Obrada prirodnih jezika

Elektrotehnički fakultet Univerziteta u Beogradu

Master akademske studije

modul Softversko inženjerstvo

2021/2022

Mašinsko učenje

Vuk Batanović, Elektrotehnički fakultet Univerziteta u Beogradu

Uvod u mašinsko učenje

- Šta je mašinsko učenje?
- Arthur Samuel (1959): Field of study that gives computers the ability to learn without being explicitly programmed
- Kombinacija primenjene statistike, veštačke inteligencije, matematičke optimizacije, računarskih nauka,…
- Čemu mašinsko učenje?
  - Problemi za koje je vrlo teško ručno definisati kako ih treba rešiti – klasično programiranje nije moguće
  - Izvlačenje korisnih informacija iz velike količine sirovih podataka ili predviđanje budućih trendova korišćenjem trenutno dostupnih podataka – istraživanje podataka (engl. *data mining* / *predictive analytics*) – ručna analiza nije moguća ili je previše spora
  - Kompleksni sistemi koji se dinamički prilagođavaju okruženju

Uvod u mašinsko učenje

- Neki problemi za koje je teško ručno definisati kako ih treba rešiti
  - Obrada prirodnih jezika (engleskog, francuskog, srpskog,…)
    - Detekcija *spam*-a, analiza sentimenta teksta, mašinsko prevođenje,…
  - Računarski vid
    - *Optical character recognition* (OCR), prepoznavanje lica ili pokreta,…
  - Robotika
    - Kretanje robota kroz prostor
  - Prepoznavanje obrazaca
    - Pomoć pri postavljanju dijagnoza u medicini (engl. *Computer Aided Diagnosis* – CAD)
  - Igranje igara
    - Go – DeepMind AlphaGo

Tipovi mašinskog učenja

- Nadgledano učenje (engl. *Supervised learning*)
- Nenadgledano učenje (engl. *Unsupervised learning*)
- Samonadgledano učenje (engl. *Self-supervised learning*)
- Polunadgledano učenje (engl. *Semi-supervised learning*)
- Učenje sa podrškom (engl. *Reinforcement learning*)

Nadgledano učenje

- Najčešće primenjivan tip mašinskog učenja
- Svakom ulaznom podatku x je pridružena željena izlazna vrednost y koju algoritam treba da predvidi
- Cilj učenja je da se na osnovu datih parova (x, y) pronađe optimalna funkcija koja mapira ulaz u izlaz
- Realna funkcija preslikavanja je nepoznata, tako da se funkcija koja se optimizuje često naziva *hipotezom* i označava sa $h(x)$
- Model nadgledanog učenja koji se primenjuje određuje opšti oblik te funkcije (npr. da se radi o linearnoj funkciji) tj. prostor mogućih hipoteza
- Konkretne vrednosti podataka koje se koriste pri obučavanju određuju tačan oblik te funkcije tj. najbolju hipotezu u datom prostoru

