

MAŠINSKO UČENJE

JELENA JOVANOVIĆ

Email: jeljov@gmail.com

Web: <http://jelenajovanovic.net>

PREGLED PREDAVANJA

- Šta je mašinsko učenje?
- Zašto (je potrebno/bitno) m. učenje?
- Oblasti primene m. učenja
- Oblici m. učenja
- Osnovni koraci i elementi procesa m. učenja
 - Podaci
 - Atributi (features)
 - Odabir algoritma m. učenja
 - Validacija i testiranje kreiranog modela

ŠTA JE MAŠINSKO UČENJE ?

Mašinsko učenje se odnosi na sposobnost softverskog sistema da:

- *generalizuje* na osnovu prethodnog *iskustva*, i
- da koristi ove generalizacije kako bi pružio odgovore na pitanja koja se odnose na entitete/pojave koje pre nije sretao

Iskustvo = skup podataka o pojavama/
entitetima koji su predmet učenja



ŠTA JE MAŠINSKO UČENJE ?

Za kompjuterski program se kaže da uči
iz iskustva ***E*** (*experience*),
vezanog za zadatak ***T*** (*task*), i
meru performansi ***P*** (*performance*),
ukoliko se njegove performanse na zadatku ***T***, merene
metrikama ***P***, unapređuju sa iskustvom ***E***

Tom Mitchell (1997)

ŠTA JE MAŠINSKO UČENJE ?

Primer: program koji označava poruke kao spam i not-spam

- Zadatak (***T***): klasifikacija email poruka na spam i not-spam
- Iskustvo (***E***): email poruke označene kao spam i not-spam; “posmatranje” korisnika dok označava email poruke
- Performanse (***P***): procenat email poruka korektno klasifikovanih kao spam/not-spam

ZAŠTO MAŠINSKO UČENJE ?

1) Neke vrste zadataka ljudi rešavaju vrlo lako, a pri tome nisu u mogućnosti da precizno (algoritamski) opišu kako to rade

Primeri: prepoznavanje slika, zvuka, govora

2) Za neke vrste zadataka mogu se definisati algoritmi za rešavanje, ali su ti algoritmi vrlo složeni i/ili zahtevaju velike baze znanja

Primeri: automatsko prevođenje

ZAŠTO MAŠINSKO UČENJE ?

3) U mnogim oblastima se kontinuirano prikupljaju podaci sa ciljem da se iz njih “nešto sazna”; npr.:

- u medicini: podaci o pacijentima i korišćenim terapijama
- u sportu: o odigranim utakmicama i igri pojedinih igrača
- u marketingu: o korisnicima/kupcima i tome šta su kupili, za šta su se interesovali, kako su proizvode ocenili,...

Analiza podataka ovog tipa zahteva pristupe koji će omogućiti da se otkriju pravilnosti, zakonitosti u podacima koje nisu ni poznate, ni očigledne, a mogu biti korisne

GDE SE PRIMENJUJE MAŠINSKO UČENJE ?

Brojne oblasti primene

- Kategorizacija teksta prema temi, iskazanim osećanjima i/ili stavovima i sl.
- Mašinsko prevođenje teksta
- Razumevanje govornog jezika
- Prepoznavanje lica na slikama
- Segmentacija tržišta
- Uočavanje paternu u korišćenju različitih aplikacija
- Autonomna vozila (self-driving cars)
- ...

OBLICI MAŠINSKOG UČENJA

Osnovni oblici mašinskog učenja:

- Nadgledano učenje (supervised learning)
- Nenadgledano učenje (unsupervised learning)
- Učenje uz podsticaje (reinforced learning)

NADGLEDANO UČENJE

Obuhvata skup problema i tehnika za njihovo rešavanje u kojima program koji uči dobija:

- skup ulaznih podataka (x_1, x_2, \dots, x_n) i
- skup željenih/tačnih vrednosti, tako da za svaki ulazni podatak x_i , imamo željeni/tačan izlaz y_i

Zadatak programa je da “nauči” kako da novom, neobeleženom ulaznom podatku dodeli tačnu izlaznu vrednost

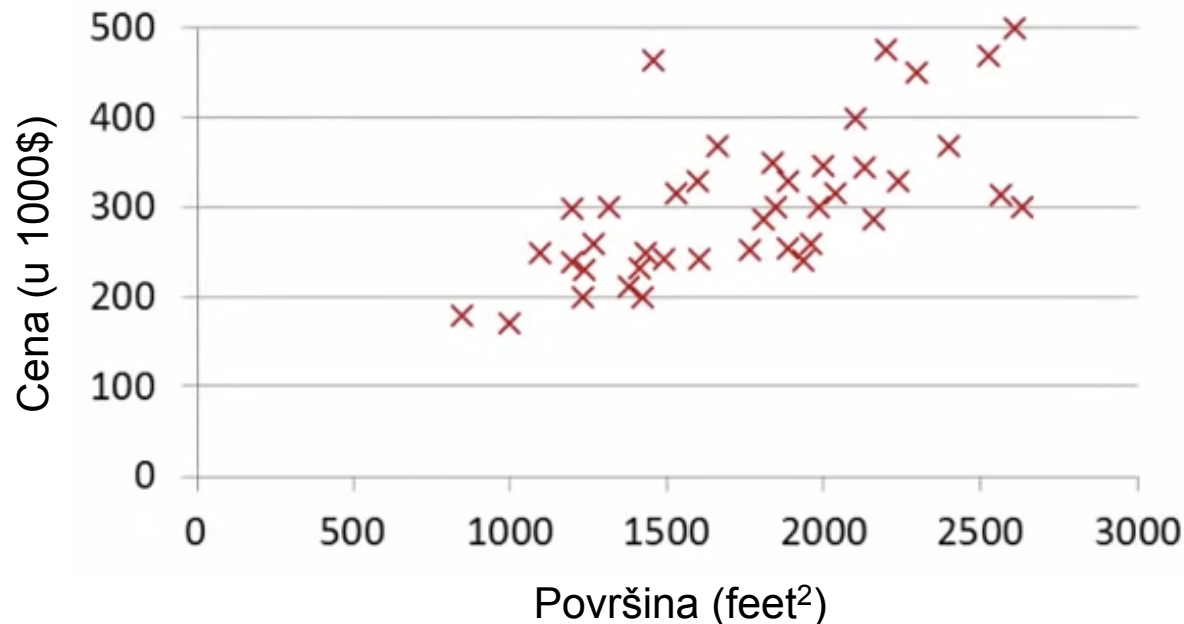
Izlazna vrednost može biti:

