MAŠINSKO UČENJE

JELENA JOVANOVIĆ

Email: jelena.jovanovic@fon.bg.ac.rs

PREGLED PREDAVANJA

- Šta je mašinsko učenje?
- Zašto / kad je potrebno / korisno m. učenje?
- Oblasti primene m. učenja
- Oblici m. učenja
- Osnovni koraci i elementi procesa m. učenja

ŠTA JE MAŠINSKO UČENJE ?

Oblast Veštačke inteligencije koja se bavi razvojem algoritama i metoda koje programima omogućuju da izvršavaju različite zadatke tako što

- uče iz primera, bez eksplicitnog programiranja
- kreiraju interne reprezentacije naučenog
- koriste te reprezentacije kako bi adekvatno odreagovali (tj. izvršili zadatak) u novoj situaciji

ŠTA JE MAŠINSKO UČENJE ?

Formalnije, za definisanje m. učenja potrebna su nam 3 koncepta:

- **T**ask: zadatak, tj. predmet i svrha učenja
- **P**erformance: mere uspešnosti učenja
- **E**xperience: iskustvo tj. podaci o pojavama / entitetima koji su predmet učenja

ŠTA JE MAŠINSKO UČENJE ?

Za kompjuterski program kažemo da uči ukoliko se njegove *performanse* na datom *zadatku* (***T***), merene izabranim *metrikama* (***P***), unapređuju sa novim *iskustvom*, tj podacima (***E***) *relevantnim* za dati zadatak

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***): skup email poruka označenih kao spam i not-spam; atributi koji opisuju te poruke
- Mere performansi (***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,…

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

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, …, 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

NADGLEDANO UČENJE

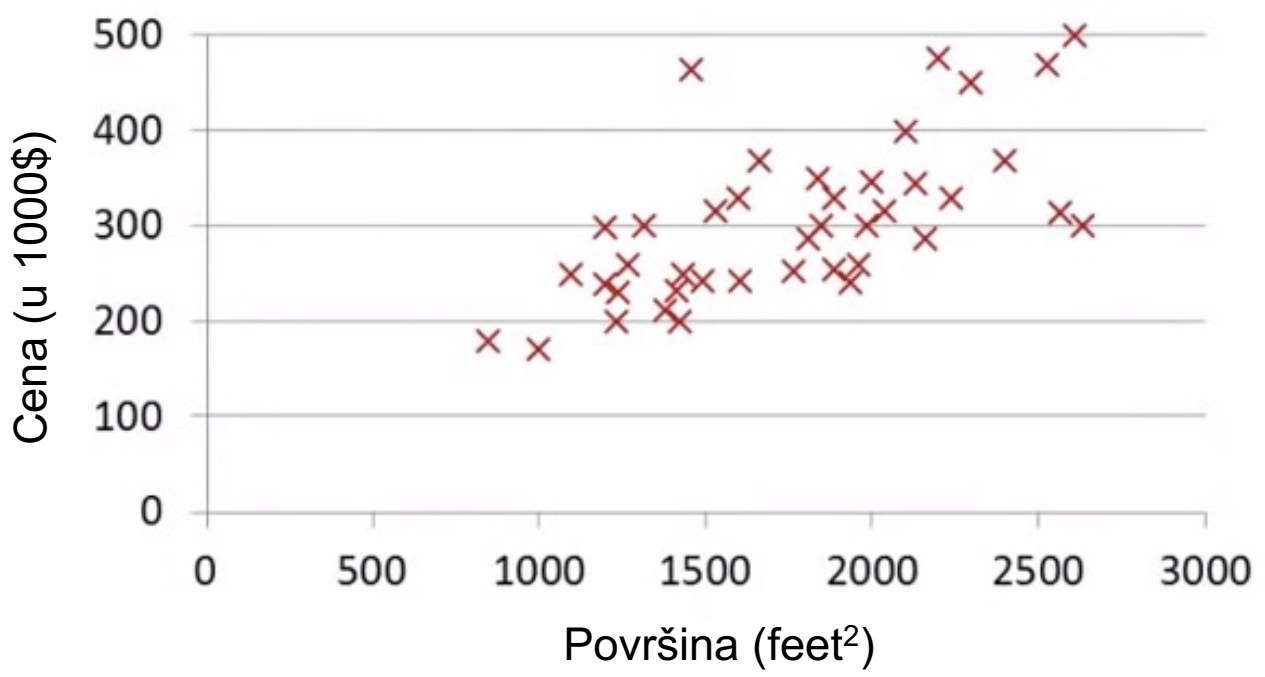
Izlazna vrednost može biti:

- 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



Izvor slike: https://www.coursera.org/learn/machine-learning

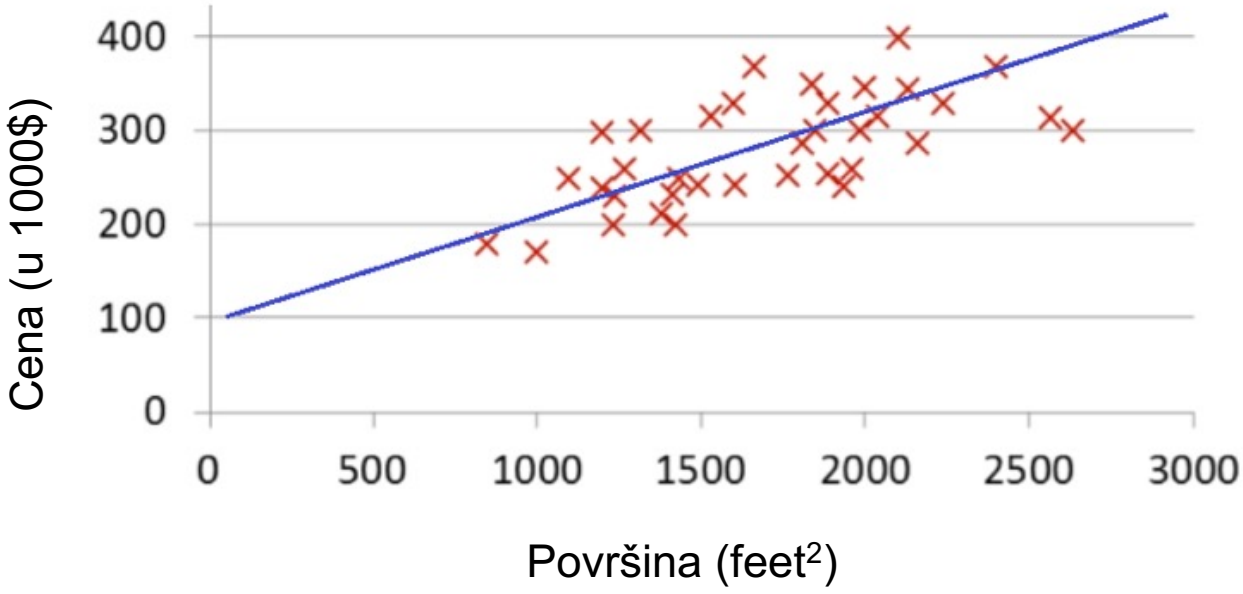
NADGLEDANO UČENJE

Primer linearne regresije (nastavak)

Funkcija koju treba „naučiti“ u ovom slučaju je:

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

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



NENADGLEDANO UČENJE

Kod nenadgledanog učenja

- nemamo informacije o željenoj izlaznoj vrednosti
- program dobija samo skup ulaznih podataka (x_1, x_2, …, x_n)

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

NENADGLEDANO UČENJE

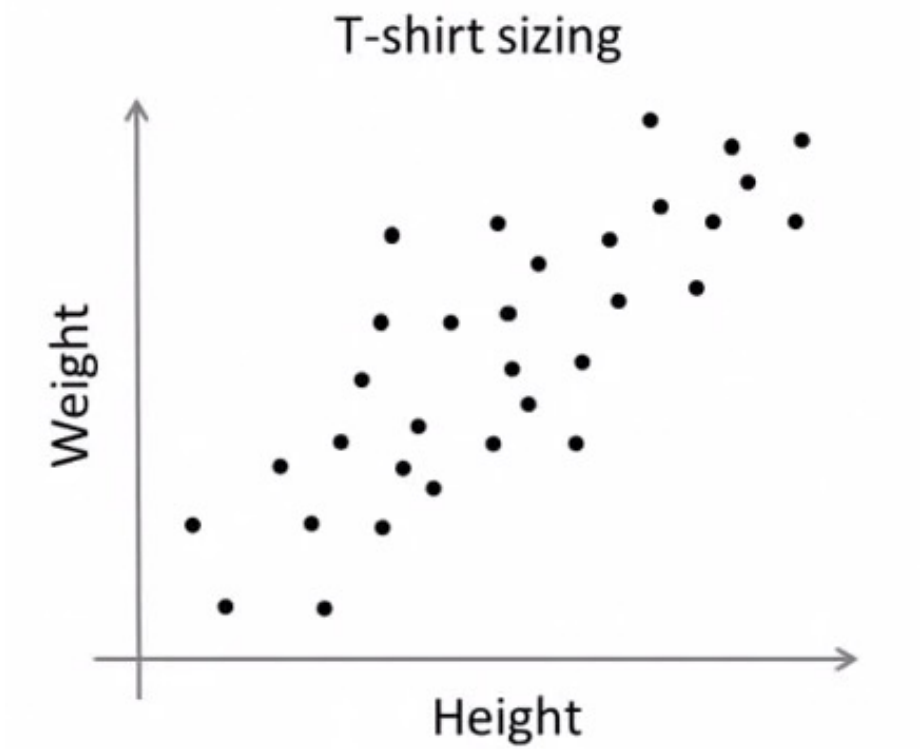
Tipični zadaci:

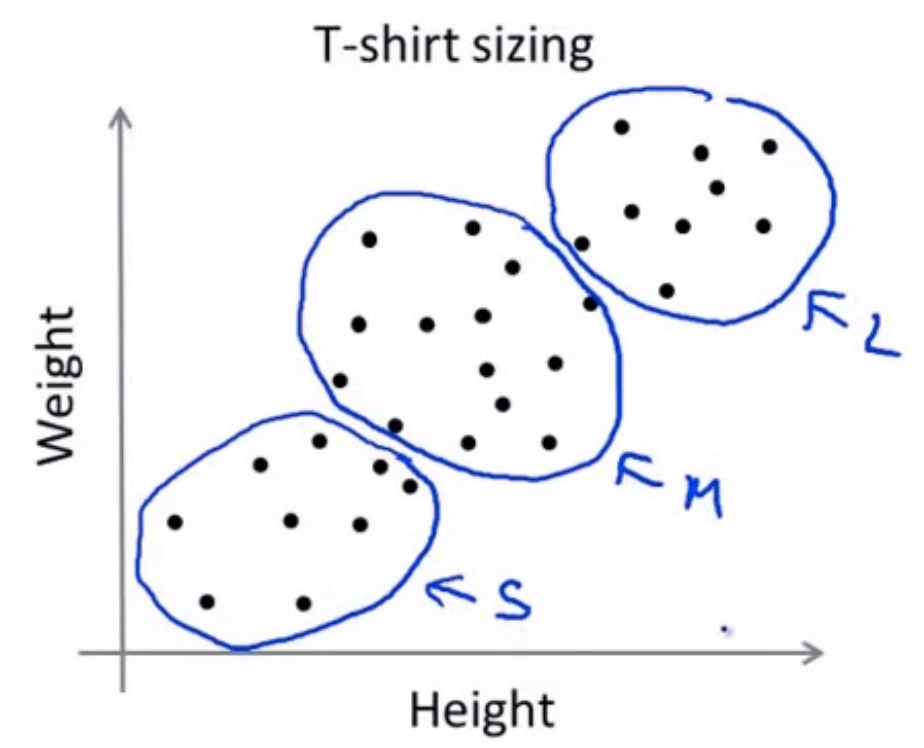
- Klasterizacija
- Redukcija dimenzionalnosti podataka
- Identifikovanje asocijativnih pravila