Odlike / atributi

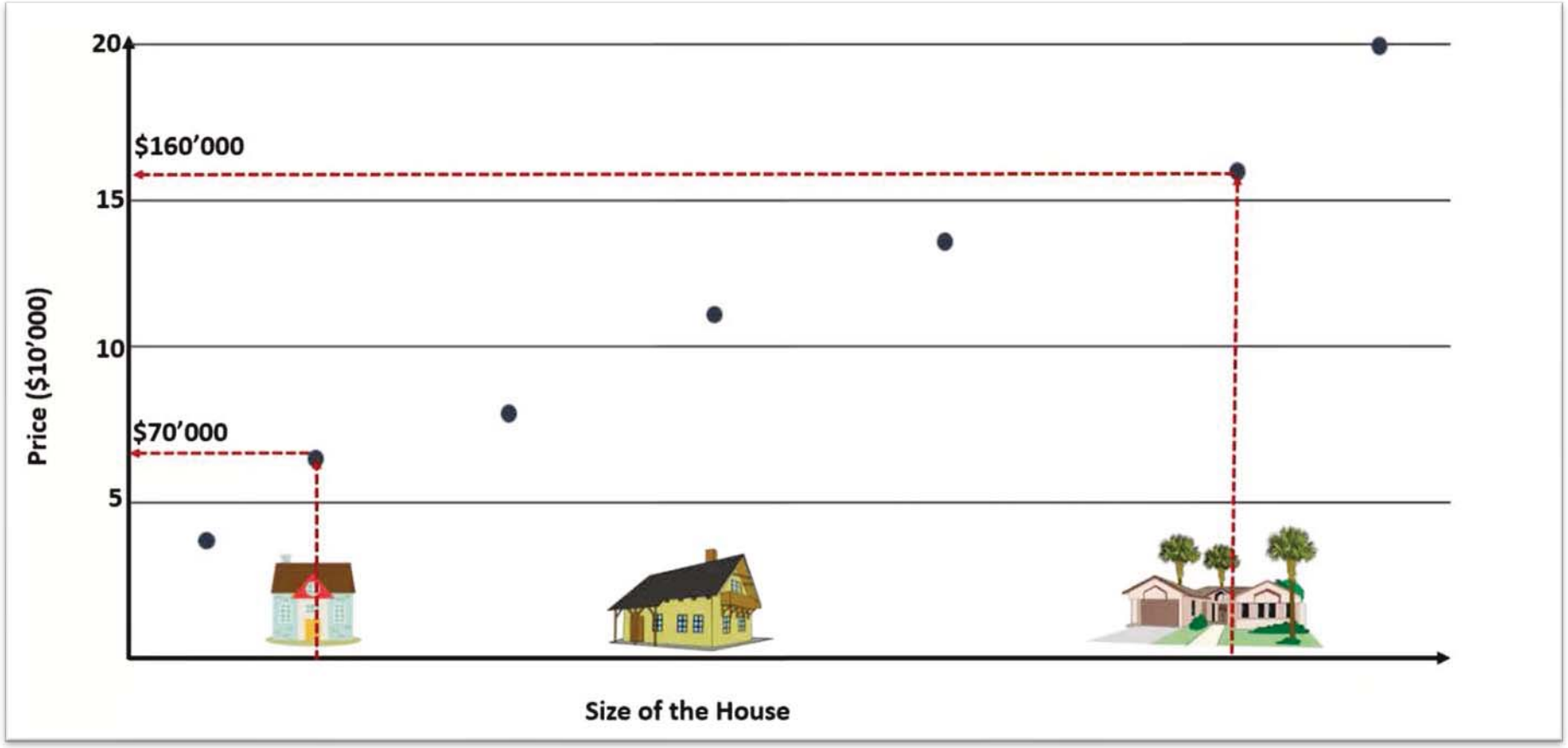
- U klasičnom mašinskom učenju potrebno je ručno specificirati koji su to faktori u ulaznim podacima koji utiču na izlaz – te faktore nazivamo *atributima* ili *odlikama* (engl. *features*)
- Pri učenju se svaki podatak tretira kao skup/vektor nekih njegovih odlika:

$$x = (x_1, x_2, \ldots, x_n)$$

- U dubokom mašinskom učenju (engl. *deep learning*) model je u stanju da sam pronađe relevantne odlike
  - Ali je zato potrebno pronaći odgovarajuću strukturu modela za svaki posmatran problem