- labela (tj. nominalna vrednost) – reč je o *klasifikaciji*
- realan broj – reč je o *regresiji*

NADGLEDANO UČENJE

Primer linearne regresije: predikcija cena nekretnina na osnovu njihove površine

Podaci za učenje: površine (x) i cene (y) nekretnina u nekom gradu



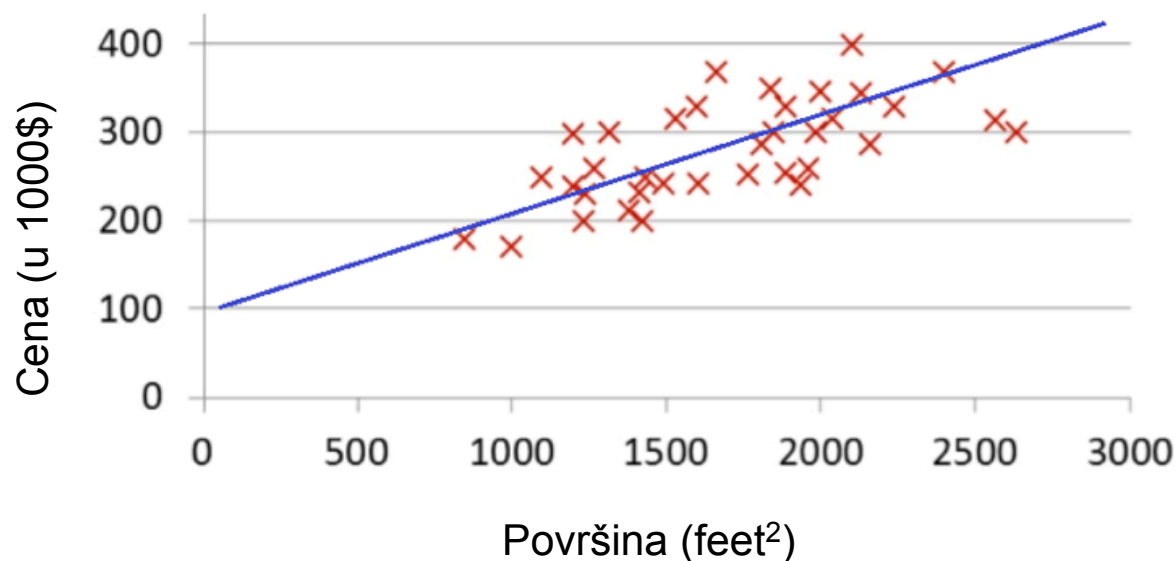
NADGLEDANO UČENJE

Primer linearne regresije (nastavak)

Funkcija koju treba „naučiti“ u ovom slučaju (samo jedan atribut) je:

$$h(x) = a + bx$$

a i b su koeficijenti koje program u procesu „učenja“ treba da *proceni* na osnovu datih podataka