NENADGLEDANO UČENJE

Primer klasterizacije:

određivanje konfekcijskih veličina na osnovu visine i težine ljudi



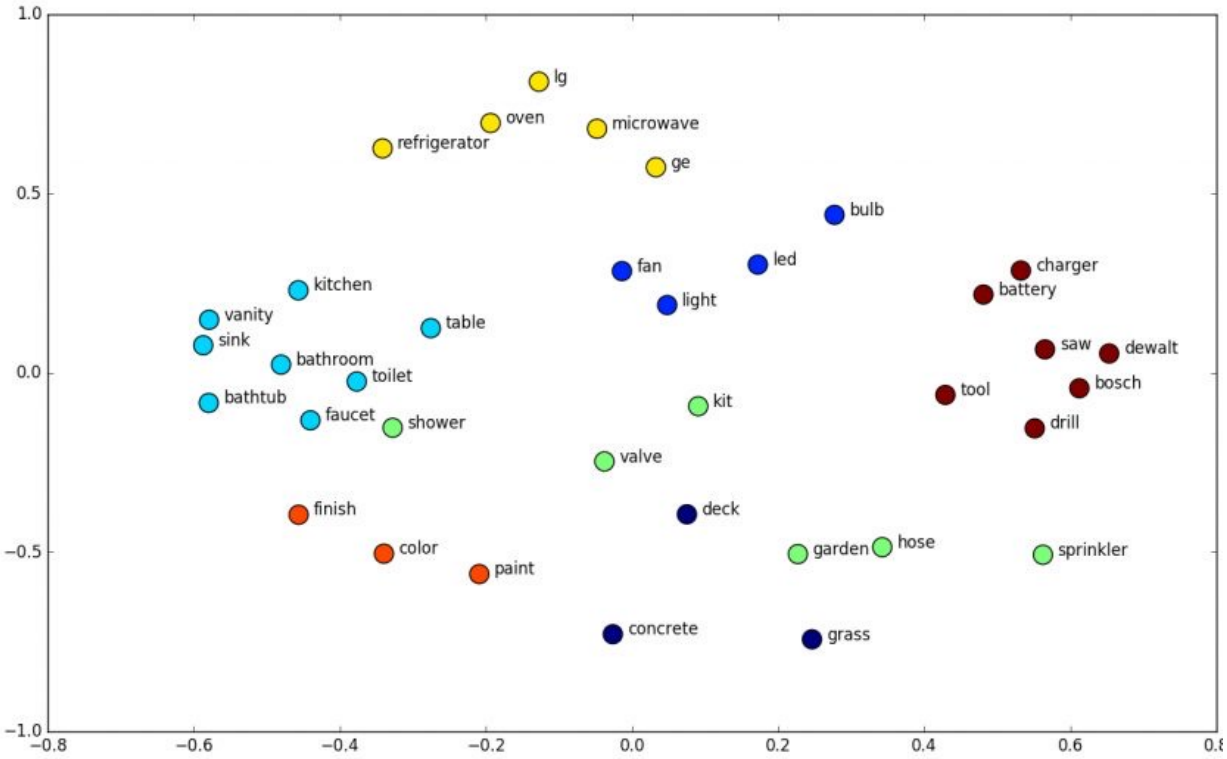


Izvor: https://www.coursera.org/learn/machine-learning

NENADGLEDANO UČENJE

Primer redukcije dimenzionalnosti:

visoko-dimenzionalna reprezentacija reči (word embeddings) svedena na 2 dimenzije, primenom t-SNE metode



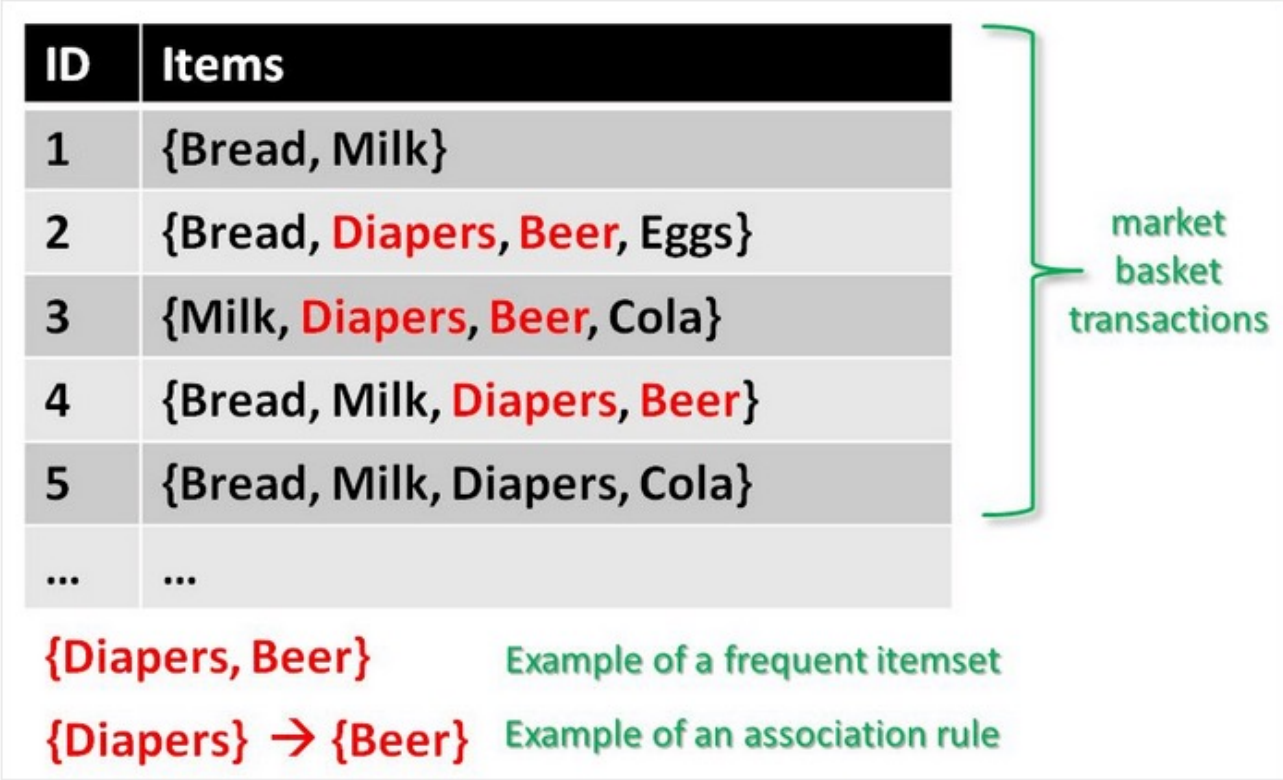
Izvor slike:
https://www.shanelynn.ie/get-busy-with-word-embeddings-introduction/

