NASLOVNICA

INFORMACIJE

6 faz (podrobno v e-učilnici)

Programska orodja
- WEKA
- MS Excel / LO Calc
- Python

Ocena:
- Izpit: 50% (lahko s kolokviji, vsak vsaj 40%, avg. 50%)
- Kratke DN: 15% (9-10 nalog a.k.a. kvizov)
- Projekt: 30% → 1. md na GH repo
- Ustni izpit: 5%
- Izračun: $0{,}15 \times DN + 0{,}5 \times PI + 0{,}3 \times PP + 0{,}05 \times UP$

Kolokviji:
1. 21.11.2024
2. 16.01.2025
- Posebne prijave (mimo ŠIS-a)

Prijava na izpit v ŠIS

10.10.24

STROJNO UČENJE, AI, PODATKOVNO RUDARJENJE

Trendi, ki vodijo v podatkovno poplavo
- Generiranje velikega št. podatkov:
  - Poslovne ustanove (npr. banke)
  - Znanstveni podatki (npr. biologija)
  - Splet (npr. socialna omrežja)

Primeri ogromnega št. podatkov
- Evropska VLBI mreža teleskopov generira 1 gigabit podatkov na sekundo za vsak teleskop v 25 dneh.
- AT&T posreduje več kot milijardo klicev na dan

Največje zbirke podatkov leta 2003
- Komercialne DB:
  - France Telecom: 30TB
  - AT&T: 26TB
- Splet:
  - Alexa internetni arhiv: 7 let = 500TB

Kaj pa danes?
- Konec junija 2017 CERN shranjeval 200PB podatkov

Področja uporabe
- Znanost (npr. bioinformatiki)
- Poslovni svet - profiliranje strank
- Splet
- Državna uprava

Prikladni problemi za podatkovno rudarjenje
- Zahtevajo odločitve na podlagi znanj
- Imajo opremljajoče okolje
- Imajo ne-optimalne trenutne rešitve
- Imajo veliko poplačilo za prave odločitve
- Paziti moramo na zasebnost!

Def. „Odkrivanja zakonitosti"
- ... v podatkih je:
  - Netrivialni proces identifikacije veljavnih, novih potencialno uporabnih in razumljivih vzorcev v podatkih

Statistika, str. učenje in PR
- Statistika:
  - Bolj usmerjena v teorijo
  - Večji fokus na testiranju hipotez
- Strojno učenje
  - Uporaba hevristik
  - Usmerjeno v izboljševanje performans učnih agentov
  - Se ukvarja z učenjem v realnem času in robotiko
- Podatkovno rudarjenje
  - Statistika & Strojno učenje
- Razmejitve so „mehke"

Zgodovina
- 1960 - Data fishing, dreding → uporabljeno s strani statistikov
- 1990 - Data mining - uporabljajo DB strokovnjaki
- 1989 - Knowledge Discovery in DB - Uporabljajo AI, ljudje s področja strojnega učenja

Tehnike str. učenja
- Klasifikacija
- Razvrščanje v skupine
- Asociacije
- Vizualizacija
- Povzemanje
- Odkrivanje odstopanj
- Regresija/ocenjevanje
- Analiza povezav

Klasifikacija
- S pomočjo statistike, odločitvenega drevesa, nevronske mreže,...

Razvrščanje v skupine
- Razvrstimo v skupine po razdalji

PR PO CRISP-DM STANDARDU

CRISP-DM
- CRoss-Industry Standard Process for Data Mining

Zakaj standarizirati?
- PR je proces, ki mora biti ponovljiv in zanesljiv
- Nudenje okvira za shranjevanje preteklih izkušenj
  - Projekte lahko repliciramo
- Lažje planiranje in vodenje projektov
- Faktor udobja za nove uporabnike
  - Dokazuje zrelost PR
  - Zmanjšuje odvisnost od tk. super-strokovnjakov

Kako je prišlo do standardizacije
- Iniciativa s strani izkušenejših podjetji
- Razvoj in dopolnitve na workshop-ih
- K procesnemu modelu prispevalo 300+ organizacij
- Leta 1999 izide CRISP-DM 1.0

CRISP-DM je:
- Ne-lastniški
- Neodvisen od aplikacije/problema
- Neodvisen od orodja/programa
- Osredotočen na poslovne probleme
- Ogrodje, ki služi kot vodilo
- Baza iz preteklih izkušenj

CRISP-DM pregled:
- Podatkovno rudarjenje kot metodologija
- Za vsakogar
- S „komplet" navodili
- V ciku iz 6 faz

CRISP-DM faze
1. Razumevanje problema
2. Razumevanje podatkov
3. Priprava podatkov
4. Modeliranje
5. Evaluacija (ocena modelov)
6. Predaja končnemu uporabniku

Razumevanje problema
- Določiti cilj projekta
- Oceniti trenutno situacijo
- Določiti cilje pod. rudarjenja
- Narediti projektni plan

Razumevanje podatkov
- Zbrati začetne podatke
- Opisati podatke
- Raziskati podatke
- Preveriti kakovost podatkov

Priprava podatkov
- Izbrati podatke
- Prečistiti podatke
- Sestaviti podatke