NENADGLEDANO UČENJE

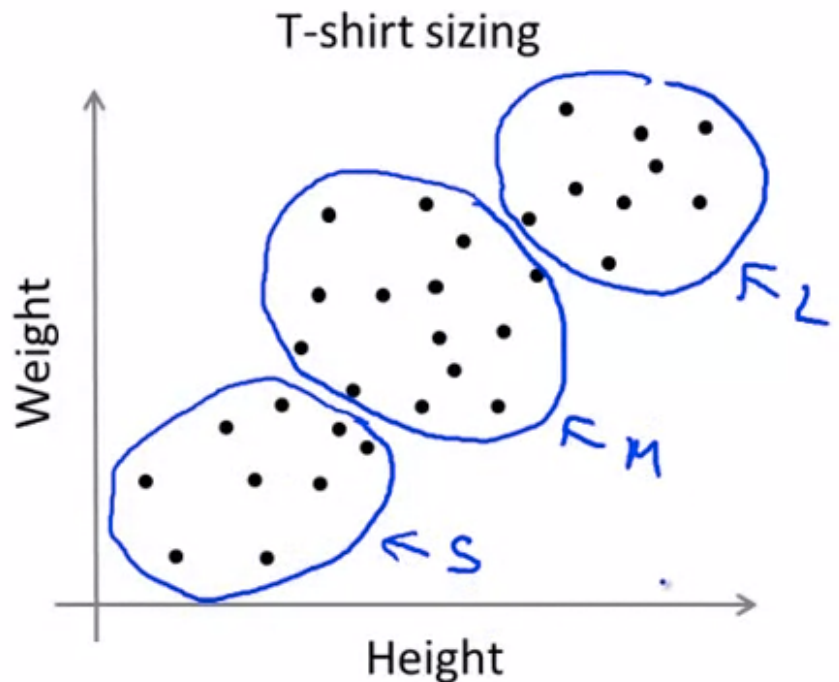
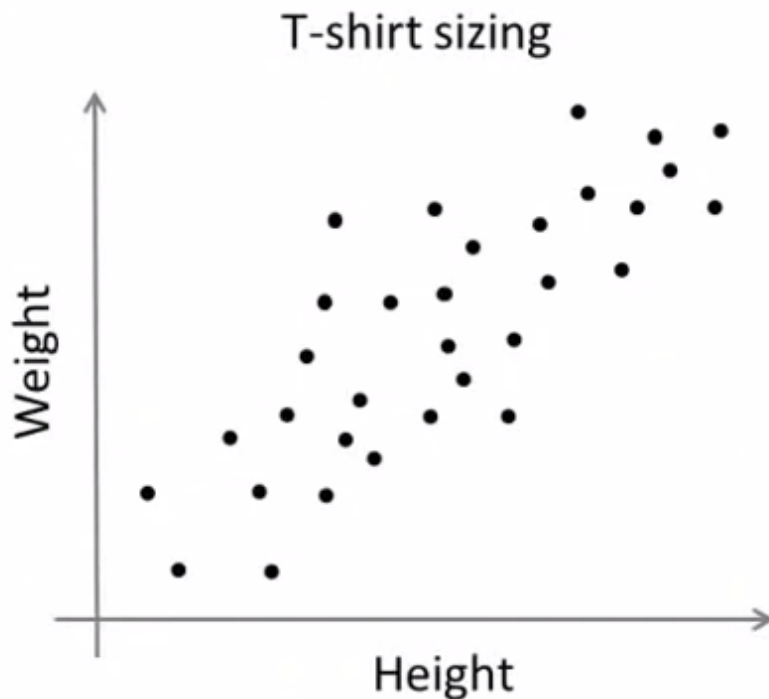
Kod nenadgledanog učenja

- nemamo informacija o željenoj izlaznoj vrednosti
- program dobija samo skup ulaznih podataka (x_1, x_2, \dots, x_n)

Zadatak programa je da otkrije paterne tj. skrivene strukture/
zakovitosti u podacima

NENADGLEDANO UČENJE

Primer: određivanje konfekcijskih veličina na osnovu visine i težine ljudi



UČENJE UZ PODSTICAJE

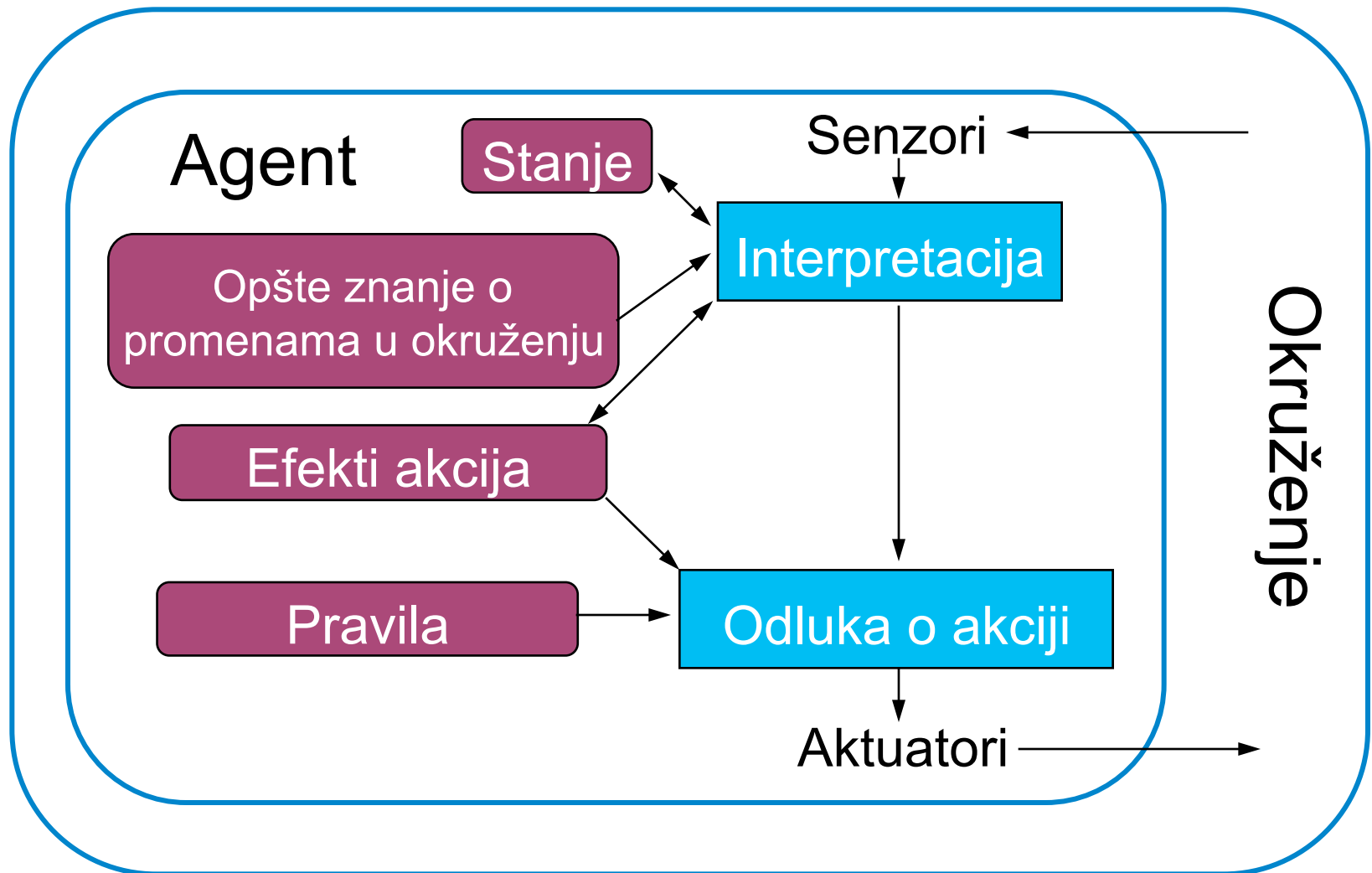
Ovaj oblik učenja podrazumeva da program (agent) deluje na okruženje izvršavanjem niza akcija