T-SNE:
https://lvdmaaten.github.io/tsne/

NENADGLEDANO UČENJE

Primer identifikovanja asocijatnih pravila:

market basket analysis



Izvor slike: https://www.datacamp.com/tutorial/market-basket-analysis-r

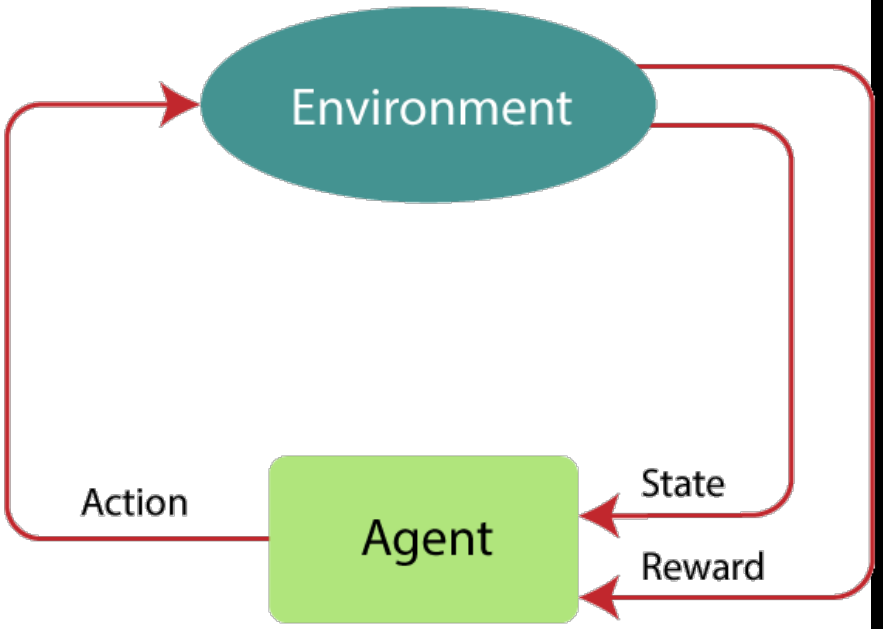
UČENJE UZ PODSTICAJE

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

Svaka akcija utiče na *okruženje* i menja *stanje* agenta

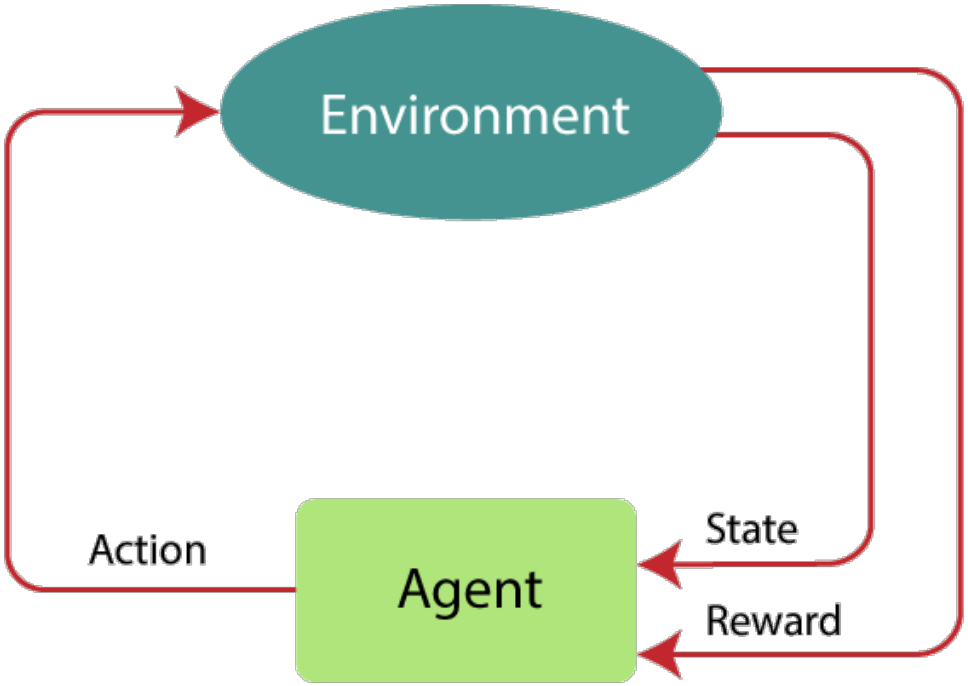
Okruženje povratno utiče na agenta pružajući mu *povratne informacije* koje mogu biti pozitivne (*nagrade*) ili negativne (*kazne*)

