953	0.889	0.915	0.866	0.923	0.862
15	OSNOVE VERJETNOSTI IN STATISTI	0.914	1	1	1	0.787	0.977	0.947	0.943	0.975
16	INFORMACIJSKI SISTEMI	0.919	0.944	0.981	0.855	0.911	0.957	0.541	1	0.923
17	PREVAJALNIKI IN NAVIDEZNI STRO	1	1	0.962	0.941	0.907	0.93	0.912	0.974	0.976
18	UPORABNIŠKA PROGRAMSKA OPREMA	1	0.938	0.967	0.75	1	0.98	0.842	1	

Podobnosti predmetov – prikaz razdalj



Podobnosti predmetov - zemljevid

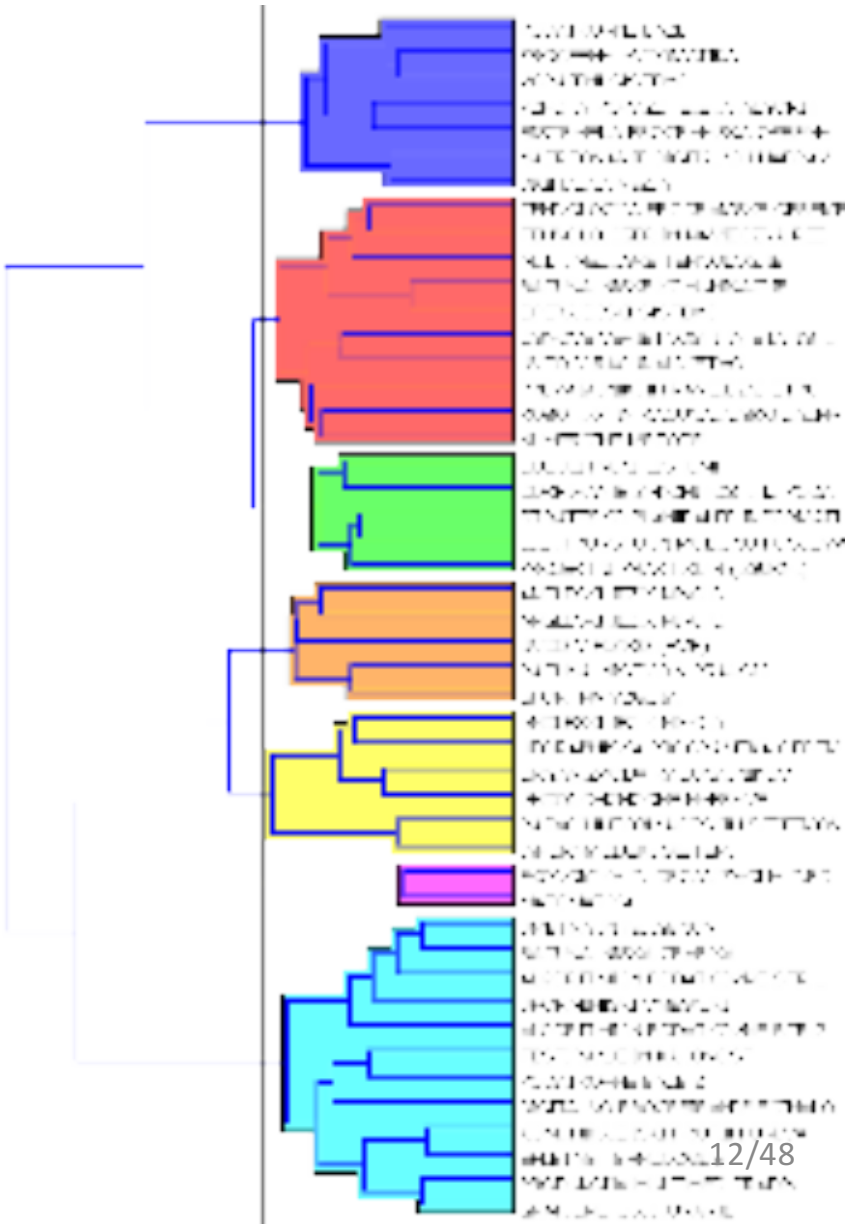
- Razdalje na sliki odgovarjajo, kolikor je mogoče, izračunanim razdaljam.
- Najbolj podobni so dodatno povezani.
- Velikost ustreza številu študentov (popularnost).
- Barve so letniki:
  - prvi zeleni,
  - drugi modri,
  - tretji rdeči



