# Uvod u strojno učenje

## Poglavlje 18.1

Prema slajdovima Stuarta Russela, Krunoslava Puljića, Tomislava Šmuca, Hantao Zhanga (hvala)!

# Uvod

Stojno učenje:

- eng. Machine Learning (ML)

Povijesno:

- razvoj ML započeo prije 50-tak godina

Problem:

- Kako indukcijom/generalizacijom izvesti novo znanje iz primjera/podataka?

# Motivacija

1. Automatizirano prikupljanje znanja
   - direktno iz podataka
2. Otkrivanje znanja u skupovima podataka
   - Ima li u njima znanja?
   - Rastuće količine informacija
   - eng. data mining, text mining
3. Klasične metode nisu dovoljne jer su neki problemi preteški
   - Napredak u izgradnji algoritama i teorije

# Automatizirano prikupljanje znanja

- Prikupljanje znanja
  - Izuzetno složen problem

- Primjer: kako zapisati znanje o igranju šaha?
  - Što znači "slaba struktura pješaka"?
  - Što znači "dobro zaštićen kralj"?
  - Kako znanje velemajstora oblikovati u ako–onda pravila

- Ideja: stvoriti sustav koji će automatski stjecati znanje
  (visoko apstraktne koncepte ili strategije rješavanja problema)
  kroz iskustvo, tj. primjere, . . . kao što to čini čovjek

# Strojno učenje

Strojno učenje bavi se izgradnjom računarskih sustava koji automatski poboljšavaju svoje performanse kroz iskustvo.

- Tom M. Mitchell — definicija strojnog učenja:
  - "A computer program is said to learn from experience $E$ with respect to some class of tasks $T$ and performance measure $P$, if its performance at tasks in $T$, as measured by $P$, improves with experience $E$."

- Formalno:
  - Računalni program uči ako se performanse izvršavanja zadatka $T$, mjerene pomoću mjere uspješnosti $P$, poboljšavaju s iskustvom $E$

- Nema univerzalnog algoritma za učenje
  - ipak, izumljeni su učinkoviti algoritmi koji (uspješno) rješavaju određen tip problema

# Primjeri

- $T$: igranje šaha / backgammona / ...
  - $P$: postotak dobivenih igara
  - $E$: igranje protiv samoga sebe
- $T$: raspoznavanje rukom pisanih slova
  - $P$: postotak točno prepoznatih slova
  - $E$: baza rukom pisanih slova
- $T$: detekcija tumorskih regija na CT-u
  - $P$: postotak točno prepoznatih regija
  - $E$: baza označenih CT slika

# Primjene strojnog učenja

- Otkrivanje znanja u (velikim) skupovima podataka
- Programske implementacije koje nije moguće riješiti klasičnim programiranjem
- Prilagodljivi programski sustavi
- Bioinformatika
- Obrada prirodnog jezika
- Raspoznavanje govora
- Raspoznavanje uzoraka
- Inteligentno upravljanje
- Predviđanje trendova
- itd.

# Vrste strojnog učenja

- Nadzirano učenje (engl. supervised)
- Nenadzirano učenje (engl. unsupervised)
- Podržano učenje (engl. reinforcement learning)
- Ostala učenja
  - Polu–nadzirano učenje
  - Transduktivno učenje
  - Relacijsko učenje
  - Genetsko programiranje
  - ...

# Strojno učenje

Podaci su u obliku $(x, y)$

– $x$ ulazna vrijednost, $y$ ciljna vrijednost

Učimo nepoznatu funkciju

$$f(x) = y$$

koja primjeru $x$ pridružuje klasu ili realnu vrijednost $y$.

Traži se aproksimacija $h$ — hipoteza.

Klasifikacija, parametarski prikaz, neparametarski prikaz podataka.

# Nenadzirano učenje

- Kao tek rođeno dijete
- Nema učitelja – nema nekoga da nam pokaže pravo / točno znanje
- Bitna je samoorganizacija
- Potrebno je:
  – pronaći pravilnosti u podacima,
  – otkriti ih automatski,
  – pronaći neku reprezentaciju,
  – odrediti što podaci znače

# Algoritmi nenadziranog učenja

- Klasteriranje (engl. clustering) — K–sredina (engl. K–means)
- Učenje asocijacijskih pravila
- Redukcija dimenzionalnosti podataka