Ove akcije utiču na stanje okruženja, koje povratno utiče na agenta pružajući mu povratne informacije koje mogu biti “nagrade” ili “kazne”

Cilj agenta je da nauči kako da deluje u datom okruženju tako da vremenom max. nagrade (ili min. kazne)

Primer: autonomna vozila

ILUSTRACIJA AGENTA KOJI UČI UZ PODSTICAJE



**OSNOVNI
KORACI I ELEMENTI
PROCESA
M. UČENJA**

OSNOVNI KORACI PROCESA M. UČENJA

- 1) *Prikupljanje podataka* potrebnih za formiranje dataset-ova za obuku, (validaciju) i testiranje modela m. učenja
- 2) Priprema podataka, što tipično podrazumeva “čišćenje” i *transformaciju podataka*
- 3) Analiza rezultujućih dataset-ova, i njihovo, eventualno, dalje unapređenje kroz *selekciju/transformaciju atributa*
- 4) *Izbor 1 ili više algoritama* m. učenja
- 5) *Obuka, konfiguracija i evaluacija kreiranih modela*
- 6) Izbor modela koji će se koristiti (na osnovu rezultata koraka 5) i njegovo *testiranje*

PODACI

Podaci su potrebni za trening, validaciju i testiranje modela

- Tipična podjela podataka kojima raspolažemo je 60% za trening, 20% za validaciju i 20% za testiranje
- Izbor uzoraka za trening, validaciju i testiranje treba da se uradi na slučajan način (random selection)

Za nadgledano učenje, moramo imati “obeležene” podatke

- Npr. obeležiti slike koje sadrže lice, elektronsku poštu koja je nepoželjna, e-mail adrese koje su lažne, i sl.

PODACI

Izvori podataka:

- Javno dostupne kolekcije podataka; sve više tzv. otvorenih podataka (open data)
 - Pogledati npr. <http://bitly.com/bundles/bigmlcom/4>
- Podaci dostupni posredstvom Web API-a
 - Pogledati npr. <http://www.programmableweb.com/>
- Sve veće tržište gde je moguće kupiti podatke
 - Pogledati npr. <http://datamarket.com/>

PODACI

Preporuka: predavanje Peter Norvig*-a na temu značaja podataka za mašinsko učenje:

The Unreasonable Effectiveness of Data

URL: <http://www.youtube.com/watch?v=yvDCzhbjYWs>

*Peter Norvig je autor jedne od najpoznatijih knjiga u domenu Veštačke inteligencije i trenutno na poziciji Director of Research u Google-u

ATRIBUTI (FEATURES)

Osnovna ideja:

- pojave/entitete prepoznamo uočavajući njihove osobine (ili izostanak nekih osobina) i uviđajući odnose između različitih osobina
- omogućiti programu da koristi osobine pojava/entiteta za potrebe njihove identifikacije/grupisanja

Izazov:

- odabrati atribute koji najbolje opisuju neki entitet/pojavu, tj. omogućuju distinkciju entiteta/pojava različitog tipa

ATRIBUTI (FEATURES)

Primeri:

- Za elektronsku poštu: naslov (tj. polje subject), reči napisane velikim slovom, dužina email-a, prva reč i sl.
- Za stan: površina, lokacija, broj soba, tip grejanja i sl.
- Za tweet poruke: prisustvo linkova, prisustvo hashtag-ova, vreme slanja, pošiljalac, ...