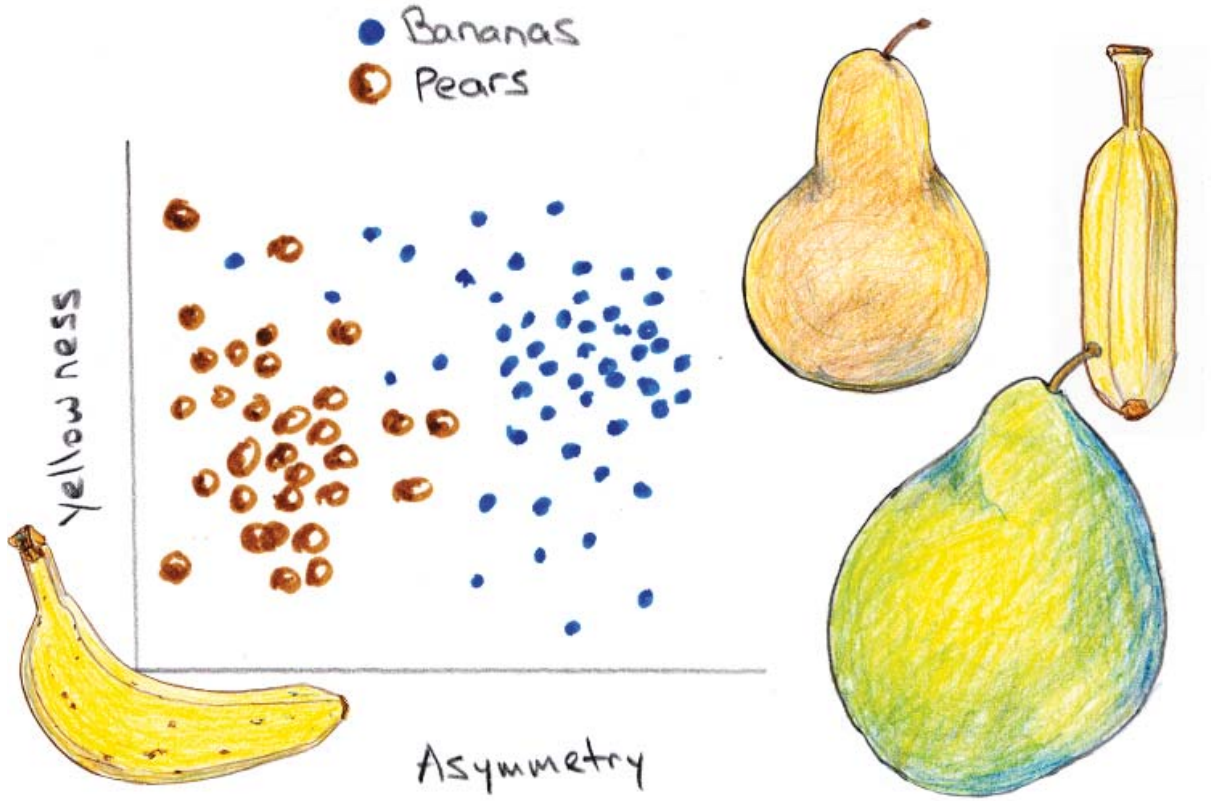
Nadgledano učenje

- Cilj učenja jeste da se dobije takav model koji će davati dobre rezultate ne samo nad podacima koji su korišćeni pri obučavanju modela već i na nekim kasnijim/drugim, dotle neviđenim podacima
  - Ova sposobnost modela se naziva sposobnost *generalizacije*
- Tip vrednosti y
  - Kontinualna vrednost – problem *regresije*
  - Diskretna / kategorička vrednost – problem *klasifikacije*
    - Ako postoje samo dve klase – binarna klasifikacija
    - Ako postoji više klasa – višeklasna klasifikacija
  - Strukturirana vrednost - problem *strukturne predikcije*
- Veliki broj zadataka u NLP-u se svodi na neki vid klasifikacije ili strukturne predikcije



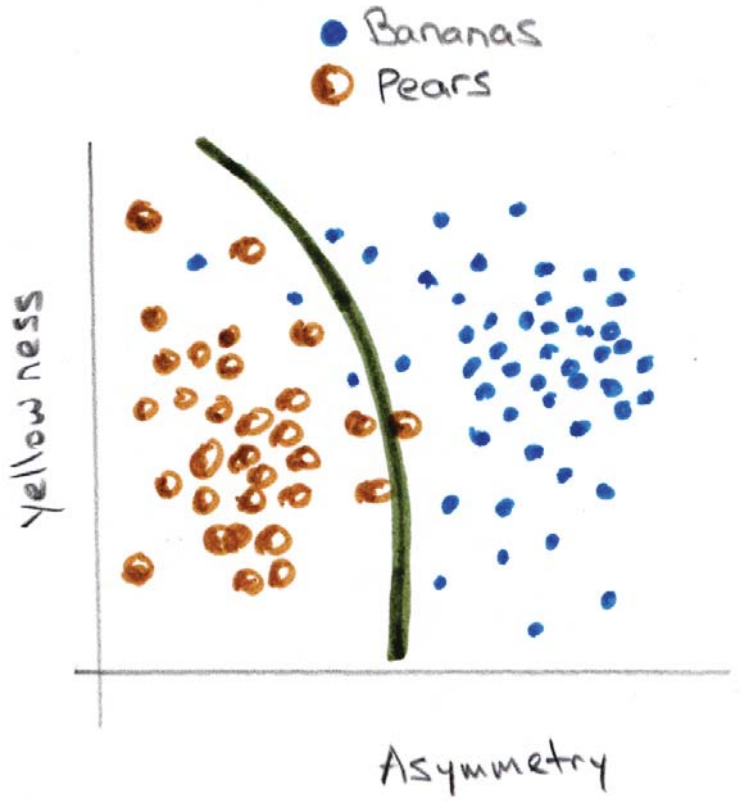
Primer problema regresije – predviđanje cene nekretnina