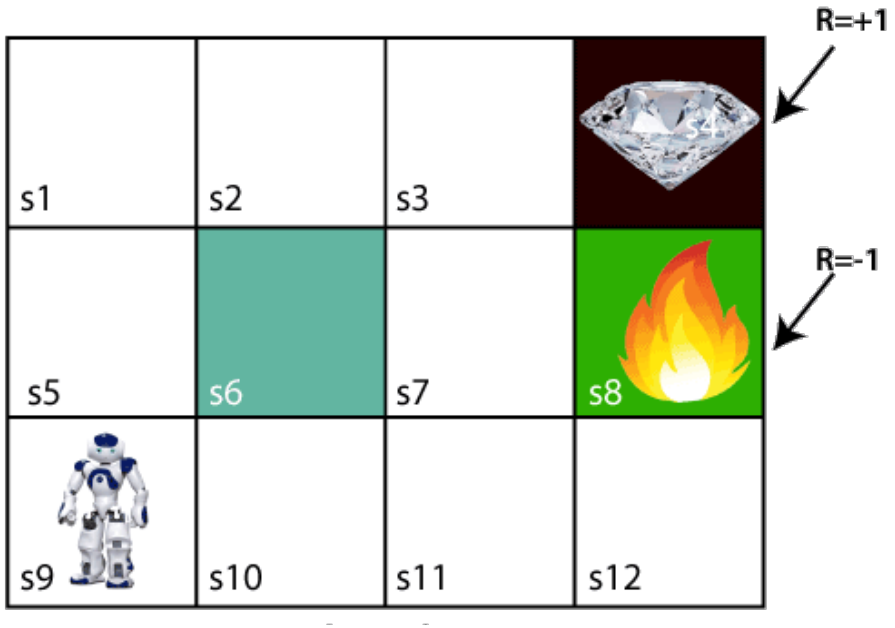
Cilj agenta je da (kroz pokušaje i greške) nauči *sekvence akcija* koje će u datom okruženju max. nagrade (ili min. kazne)