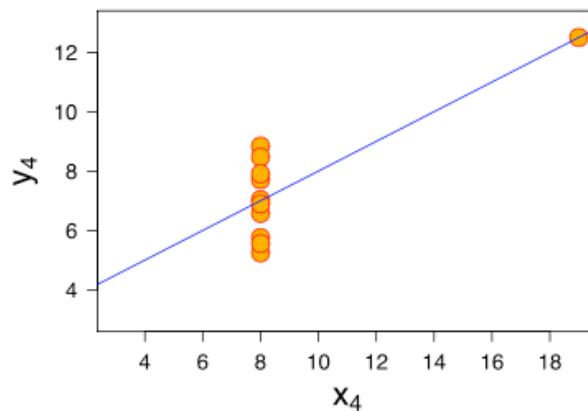
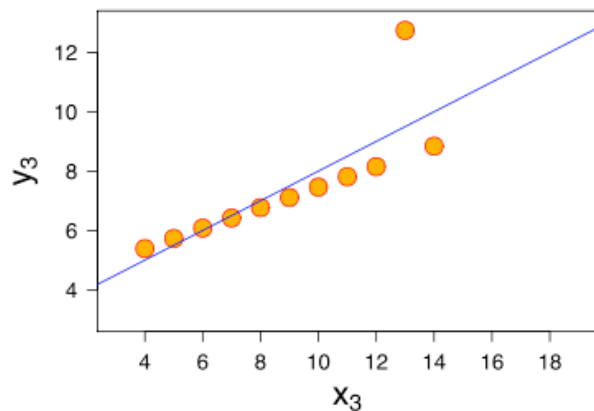
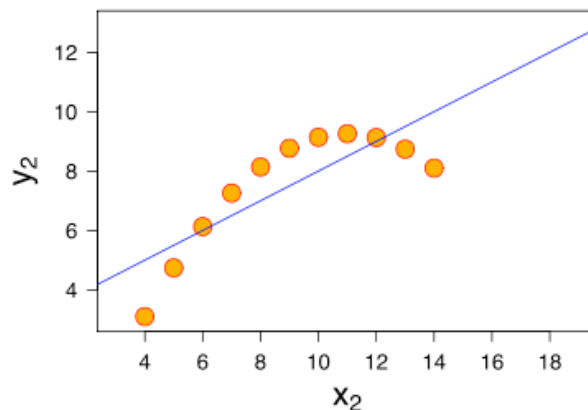
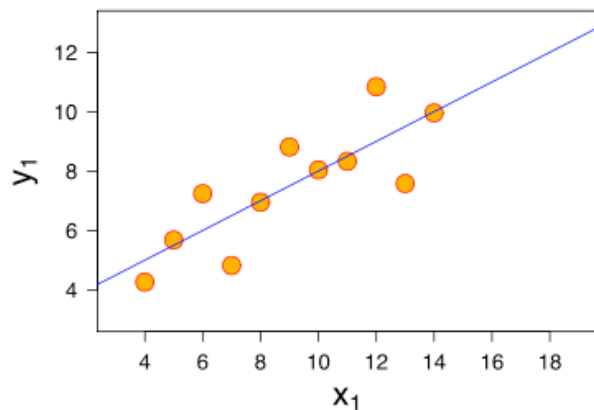
ODABIR ALGORITMA

Generalno, zavisi od:

- vrste problema koji rešavamo,
- karakteristika skupa atributa (features)
 - tip atributa i stepen homogenosti tipova i opsega vrednosti atributa
 - stepen međuzavisnosti (korelisanosti) atributa
- obima podataka koji su nam na raspolaganju

ODABIR ALGORITMA

Primer: pokušaj aproksimacije četiri različita skupa podataka primenom iste linearne funkcije (tj. linearne regresije)



TESTIRANJE

Za procenu uspešnosti modela, potrebni su podaci koje model nije imao prilike da “vidi” u fazi učenja

Reč je o podacima za testiranje, za koje se obično izdvaja 20-30% ukupnih podataka

Uspešnost modela se utvrđuje različitim metrikama: tačnost, preciznost, odziv, ...

TRAIN/VALIDATE/TEST

Pored treniranja i testiranja modela, najčešće se radi i validacija modela kako bi se:

- a) izabrao najbolji model između više kandidata
- b) odredila optimalna konfiguracija parametara modela
- c) izbegli problemi *over/under-fitting-a*

U ovim slučajevima, ukupan dataset deli se u odnosu 60/20/20 na podatke za trening, validaciju i testiranje

Podaci za validaciju koriste se za poređenje performansi

- različitih modela (a);
- izabranog modela sa različitim vrednostima parametara (b)