Podobnosti predmetov - skupine

Razdelitev v skupine lažje prikažemo s hierarhičnim razvrščanjem v skupine in dendrogramom.

- med hardverskimi predmeti so se priteple podatkovne baze,
- rdeča je zmešnjava,
- zelena bi sodila skupaj, vendar glede na dejansko snov niti ne,
- oranžna so očitno izbirni predmeti,
- rumena so hardverski informatiki,
- …



Kaj gre narobe?

Dve težavi:

- podatki: vsebujejo le izbor predmetov, ki jih študenti poslušajo eno leto.
  - študenti drugega le predmete drugega, predmete tretjega pa bodo naslednje leto.
  - predmeti tretjega so si podobni, ker pač izbirajo le študenti tretjega letnika.
  - stari podatki, iz časa prehoda na bolonjski študij, zato navidezno predmeti prvega letnika podobni tistim iz tretjega.
- mera podobnosti: npr, angleščine se navidezno razlikujejo med seboj, ker vsak študent izbere le eno. Podobno, če dva predmeta enaka a izbrati možno le enega, potem navideno različna.

Glavna težava: ne vemo, kaj je sploh namen te analize?!

Svetovalni sistem?

Recimo, da želimo zgraditi svetovalni sistem, ki daja nasvete v slogu:

“Študenti, ki so izbrali te predmete, so izbrali tudi...”

Kako svetovati:

- Svetujemo predmete, ki so po razdalji blizu že izbranim predmetom.
- Svetujemo predmete, ki so v istih skupinah kot že izbrani predmeti.

Svetovalni sistem?