Izvor: https://www.javatpoint.com/reinforcement-learning

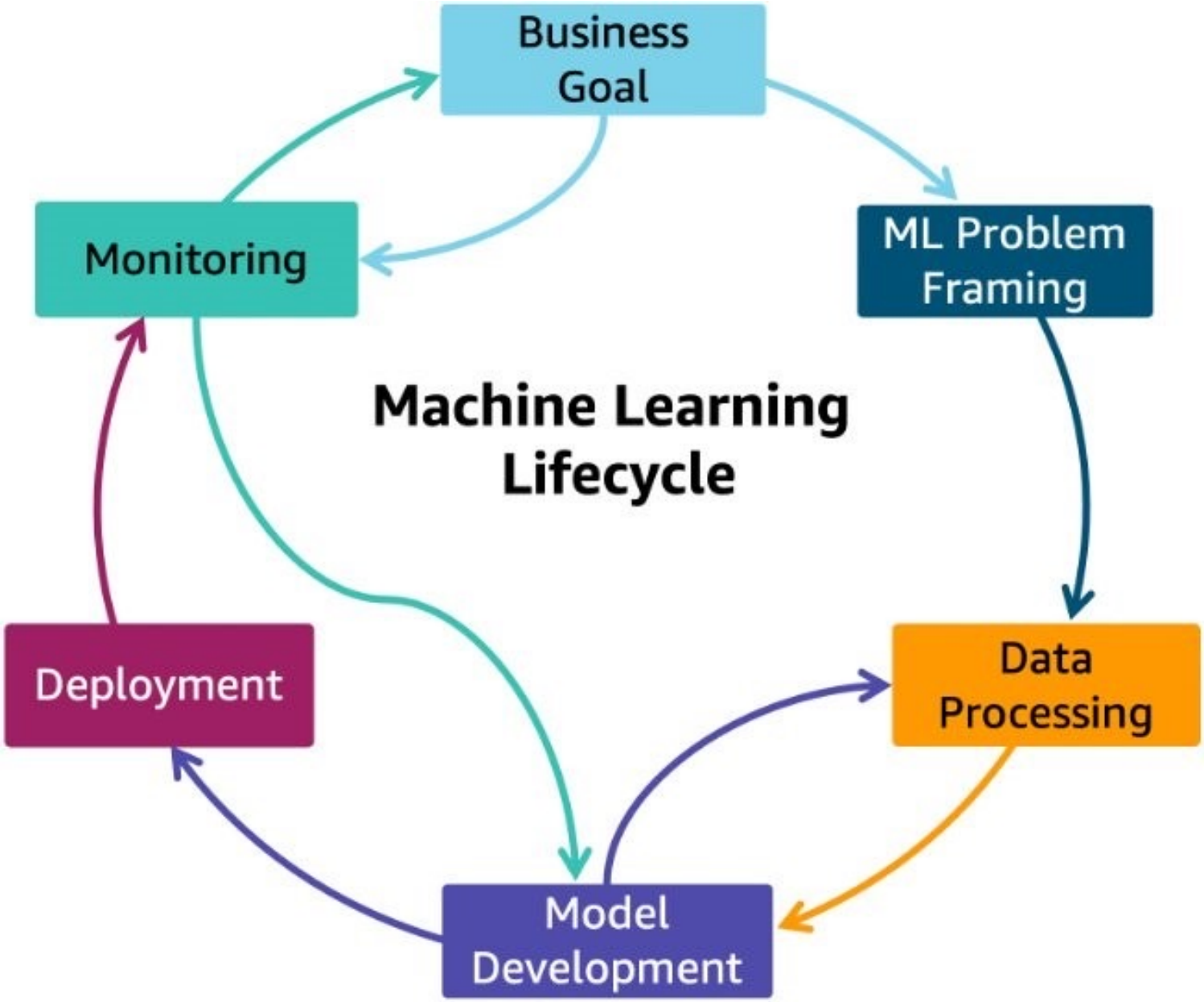
UČENJE UZ PODSTICAJE



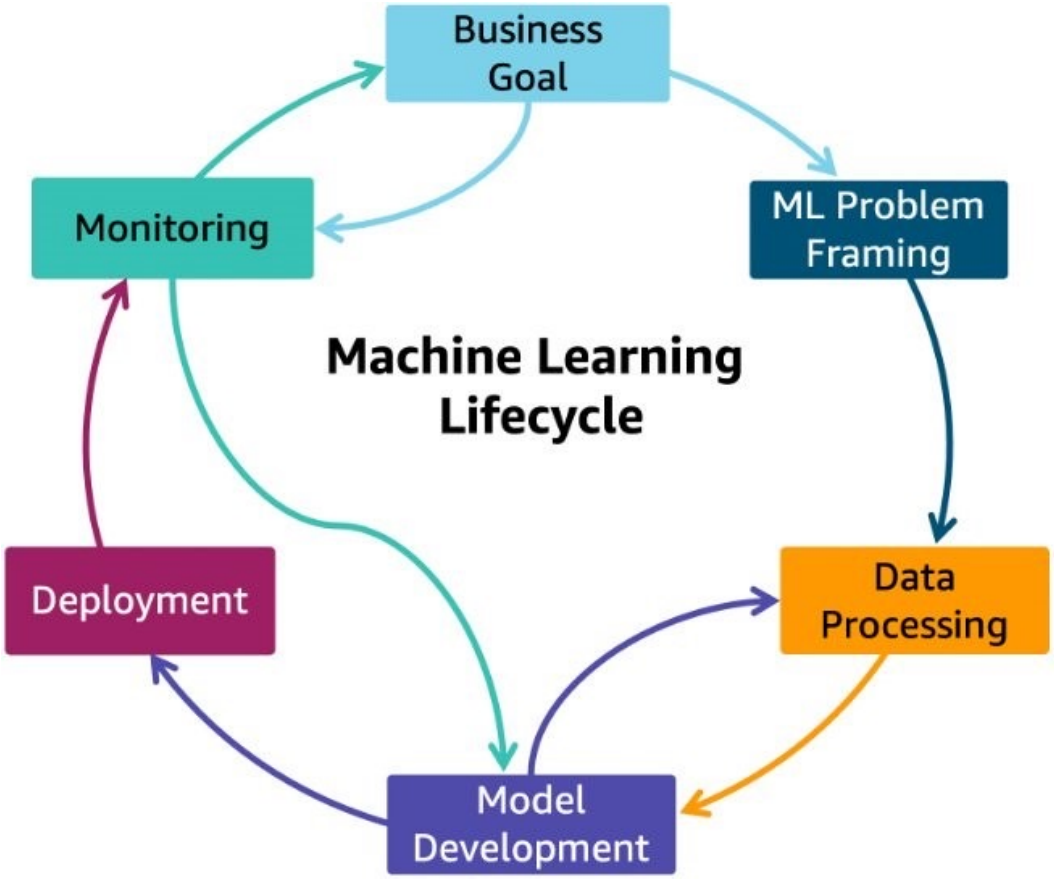


Izvor: https://www.javatpoint.com/reinforcement-learning

OSNOVNI KORACI I ELEMENTI PROCESA M. UČENJA



Izvor slike: https://aws.amazon.com/blogs/architecture/optimize-ai-ml-workloads-for-sustainability-part-2-model-development/



Definisanje zadatka m. učenja i prikupljanje potrebnih podataka

Priprema i transformacija podataka, kreiranje i selekcija atributa

Selekcija algoritama m. učenja
Kreiranje (obuka) modela
Evaluacija modela

Izvor slike: https://aws.amazon.com/blogs/architecture/optimize-ai-ml-workloads-for-sustainability-part-2-model-development/

PODACI

"*The value of data in machine learning cannot be overstated*"

Jedan od primarnih načina za unapređenje performansi modela m. učenja je obezbediti *što više kvalitetnih* podataka

PODACI

Preporuka: predavanja o značaju podataka za m. učenje

The Unreasonable Effectiveness of Data

Peter Norvig

URL: https://www.youtube.com/watch?v=yvDCzhbjYWs

PODACI

Preporuka: predavanja o značaju podataka za m. učenje

The Unreasonable Effectiveness of Data

Peter Norvig

URL: https://www.youtube.com/watch?v=yvDCzhbjYWs

Data centric AI development: From Big Data to Good Data

Andrew Ng

URL: https://www.youtube.com/watch?v=avoijDORAIc

PODACI

Izvori podataka:

- Javno dostupne kolekcije podataka, npr.
  - https://github.com/awesomedata/awesome-public-datasets
  - https://www.kaggle.com/datasets
- Podaci dostupni posredstvom Web API-a
- Sve veće tržište gde je moguće kupiti podatke
  - Pogledati npr. https://about.datarade.ai/data-marketplaces

PODACI

Za nadgledano učenje, moramo imati “obeležene” podatke

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

PODACI

Za nadgledano učenje, moramo imati “obeležene” podatke

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

Ilustracija kako to izgleda u dataset-u:

	Hits	Years	WellPaid
1	81	14	No
2	130	3	No
3	141	11	No
4	87	2	No
5	169	11	Yes
6	37	2	No
7	73	3	No
8	81	2	No
9	92	13	Yes
10	159	10	No
11	53	9	No
12	113	4	No