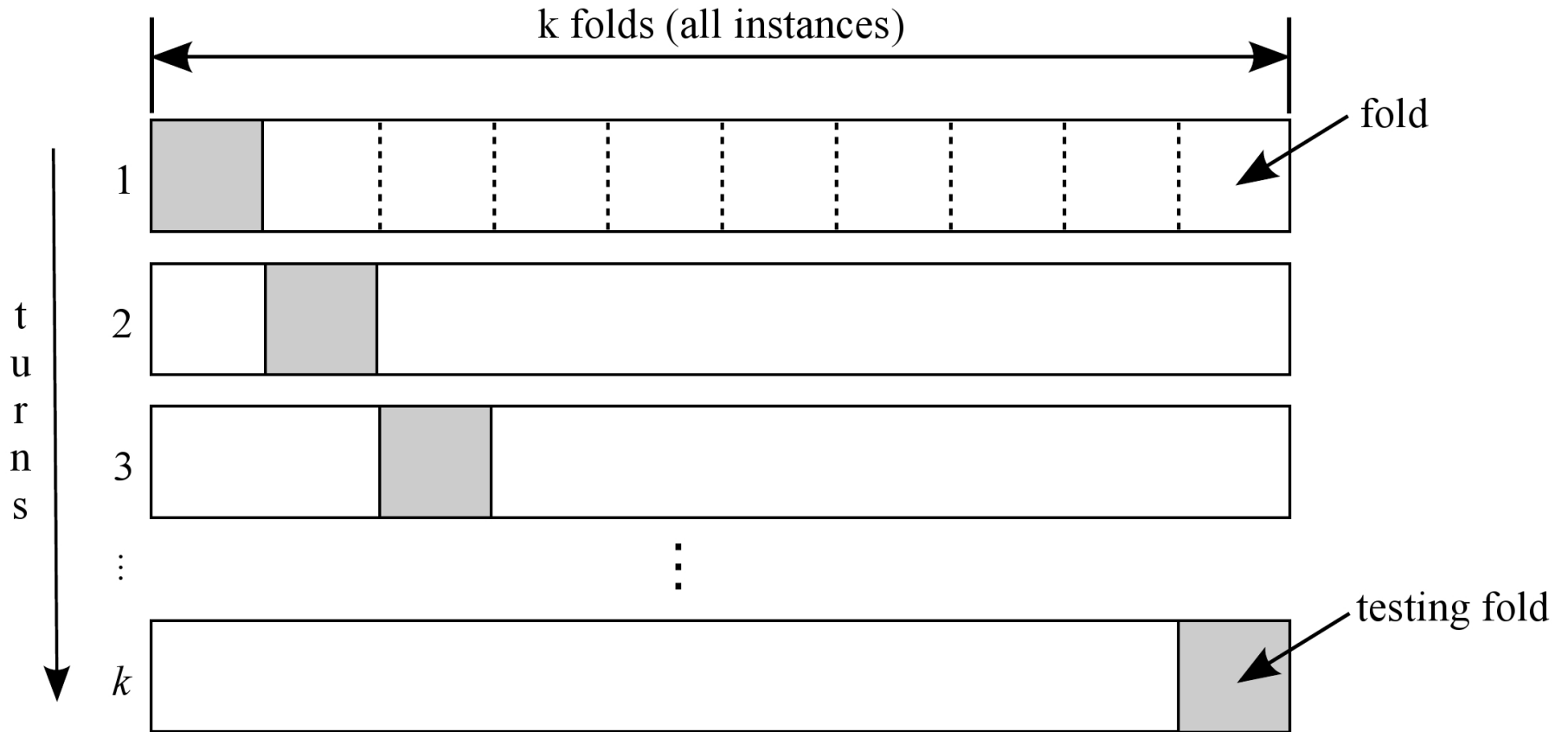
CROSS-VALIDATION

Čest pristup za efikasno korišćenje raspoloživih podataka

Kako funkcioniše:

- raspoloživi skup podataka za trening se podeli na K delova ili podskupova (*folds*)
 - najčešće se uzima 10 podskupova (*10 fold cross validation*)
- zatim se obavlja K iteracija treninga + validacije modela ; u svakoj iteraciji:
 - uzima se 1 deo podataka za potrebe validacije, a ostatak (K-1 deo) se koristi za učenje
 - bira se uvek različiti podskup koji će se koristiti za validaciju

CROSS VALIDATION



CROSS VALIDATION

Pri svakoj iteraciji računaju se performanse modela

Na kraju se računa prosečna uspešnost na nivou svih K iteracija – tako izračunate mere uspešnosti daju bolju sliku o performansama modela

Ukoliko su rezultati u svih K iteracija vrlo slični, smatra se da je procena uspešnosti modela pouzdana

ANALIZA GREŠKE

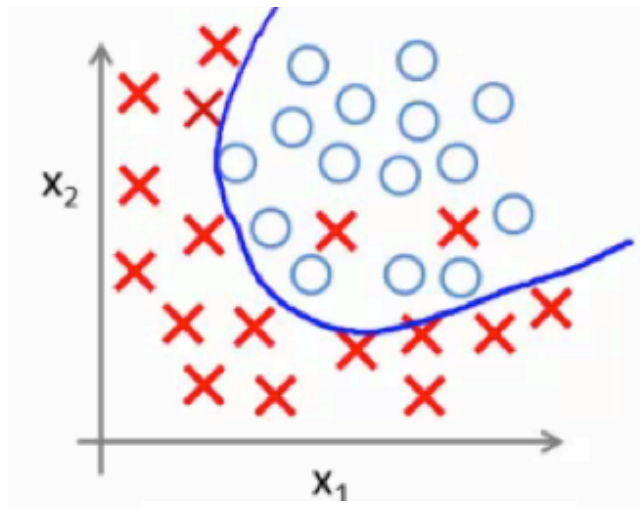
Podrazumeva “ručno” pregledanje primera na kojima je model pravio greške i uočavanje paterna u tim primerima

Pomaže da se stekne osećaj zbog čega model greši, i šta bi se moglo uraditi da se greške otklone; Npr.

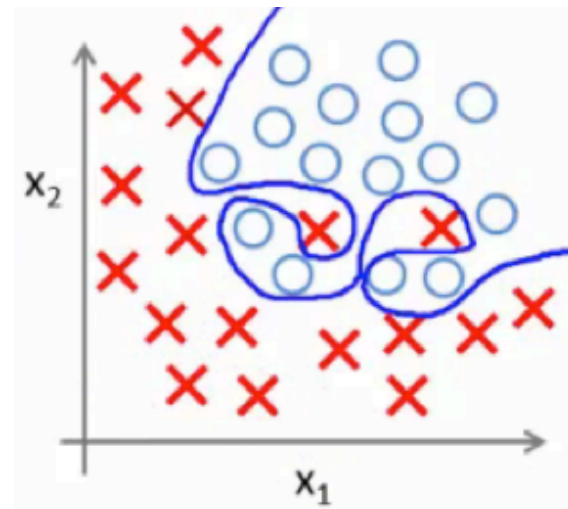
- identifikovati suvišne attribute
- identifikovati attribute koji nedostaju
- drugačije podesiti parametre modela
- ...