Modifikovana slika preuzeta sa: http://www.youtube.com/watch?v=dnKETo-gWbc



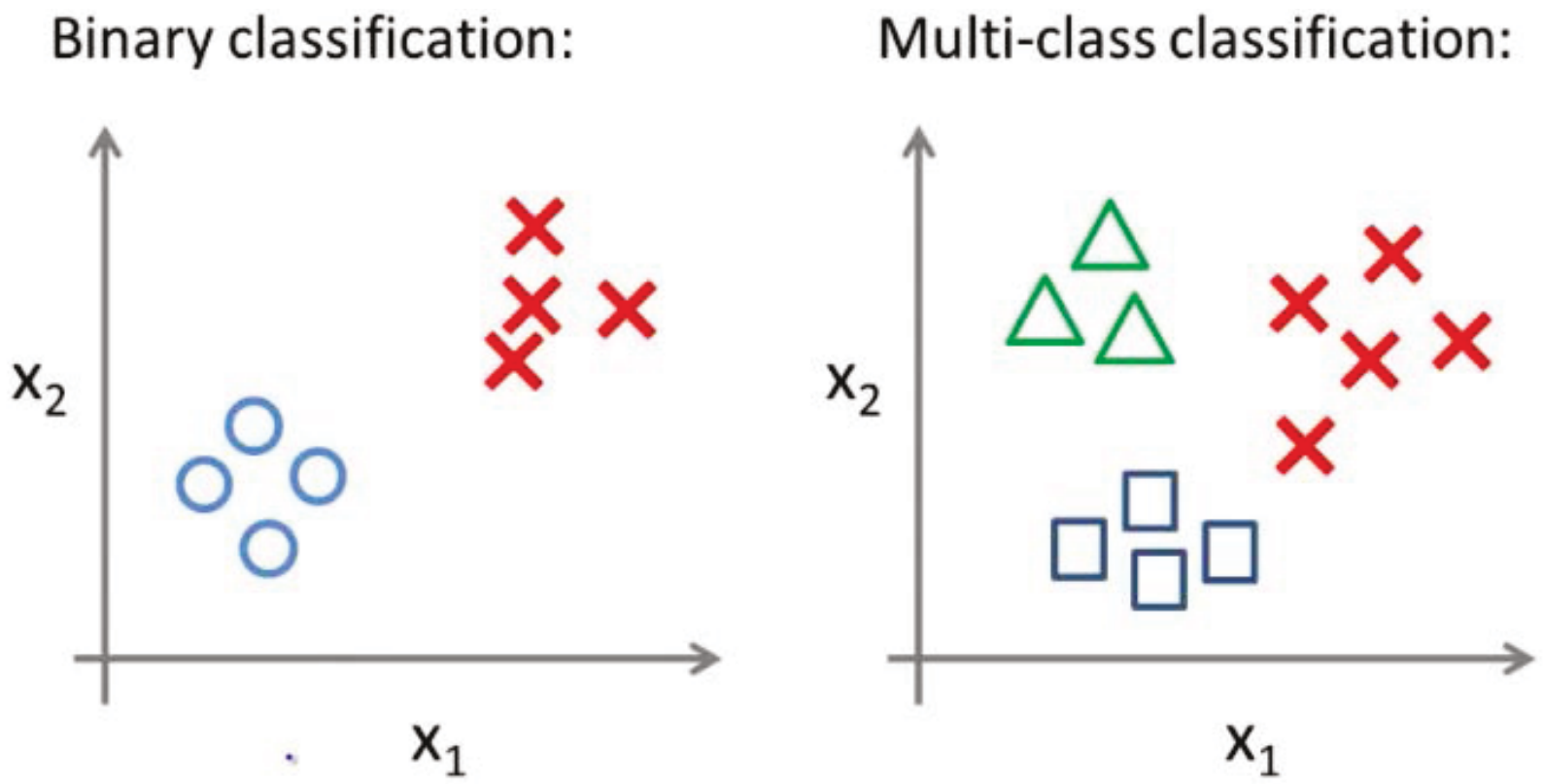
Primer problema klasifikacije – razdvajanje objekata na osnovu njihovih slika