PODACI

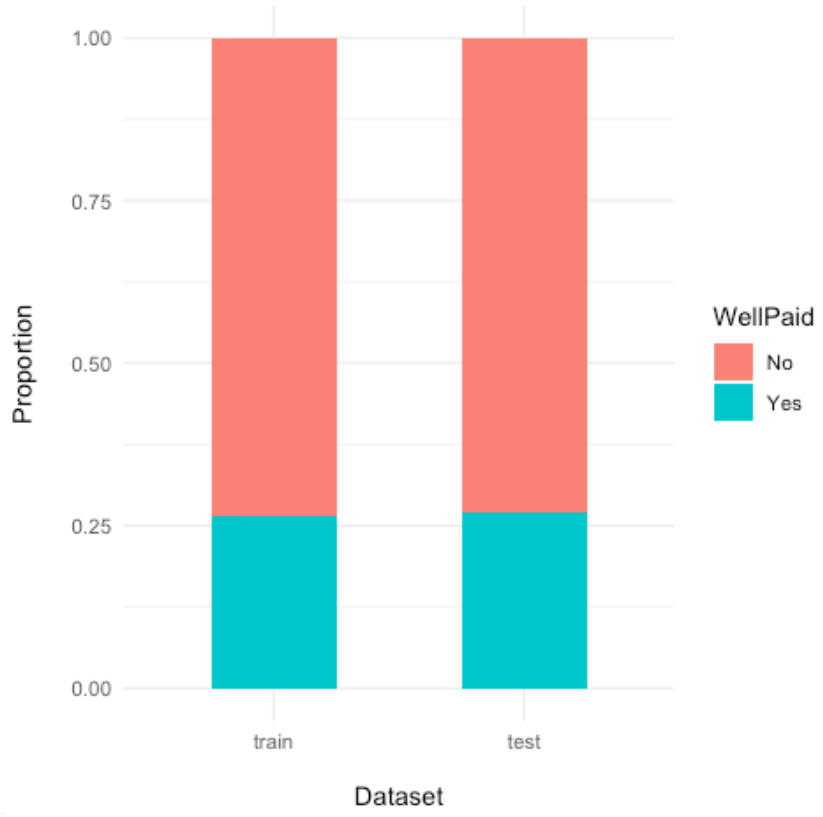
Za nadgledano učenje, raspoloživi skup podataka se deli na:

- Podatke za kreiranje (trening) modela, obično 70-80% podataka
- Podatke za testiranje modela, preostalih 20-30% podataka

PODACI

Pri izboru uzoraka za trening i testiranje, potrebno je

- selekciju uraditi na slučajan način (random selection)
- obezbediti jednaku distribuciju
  izlazne varijable u trening i test setu



ATRIBUTI (FEATURES)

Osnovna ideja:

- Pojave / entitete prepoznajemo uočavajući njihove osobine (ili izostanak nekih osobina) i uviđajući odnose između različitih osobina
- Omogućiti programu da koristi osobine pojava / entiteta za potrebe identifikacije / grupisanja / predviđanja

ATRIBUTI (FEATURES)

Primeri:

- Zadatak predviđanja cene stana:
  - površina, lokacija, broj soba, tip grejanja i sl.

ATRIBUTI (FEATURES)

Primeri:

- Zadatak predviđanja cene stana:
  - površina, lokacija, broj soba, tip grejanja i sl.
- Zadatak identifikovanja spam poruka:
  - domen pošiljaoca poruke, broj specijalnih znakova u tekstu poruke, dužina poruke, i sl.

ATRIBUTI (FEATURES)

Primeri:

- Zadatak predviđanja cene stana:
  - površina, lokacija, broj soba, tip grejanja i sl.
- Zadatak identifikovanja spam poruka:
  - domen pošiljaoca poruke, broj specijalnih znakova u tekstu poruke, dužina poruke, i sl.
- Zadatak profilisanja klijenata kompanije (klasterizacija):
  - učestanost i obim kupovine proizvoda različite kategorije, raspoloživi demografski podaci i sl

ATRIBUTI (FEATURES)

Izazov:

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

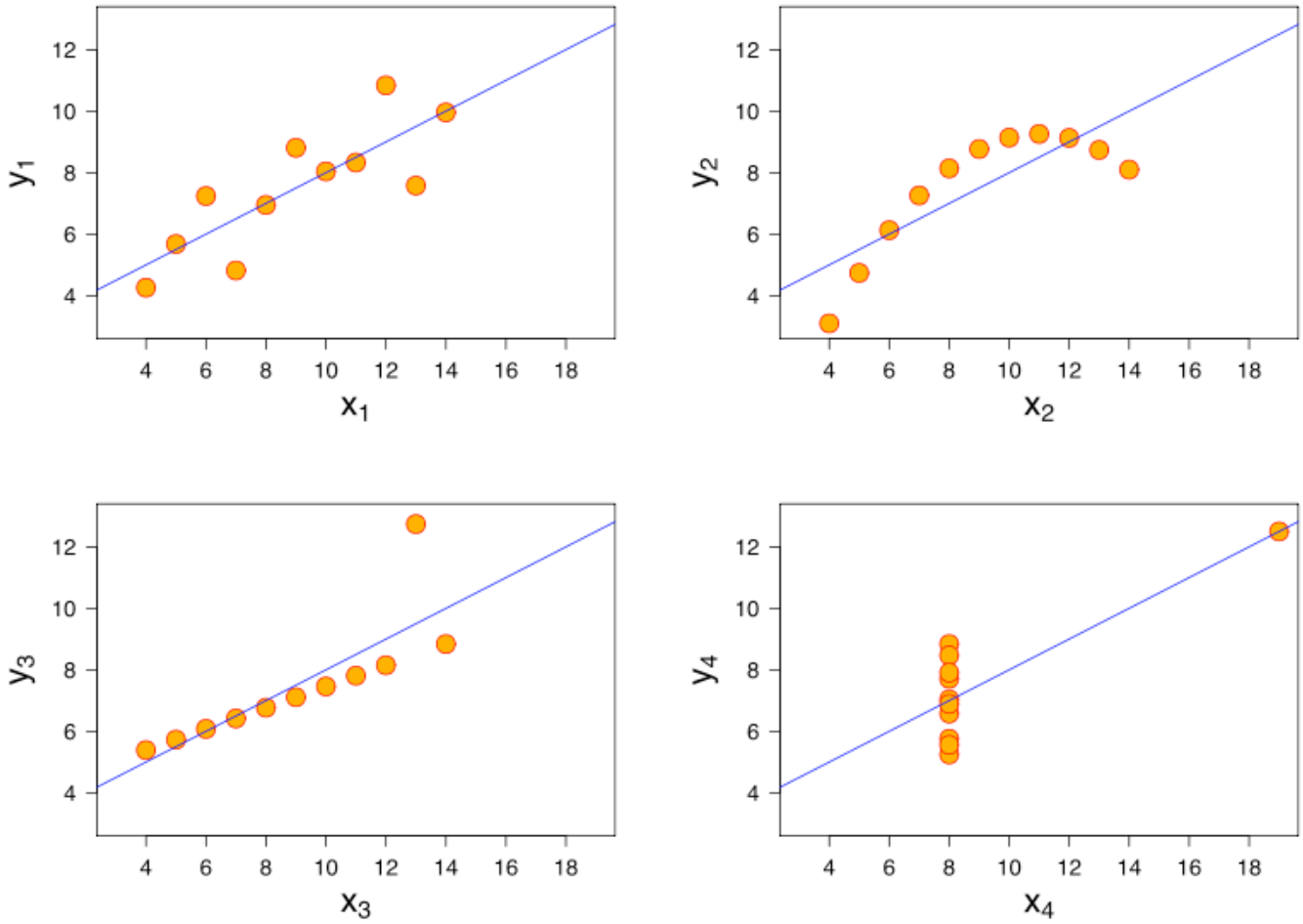
ODABIR METODE M. UČENJA

Generalno, zavisi od:

- Vrste problema koji rešavamo
- Obima podataka koji su nam na raspolaganju
- Karakteristika skupa atributa (features)
  - tip atributa
  - stepen međuzavisnosti (korelisanosti) atributa
  - distribucija vrednosti atributa
- Značaja mogućnosti objašnjenja rezultata

ODABIR METODE M. UČENJA

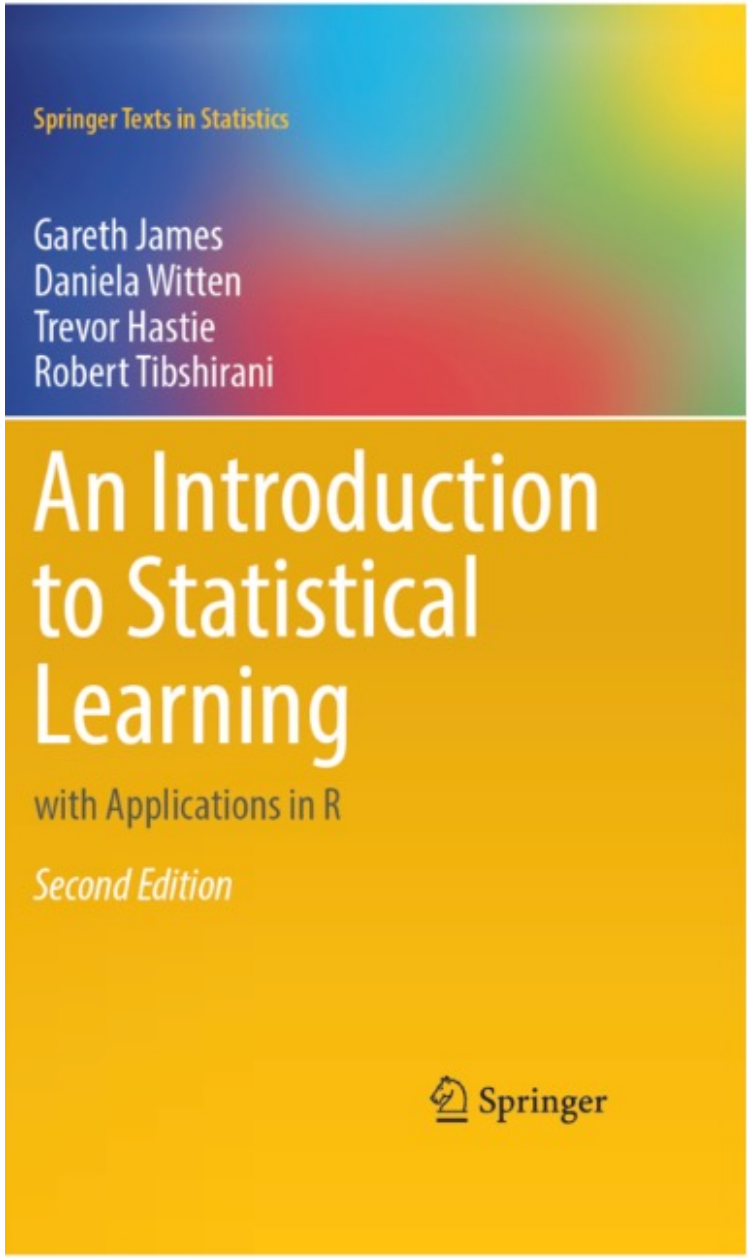
Primer: pokušaj aproksimacije četiri različita skupa podataka primenom iste linearne funkcije (tj. linearne regresije)



Očigledno razičiti podaci traže različite funkcije tj algoritme učenja

Izvor: http://goo.gl/yjm0El

Preporuka dodatne (neobavezne) literature



Besplatno (legalno) dostupna na: https://www.statlearning.com/

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

https://forms.gle/7vdwezRAumx72VnXA