Študent je izbral k predmetov: s_1, s_2, ..., s_k

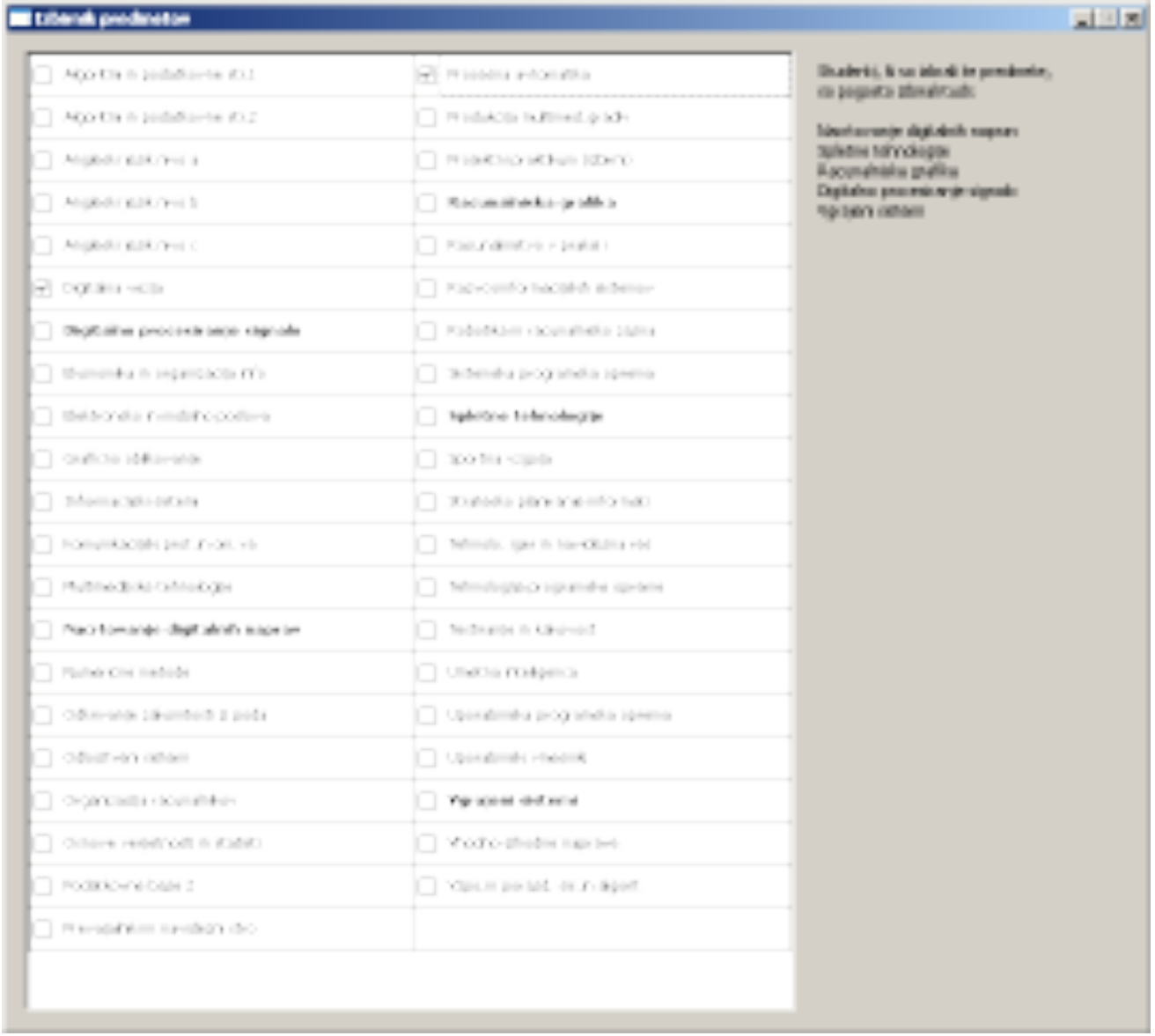
Za vse ostale (t) napovemo verjetnost, da ga zanimajo, če je izbral onih k:

$$p(s_t|s_1,s_2,\ldots,s_k) = \frac{p(s_1,s_2,\ldots,s_k|s_t)p(s_t)}{p(s_1,\ldots,s_k)} \approx p(s_t)\prod_{i=1}^{k} p(s_i|s_t)$$

[poenostavljen naivni Bayesov klasifikator]

Prikažemo le prvih pet najbolj verjetnih.

Svetovalni sistem - aplikacija



Podobnost med študenti?

Dva študenta podobna, če izbirata podobne predmete.

Uporabimo lahko iste pristope kot za predmete.

Glavni vzorec, zaradi nerodno sestavljenih podatkov:

- drugi letnik modri,
- tretji letnik rdeči.

Kaj je podatkovno rudarjenje?

http://goo.gl/forms/8yXby6aLC7

Podatkovno rudarjenje

- Data mining – odkrivanje zakonitosti podatkov, odkrivanje znanja iz podatkov, podatkovno rudarjenje
- **Iskanje uporabnih vzorcev v podatkih.**
  - odkrite vzorce bomo uporabili pri odločanju, razumevanju,...
- Podatkov je navadno veliko => uporaba računalnika.
- Poudarek na praktičnosti in ne toliko na matematični rigoroznosti:
  - domiselna uporaba podatkov
  - domiselna vizualizacija
  - domiselna izbira metod statistike, strojnega učenja, ...

Kje lahko uporabimo podatkovno rudarjenje?

Uporabnost

- **Poslovna inteligenca**: poslovne odločitve temeljijo na podatkih. Surove podatke je potrebno obdelati, vizualizirati, modelirati.
- **Sistemi za odnose s strankami**: razumevanje strank in izbira strank za določen produkt/storitev, napovedovanje prebegov strank.
- **Analiza nakupovalne košarice**: "svetovanje" kupcem kaj vse še potrebujejo, prilagajanje ponudbe.