# K–sredina

- K–sredina (engl. K–means)
- metoda klasteriranja koja pokušava rasporediti $n$ mjerenja u $k$ klastera, tako da svako mjerenje smjesti u klaster s najbližom srednjom vrijednošću (centrom klastera)
  – Minimizira se suma kvadrata udaljenosti unutar klastera

$$\arg\min_{S} \sum_{i=1}^{k} \sum_{x_j \in S_i} \|x_j - \mu_i\|^2$$

- Problem je NP–težak
  – Ipak, postoje efikasni heuristički algoritmi

# Učenje asocijacijskih pravila

- Učenje asocijacijskih pravila (engl. association rule learning)
- Potraga za vezom između varijabli
- Primjer: supermarket može odrediti koji se proizvodi često kupuju zajedno te iskoristi tu informaciju u marketniške svrhe – analiza kupovne košarice

# Učenje asocijacijskih pravila

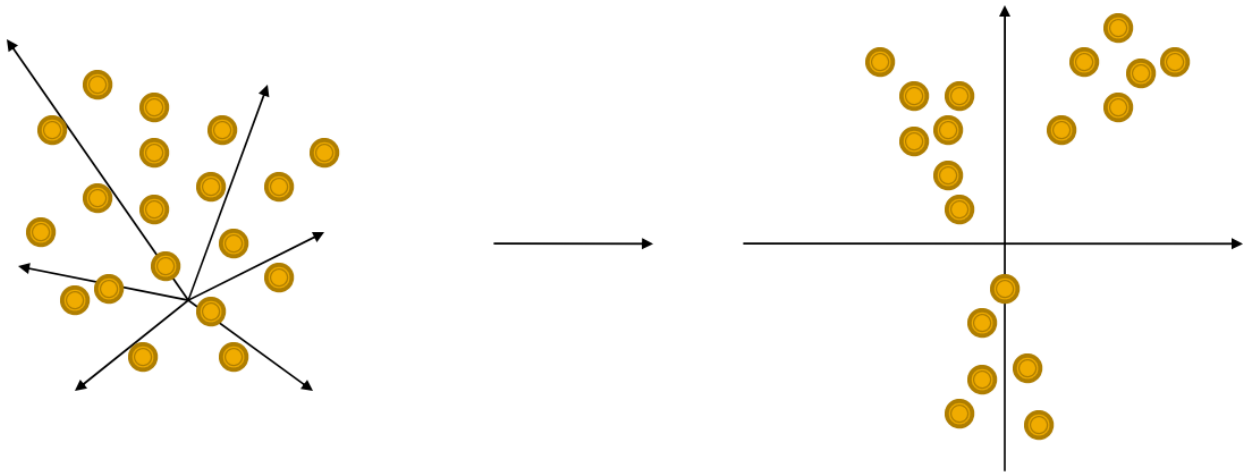
- Skup transakcija iz Konzuma

| ID | mlijeko | kruh | pivo | pelene | riba | meso |
|---|---|---|---|---|---|---|
| 1 | 1 | 1 | 1 | 1 | 0 | 0 |
| 2 | 0 | 1 | 1 | 0 | 1 | 1 |
| 3 | 0 | 1 | 1 | 1 | 1 | 1 |
| 4 | 0 | 0 | 1 | 1 | 0 | 1 |

- Učenje pravila:
  - $\{pelene \rightarrow pivo\}$
  - $\{meso, riba \rightarrow kruh\}$

# Analiza glavnih komponenti

- Analiza glavnih komponenti (engl. Principal Component Analysis)
- Metoda za redukciju dimenzionalnosti



- Matematička procedura koja koristi ortogonalne transformacije da pretvori skup mjerenja moguće koreliranih varijabli u skup vrijednosti nekoreliranih varijabli
  - Nekorelirane varijable se nazivaju glavnim komponentama
  - Broj glavnih komponenti je $\leq$ broj originalnih varijabli.

# Primjena učenja bez nadzora

- Analiza ponašanja kupaca
- Segmentiranje tržišta
- Grupiranje teksta po sličnosti
- Grupiranje fotografija po sadržaju
- Vizualizacija podataka
- Početna analiza podataka (otkrivanje neznane strukture)
- ...