Slika preuzeta sa: http://eighteenthelephant.wordpress.com/2015/10/23/learning-about-machine-learning-part-i/



Primer problema klasifikacije – razdvajanje objekata na osnovu njihovih slika

Slika preuzeta sa: http://eighteenthelephant.wordpress.com/2015/10/23/learning-about-machine-learning-part-i/

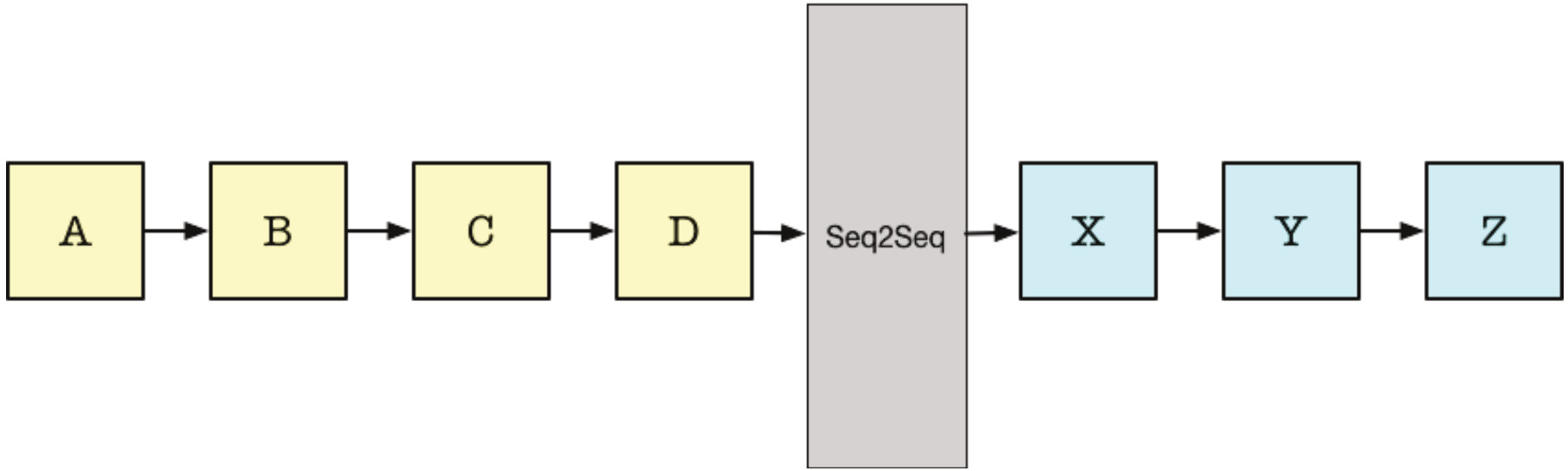


Ilustracija razlike između binarne i višeklasne klasifikacije

Slika preuzeta sa: Andrew Ng, Machine Learning, Coursera

Strukturna predikcija

- Strukturirani podatak
  - Sastoji se iz nekoliko delova
  - Nisu samo delovi ti koji sadrže korisne informacije već je važan i odnos između njih u okviru posmatrane strukture
- Primeri strukturiranih podataka
  - Sekvence
  - Stabla
  - Slike
  - Tekstualni dokumenti
  - …