Uporabnost II

- **Odkrivanje zlorab**: detekcija sumljivih bančnih transakcij (ukradene bančne kartice), goljufanje zavarovalnic (režirane-izmišljenje prometne nesreče, okvare in poškodbe), nadzor spletnega prometa, spam filtri, ...
- **Družbeno-ekonomske analize**: gore zbranih podatkov o soci-demografskih gibanjih, gospodarsko-finančni rezultati.
- **Analiza prostorskih podatkov**: zemljevidi omogočajo opazovanje v kontekstu.

Uporabnost III

- **Znanost & tehnika**: tehnologija omogoča znanstvenikom zbiranje gore podatkov.

  Genetika: odkrivanje funkcij genov in analiza reakcij celic/organizmov na dražljaje.

  Farmacija: odkrivanje novih zdravil.

  Astronomija: avtomatiziranje razlikovanja med različnimi objekti, ki jih najdejo s teleskopi.

  Medicina: odkrivanje bolezni, napovedovanje potekov in najprimernejših metod zdravljenja.

  ...

Vsebina predmeta

- Metode
  - Podatki in problemi
  - Vizualizacija
  - Modeliranje
  - Vrednotenje modelov
  - Uporaba rezultatov
- Konkretna področja uporabe metod

Proces odkrivanja znanja iz podatkov in podatkovnega rudarjenja

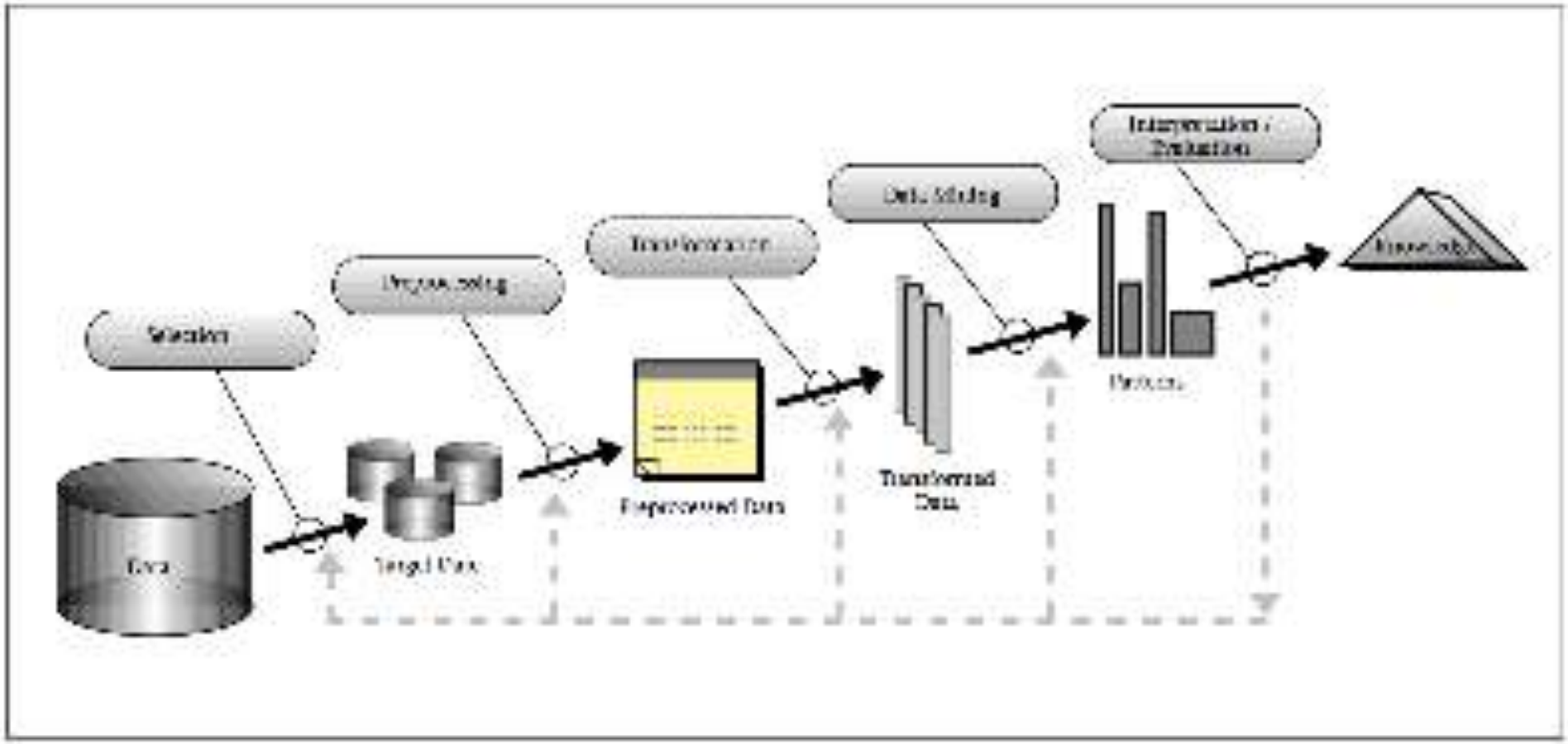


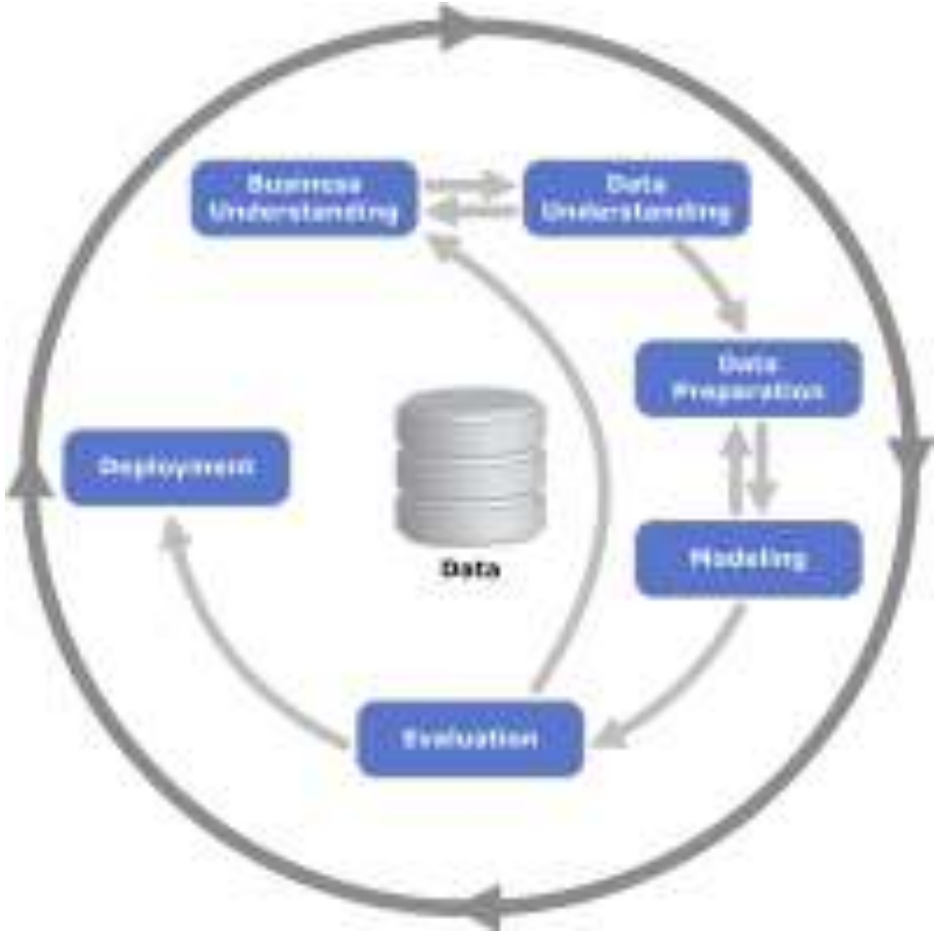
Figure 1. An Overview of the Steps That Compose the KDD Process

Fayyad et al (1996): From data mining to knowledge discovery in databases, AI magazine, 17(3):37

CRISP-DM

Cross Industry Standard Process for Data Mining

- Začetki v 1996.
- Nikoli zares zaključen proces, temveč iteracije.
- IBM.