# Učenje s nadzorom

- Učenje s nadzorom (engl. Supervised learning)
- Kao dijete koje je ovladalo komunikacijom
- Sada dobivamo eksplicitnu informaciju za svaki primjer koji učimo
- Cilj nam je naučiti raditi predikciju na novim,
  još neviđenim primjerima

# Vrste učenja s nadzorom

- Regresija
  - učenje funkcija
  - Eksplicitna informacija o objektu je numerička
- Na primjer: cijena zlata, temperatura, tlak, broj cipela, broj otkucaja srca, ...
  - Cilj je aproksimacija funkcije
- Klasifikacija
  - učenje klasificiranja (raspoznavanje uzoraka)
  - jedna ili više klasa
  - Eksplicitna informacija o objektu je kategorijska
- Na primjer: Dobar / Loš, Tumor / nije tumor, Crven / Zelen / Žut / Plav, ...
  - Cilj je izgraditi funkciju koja će za svaki objekt reći kojoj klasi tj. kategoriji pripada

# Algoritmi učenja s nadzorom

- Stabla odluke
- Štreber
- K najbližih susjeda
- Naivni Bayesov klasifikator
- Učenje pravila

# Primjena učenja s nadzorom

- Predviđanje kretanja cijena dionica
- Klasifikacija teksta
- Detekcija tumorskih regija
- Detekcija bolesti prema analizi DNA
- Prepoznavanje lica
- Prepoznavanje govornika
- Klasifikacija prebjega u drugi telekom
- Podrška u odlučivanju pri izdavanju kredita
- ...

# Podržano učenje

- Podržano učenje (engl. Reinforcement learning)
- učenje podrškom
- Kao dijete koje uči voziti bicikl
- Učenje u kojem nam učitelj daje ocjenu koliko smo dobro nešto obavili
  – pali smo s bicikla ili nismo
- Prolazimo seriju stanja i akcija i tek na kraju dobivamo (ili ne dobivamo) nagradu
- Istražiti stanja i akcije koje vode do cilja
- Cilj nam je maksimizirati ukupnu sumu "nagrada" na kraju

# Algoritmi podržanog učenja

- Q-learning
  - Svodi se na metodu pokušaja i pogrešaka
  - Pretraživanje radi dostizanja nagrade
  - Tražimo optimalnu akciju u trenutnom stanju
  - To je akcija koja maksimizira sumu trenutne i odgođene nagrade u slučaju da slijedimo optimalnu strategiju
- Temporal difference learning
  - Kombinacija Monte Carlo ideje i dinamičkog programiranja

# Primjena podržanog učenja

- Igranje igara
- Robotsko kretanje
- Autonomna navigacija
- Učenje kontrolnih strategija
- ...

# Induktivno učenje

- . . . ili učenje iz primjera
- $h$ je hipoteza (model) koja reprezentira/aproksimira ciljni koncept $c$
  – u idealnom slučaju je $h(x) = c(x)$
- ono što u najboljem slučaju možemo garantirati učenjem nekim algoritmom strojnog učenja jest da naučeni $h$ dobro aproksimira ciljni koncept $c$ nad skupom primjera za učenje $T$
- Osnovna hipoteza induktivnog učenja:
  – Bilo koja hipoteza koja dobro aproksimira ciljni koncept na dovoljno velikom skupu primjera dostupnih za učenje, isto će tako dobro aproksimirati ciljni koncept i na novim, još nedostupnim primjerima.

# Redoslijed nastavka o strojnom učenju

Redoslijed u nastavku priče o strojnom učenju:

dijelom ide prema vrstama učenja — pod nadzorom, bez nadzora.

Matematički i algoritamski — prirodnija podjela je prema:

načinima zaključivanja — "egzaktno" i probabilističko,

modelima učenja — klasifikacija, parametarski i neparametarski.

Poredak

1. Stabla odlučivanja — klasifikacija, "egzaktno"
2. Regresija — parametarski, "egzaktno"
3. Najbliži susjedi — neparametarski, "egzaktno"
4. k-sredine — bez nadzora, klasifikacija, "egzaktno"
5. Bayesova (opća i naivna) klasifikacija — probabilističko

Posebna stvar na kraju

6. Neuralne mreže — parametarski.

# Izvorni materijali za ove slajdove

- Krunoslav Puljić
  – http://web.math.hr/nastava/ui/
- Tomislav Šmuc, IRB
  – http://web.math.hr/nastava/su/
- Jan Šnajder, Bojana Dalbelo–Bašić, FER
  – http://www.zemris.fer.hr/predmeti/su/
- Hantao Zhang, University of Iowa, CS4420 (22C:145)
  – http://www.cs.uiowa.edu/~hzhang/c145/
  – http://homepage.cs.uiowa.edu/~hzhang/c145/notes/