PROBLEM PREVELIKOG PODUDARANJA (OVER-FITTING)

Odnosi na situaciju u kojoj model savršeno nauči da prepozna instance iz trening seta, ali nije u mogućnosti da prepozna instance koje se i malo razlikuju od naučenih



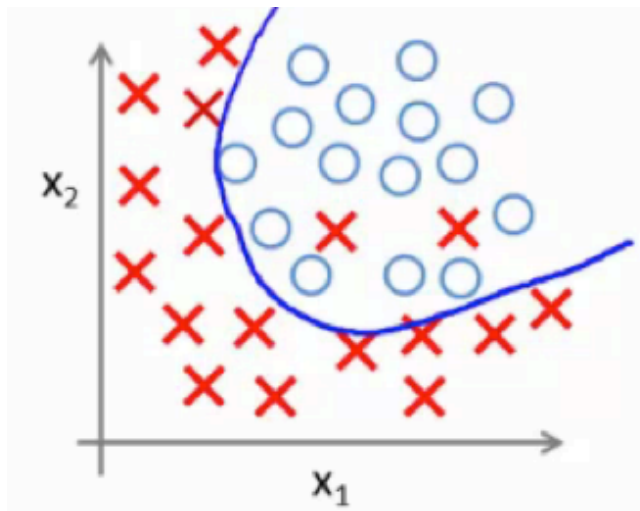
poželjno rešenje



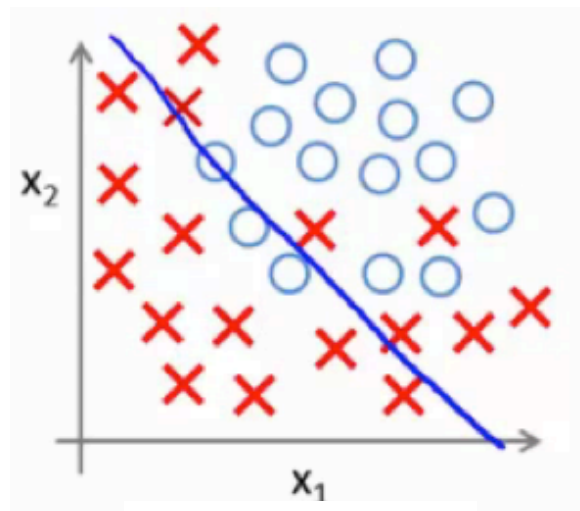
over fitting

PROBLEM NEDOVOLJNOG PODUDARANJA (UNDER-FITTING)

Under-fitting se odnosi na slučaj kad model ne uspeva da aproksimira podatke za trening, tako da ima slabe performanse čak i na trening setu

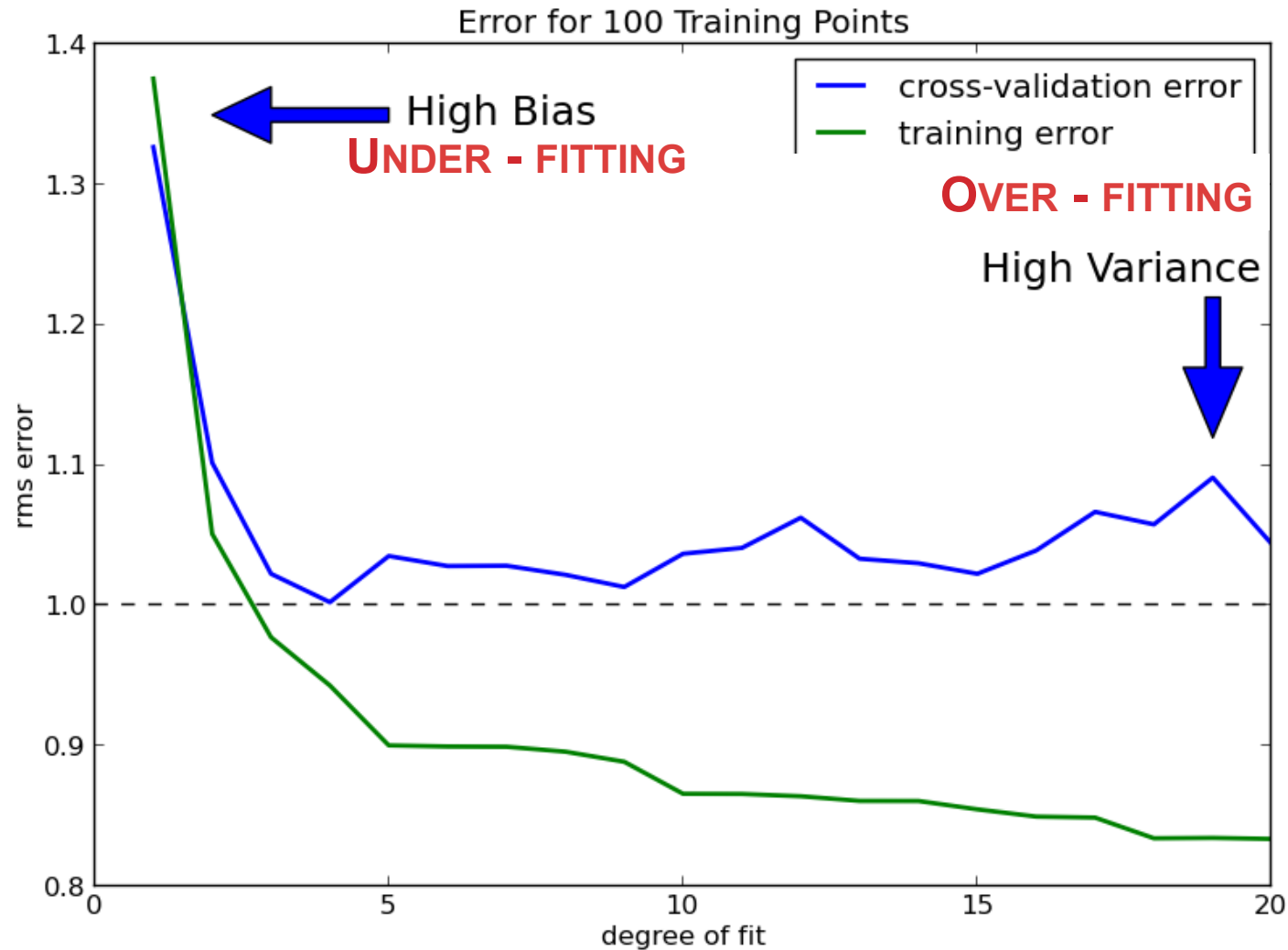


poželjno rešenje



under fitting

OVER-FITTING VS. UNDER-FITTING



ZAHVALNICE I PREPORUKE

Stanford

Machine Learning

Andrew Ng

Learn about the most effective machine learning techniques, and gain practice implementing them and getting them to work for yourself.