[http://en.wikipedia.org/wiki/File:CRISP-DM_Process_Diagram.png]

SEMMA

Sample: vzorčenje podatkov

Explore: razumevanje podatkov z vizualizacijo, pričakovani, nepričakovani podatki

Modify: izbiranje in transformiranje atributov

Model: modeliranje - uporaba metod podatkovnega rudarjenja in strojnega učenja

Assess: ocenjevanje zanesljivosti in uporabnosti modelov

- SAS Institute Inc., statistics & business intelligence software.
- Zaporedje korakov.
- Poslovni del izvzet.

Vsebina predmeta - metode

- **Podatki in problemi**
  - naročnik/stranka želi, da podatke “analiziramo”
    - (včasih) ni očitno kaj zanimivega se da z njimi narediti
  - spoznavanje področja, konkretnega problema in podatkov na razpolago:
    - kako so bili podatki zbrani?
    - kakšen je pomen posameznega podatka?
    - kakšne so običajne vrednosti? (čudne, nemogoče)
    - distribucije vrednosti, osamljene primere
    - pomen neznanih vrednosti?

Vsebina predmeta - metode

- **Podatki in problemi - II**
  - podatke obdelamo:
    - neznane vrednosti zamenjamo z najpogostejšimi, povprečnimi, optimističnimi, pesimističnimi vrednostmi
    - zvezne vrednosti včasih diskretiziramo ali obratno
  - nekatere primere ali spremenljivke zavržemo, da nas ne motijo pri nadaljnjem delu
  - ...

Vsebina predmeta - metode

- **Vizualizacija**
  - navaden izris pogosto zadošča za analizo:
    - “pametna analiza” in “pametne slike” dovolj, da izvemo, kar nas zanima
  - diagrami, zemljevidi, grafi, histogrami, razsevni diagrami, mozaični (parketni) diagrami, ...

Vsebina predmeta - metode

- **Modeliranje**
  - model = klasifikator ali regresor
    = formalno zapisano pravilo, s katerim lahko iz vrednosti znanih spremenljivk napovemo vrednosti neznanih
  - poudarek na matematično preprostih modelih: naivni Bayesov klasifikator, drevesa, povezovalna pravila (kdor kupi šunko, kupi tudi sir), linearna regresija (iz statistike)
  - izogibali se bomo črnim škatlam, ki ne nudijo vpogleda v odločitev: metoda podpornih vektorjev, nevronske mreže, ...
  - razvrščanje v skupine

Vsebina predmeta - metode

- **Vrednotenje modelov**
  - priprava ločenih testnih podatkov
  - preverjanje modelov na testnih podatkih
  - uporaba statistik primernih za problem:
    - odstotek pravilnih napovedi in pravilnost rangiranja
      (klasifikacijska točnost in površina pod krivuljo ROC)
    - točnost rangiranja
      (Google, težko govorimo o deležu pravilnih odgovorov, ker na stotine, morda bolje osredotočiti na nekaj prvih dest in uporabnike prositi, da ocenijo kvaliteto zadetkov Googla)

Vsebina predmeta - metode

- **Uporaba rezultatov**
  - poročilo o analizi in modeliranju
  - uporaba modela v aplikaciji

Povezava z drugimi predmeti

- **Želeno predznanje**
  - statistika (zahteva *matematično pravilen* model, pravilne predpostavke)
  - umetna inteligenca (vseeno za pravilnost in resnico, zahteva *delujoč* sistem)
  - strojno učenje (vseeno za pravilnost predpostavk, zahteva *točen* model, preverimo empirično in ne toliko formalno)
  - matematika
  - podatkovne baze

- Glavna razlika:
  - Podatkovno rudarjenje počne podobne reči kot umetna inteligenca, strojno učenje in statistika, a vendar vse iz vidika podatkov: **Odkod izvirajo podatki, kako jih kombinirati, kaj narediti z njimi....**
  - Model mora biti **uporaben**.