Primer strukturne predikcije – mašinsko prevođenje

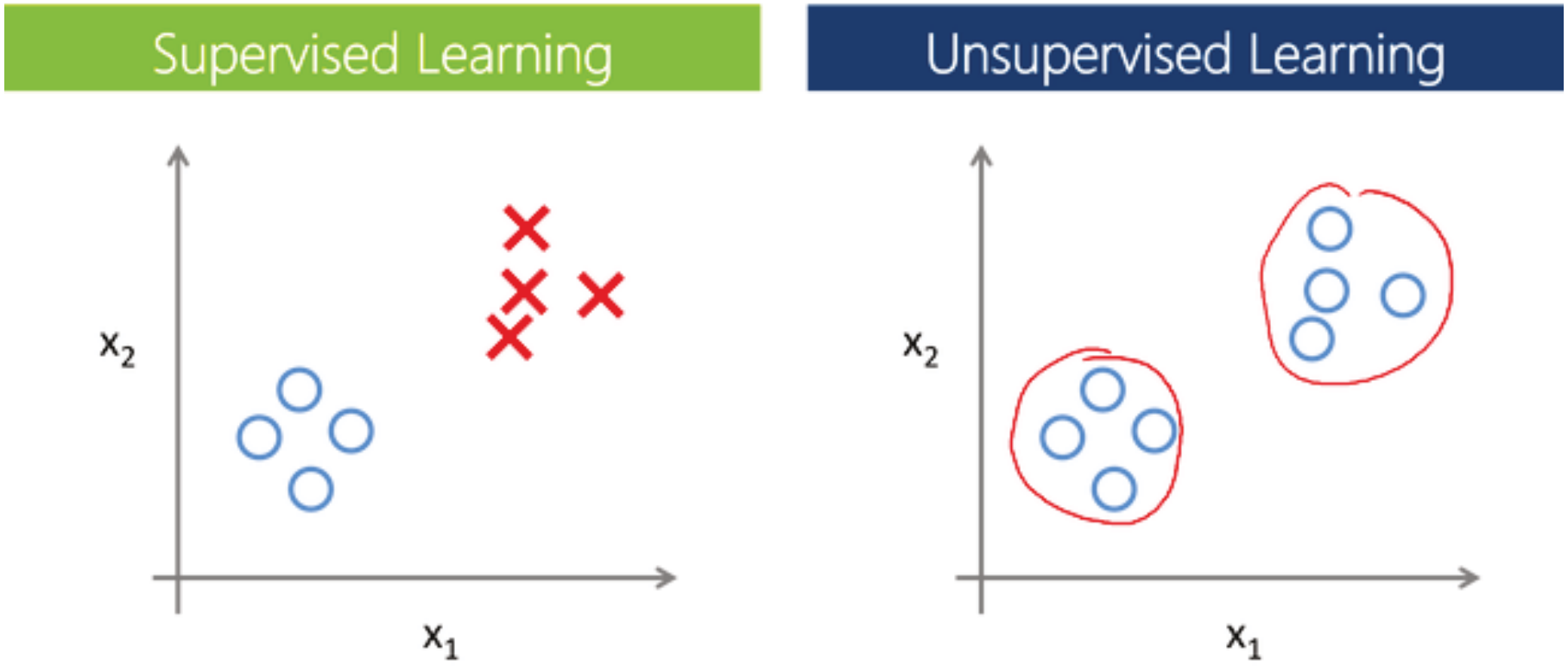
Slika preuzeta sa: http://asivokon.github.io

Nenadgledano učenje

- Dati su samo ulazni podaci x
- Ne postoji željena izlazna vrednost
- Potrebno je pronaći neku pravilnost u podacima
- Tipični zadaci nenadgledanog učenja
  - Grupisanje (engl. *clustering*) – podaci se svrstavaju u grupe koje maksimizuju neki kriterijum sličnosti ili minimizuju neki kriterijum različitosti
  - Smanjenje dimenzionalnosti (engl. *dimensionality reduction*) – pronalaženje manjeg skupa promenljivih koje zadržavaju glavne obrasce i varijacije u skupu početnih promenljivih