Workload: 5-7 hours/week

Taught In: English

Subtitles Available In: English

Preview



Sessions:

Oct 14th 2013 (10 weeks long)

Sign Up

Apr 22nd 2013 (10 weeks long)

Sign Up

3,484

12k

Like 13k

Tweet

+1

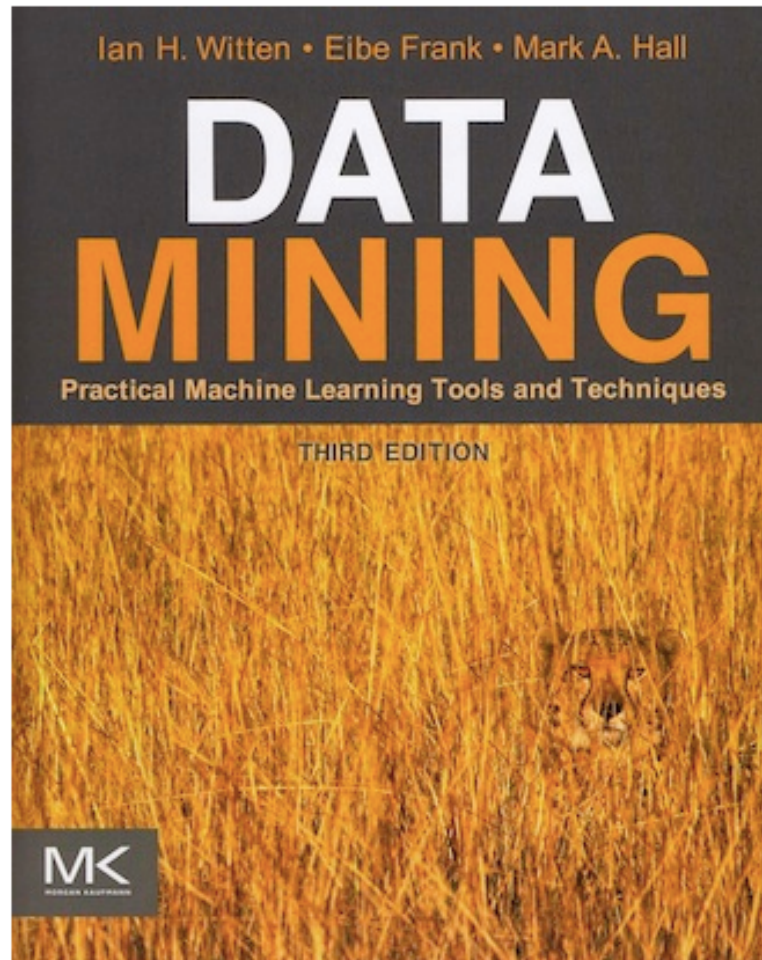
Like

Coursera:

<https://www.coursera.org/course/ml>

Stanford YouTube channel:

http://www.youtube.com/view_play_list?p=A89DCFA6ADACE599



<http://www.cs.waikato.ac.nz/ml/weka/book.html>

Springer Texts in Statistics

Gareth James
Daniela Witten
Trevor Hastie
Robert Tibshirani

An Introduction to Statistical Learning

with Applications in R

 Springer

<http://www-bcf.usc.edu/~gareth/ISL/index.html>

CPSC540

Machine Learning

UBC Nando de Freitas
January, 2013
University of British Columbia

UBC Computer Science
© Copyright

CPSC 540
Nando De Freitas

Jan 08 2013
12:42

03:53 / 59:59

Machine learning - introduction

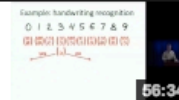


Nando de Freitas · 81 videos

Subscribe 1,514

3,357

16 0



Machine Learning: The Basics, with Ron Bekkerman
by LinkedInTechTalks
16,919 views



Andrew Ng: Deep Learning, Self-Taught Learning and Unsupervised Feature
by

Predavanja Nando de Freitas-a na UBC-u

<http://www.youtube.com/watch?v=w2OtwL5T1ow>

NEKI (POTENCIJALNO) INTERESANTNI PRIMERI PRIMENE MAŠINSKOG UČENJA

- NELL - Never Ending Language Learner ([website](#)) ([NYT article](#)) ([video lecture](#))
- [Relationship mining on Twitter](#)
- [What Facebook Knows](#) (data analysis at Facebook)
- [Using Location Data to Predict the Events You Will Want to Attend](#)
- [Smart Autofill - Harnessing the Predictive Power of Machine Learning in Google Sheets](#)
- [Deep Learning](#) (what it is about)
- Learning Analytics and MOOCs ([TED talk](#)) ([research paper](#) on the use of ML to predict dropouts in MOOCs)

(Anonimni) upitnik za vaše kritike,
komentare, predloge:

<http://goo.gl/cqdp3l>