Podatki

http://goo.gl/forms/XWqp1de5wM

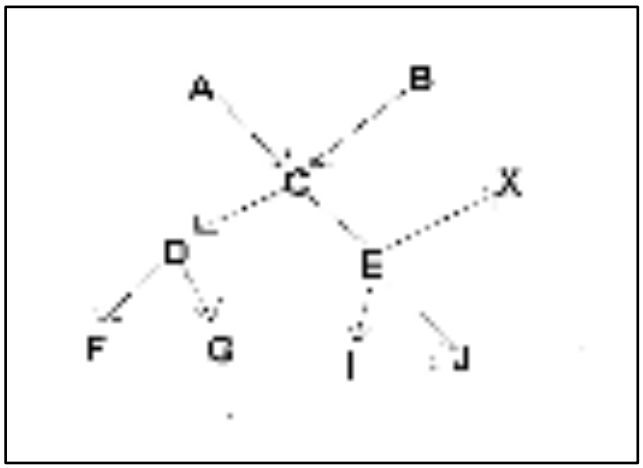
Podatki v raznih oblikah

- zbirke besedil, na primer maili
- genetski podatki, na primer sekvence (AAGCCTTCTA) ali aktivnost genov
- gibanje delnic, deviznih tečajev
- ocene filmov; Netflixovo tekmovanje, kjer pol milijona ljudi ocenjuje 18000 filmov
- pogostost priimkov po regijah
- seznami prijateljev na Facebooku
- zbirke slik, na primer lokalno, na disku, kot jih vidi Picasa, ali na spletu
- trgovinski računi
- lokacije izbruhov bolezni
- posnetki, ki jih dela Hubbleov teleskop
- socioekonomski podatki
- posnetki EKG

Ponavadi delamo s podatki v dveh oblikah

Tabelarični

Relacije med pari objektov (npr, socialna omrežja, ..)



Tabelarični podatki

stolpci: atribut, spremenljivka, lastnost, značilka (attribute, variable, feature)

- zvezna (višina, starost, ..),
- diskretna (naštevna – spol, barve, …; urejena – majhno, srednje, veliko)

vrstice: primeri, meritve, vzorci (example, data instance)

- *razred, oznaka (label)*: posebna spremenljivka, ki jo želimo napovedovati (odvisna spremenljivka)

vzorec: množica učnih in testnih primerov (sample, data set, training data, test data)

Tabelarični podatki – pretvorba v “standardno” obliko

- **Maili**: spam ali ne? Vsaka vrstica ustrezala enemu mailu, stolpci ustrezajo eni besedi. Vsak stolpec pove, ali se je določena beseda pojavila ali ne (0 ali 1) ali pa, kolikokrat se je pojavila. Razred pove ali gre za spam ali ne. Namesto besedam lahko stolpci ustrezajo n-gramom črk ali besed (n-terice zaporednih črk oz. besed v mailu).
- **Genetika**: podobno kot pri besedilih, le da težje definiramo "besedo"; iščemo lahko pogosta podzaporedja in to uporabljamo kot besede.
- **Gibanje delnic**: opazujemo lahko, recimo, kako si sledijo vrhovi; s časovnimi podatki je sicer kar zoprno delati – kako jih predelamo zelo odvisno od tega, kaj nas zanima.