Klasifikacija vs grupisanje

Klasifikacija	Grupisanje
Broj klasa poznat unapred	Broj klasa nije poznat unapred
Postojeći podaci imaju označenu pripadnost klasi	Postojeći podaci nemaju označenu pripadnost klasi
Model se koristi za klasifikovanje novih podataka	Model se koristi za razumevanje/istraživanje postojećih podataka
Spada u probleme nadgledanog mašinskog učenja	Spada u probleme nenadgledanog mašinskog učenja



Ilustracija razlike između klasifikacije i grupisanja

Slika preuzeta sa: Andrew Ng, Machine Learning, Coursera

Samonadgledano učenje

- Po nekim izvorima smatra se tipom nenadgledanog učenja, a po drugima prelazom između nenadgledanog i nadgledanog
- I u ovom tipu učenja dati su samo ulazni podaci x
- Osnovna ideja je da željena izlazna vrednost nije data, već se ona izvodi iz ulaznih podataka, često na osnovu njihove strukture
- Primer: jezički modeli – modeli čiji je cilj da predvide sledeću/izostavljenu reč u nekoj sekvenci reči
  - Tačne izlazne vrednosti za ovaj problem se mogu „veštački“ generisati korišćenjem velikih korpusa neobeleženih tekstova
- Dosta zastupljeno u savremenoj obradi prirodnih jezika, naročito u domenu distribucione semantike
  - Pristup koji se koristi za generisanje vektora značenja reči (*word embeddings*) i kontekstno-osetljivih vektora značenja (*contextual embeddings*)

Polunadgledano učenje

- Kombinacija nadgledanog i nenadgledanog učenja
- Za manji deo ulaznih podataka x_s obeležene su željene vrednosti izlaza y_s
- Za veći deo ulaznih podataka x_u željene vrednosti izlaza nisu obeležene
- Cilj polunadgledanog učenja jeste iskorišćavanje neobeleženih ulaznih podataka radi boljeg obučavanja modela
- Ima veliku vrednost u praksi jer je za mnoge probleme teško dobiti dovoljne količine podataka sa obeleženim željenim izlazom

Učenje sa podrškom

- Uglavnom se primenjuje na obučavanje softverskih agenata koji deluju u nekom prostoru akcija
- Učenje se vrši na osnovu datih ulaznih podataka (koji predstavljaju akcije agenta) i signala podrške
- Signal podrške stiže tek na kraju nekog skupa akcija agenta
- Signal podrške može biti pozitivan ili negativan
  - Predstavlja željeni ili neželjeni ishod ponašanja agenta
- Na algoritmu učenja je da iskoristi signal podrške da utvrdi koja tačno akcija ili koji skup akcija je na kraju doveo do pozitivnog/negativnog signala podrške i da shodno tome koriguje ponašanje agenta

Učenje sa podrškom

- Zbog svoje prirode učenje sa podrškom je jako pogodno za probleme gde je „nagrada“ za uspeh dugoročna, a ne kratkoročna
- Često se primenjuje u
  - NLP – predominantno u izradi *chatbot*-ova
  - Robotici
  - Igranju igara
    - Šah
    - Go – *DeepMind AlphaGo* (pobedio svetskog šampiona u igri Go)
    - Video igre
- Za ovakve probleme klasično nadgledano učenje je veoma nepogodno
  - Npr. za igranje šaha bi trebalo da se sistemu za svaku moguću poziciju specificira najbolji sledeći potez

Mašinsko učenje u ovom kursu

- U ovom kursu će akcenat biti na tehnikama nadgledanog učenja
  - Pre svega na algoritme klasifikacije
- Nadgledano učenje trenutno ima najširi spektar upotrebe
  - Većina NLP sistema koji se danas oslanjaju na mašinsko učenje koriste predominantno ili bar delimično neki oblik nadgledanog mašinskog učenja
- Klasifikacija je jedan od najlakših problema za razumevanje
  - Dobro polazište za kasnije izučavanje kompleksnijih problema, poput strukturne predikcije
- Takođe će se razmotriti osnove samonadgledanog učenja
  - Izuzetno važna paradigma u osnovi mnogih *state-of-the-art* NLP sistema