Tabelarični podatki – pretvorba v "standardno" obliko

- **Ocene filmov**: v bistvu že imamo tabelo – primeri so filmi, vsak atribut predstavlja enega ocenjevalca; vendar je ta tabela preogromna za normalno delo, zato jo je potrebno zmanjšati; k temu, kako se to počne, se bomo še vrnili.
- **Pogostost priimkov**: tudi to dobimo kar v tabelarični obliki
- **Seznami prijateljev**: seznam parov prijateljev.
- **Zbirke slik**: vsako sliko moramo pretvoriti v nekakšne značilke (številke, ki dobro opišejo sliko - odvisno od tega, kaj pravzaprav želimo narediti z zbirko).
  Recimo, obrazi: številke takšne, da imajo obrazi različnih oseb čim bolj različne, obrazi enakih oseb pa čim bolj enake številke. To se sicer danes dela rutinsko in ni tako težko, je pa matematično prezahtevno za ta predmet.

Tabelarični podatki – pretvorba v "standardno" obliko

- **Računi**: Predelava v tabelarično obliko je spet precej odvisna od tega, kaj bomo počeli; lahko je vsak primer en nakup, atributi predstavljajo vse artikle; lahko pa so en primer vsi nakupi ene stranke v enem tednu (mesecu); atributi so lahko kategorije artikov...
- **Hubble**: vsaj primer je en objekt na sliki (zvezda, galaksija), atributi so moč frekvence svetlobe; razred je, ali gre za zvezdo, galaksijo, kvazar...
- **EKG**: imamo čas in napetost, iz tega moramo naračunati neke številke, recimo naklone krivulje v različnih segmentih, ki bodo povedale kaj o bolezni, ki jo želimo diagnosticirati.

Preverjanje pravilnosti podatkov

- Običajni obsegi vrednosti, možna odstopanja?
- Očitno napačni (deli) podatkov?
- Konsistentnost? (enake merske enote, …)
- Uporabni za modeliranje? So vsi podatki dosegljivi v času uporabe modela? (npr, ob nalaganju pacienta v rešilca laboratorijski testi niso še dosegljivi)
- Povezave med atributi, ki se jih moramo znebiti?

Preverjanje pravilnosti podatkov - II

- Korelacije med podatki? (npr, starost in krvni pritisk, starejši večji pritisk; smiselno gledati pritisk glede na starost)
- Redundatni atributi.
- Interakcije med atributi? (npr, bolezen prizadane mlajše moške in starejše ženske; starost in spol tako sama zase ne povesta nič o bolezni)

Preverjanje pravilnosti podatkov - III

- Atribut z veliko napakami.
  - zvezne diskretiziramo, pretvorimo v intervale (npr, višino, čeprav zgrešimo za 10cm, bo visoka oseba vedno označena kot takšna)
  - v celoti izpustimo, raje kot da nas napačna vrednost zavede
  - združimo z atributi s podobnim pomenom, ki so prav tako nezanesljivi

Preverjanje pravilnosti podatkov - IV

- Atribut z manjkajočimi vrednostmi.
  - uporabimo metode, ki znajo takšne ignorirati (npr, naivni Bayesov klasifikator),
  - če veliko manjkajočih, ga raje izpustimo
  - uganemo manjkajoče vrednosti oz. vprašamo tistega, ki podatke pridobil
  - napovemo vrednost iz vrednosti drugih atributov
  - vstavimo najpogostejšo ali povprečno ali pesimistično ali optimistično vrednost

Preverjanje pravilnosti podatkov - V

- Pretvorba atributov.
  - številske diskretiziramo:
    - na enako široke intervale
    - enako pogoste intervale
    - na intervale z enakimi lastnostmi (npr, če se bolezen začne pojavljati pri določeni starosti, paciente razdelimo na mlajše in starejše od te meje)
  - diskretne pretvorimo v številske:
    - binarni postanejo 0 in 1
    - iz večvrednostih naredimo toliko novih atributov kolikor imajo vrednosti, eden od teh bo 1, vsi ostali 0.

Primeri lanskih projektov

https://goo.gl/whJGNw

Feedback

http://goo.gl/forms/JQgtEBnvW9