Umetna inteligenca

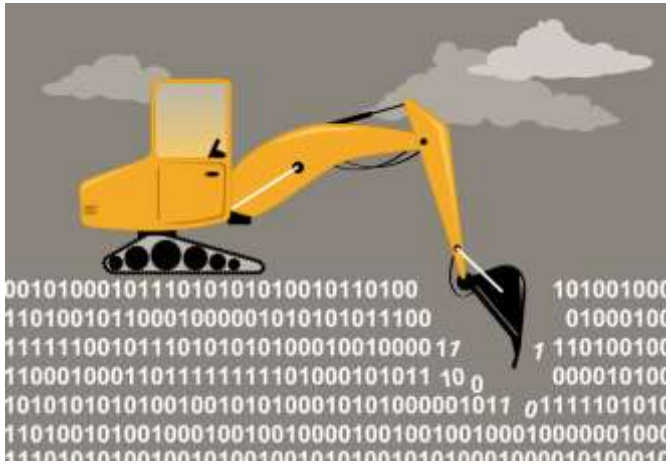
Pregled metod strojnega učenja:

- **Kaj je strojno učenje**
- **Osnovne metode strojnega učenja za napovedovanje**

Kaj je strojno učenje?

Prišel bo čas, ko bomo morali pozabiti vse, kar smo se naučili. (Ramana Maharshi)

- Strojno učenje (machine learning)
- Odkrivanje zakonitosti v podatkovnih bazah (knowledge discovery in databases)
- Podatkovno rudarjenje (data mining)

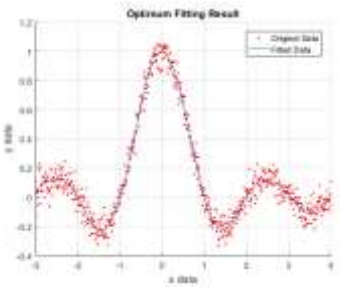


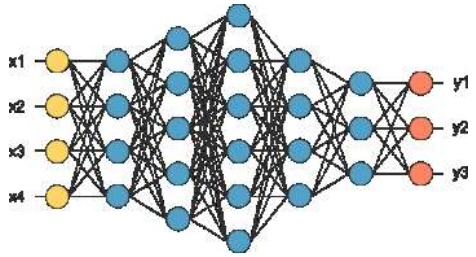
Strojno učenje

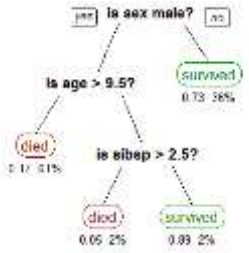
- Učenje iz podatkov - podatki so opisi problemov in njihove rešitve
- Modeliranje neznanega procesa, ki generira podatke.
- Rezultat je model, ki nudi razlago podatkov.
- Model (hipoteza, teorija) razlaga podatke in se lahko uporabi za:
  - napovedovanje
  - simulacijo
  - preverjanje
  - nadzor
  - diagnostiko
  - itd.
  - SKRATKA: za podporo odločanju

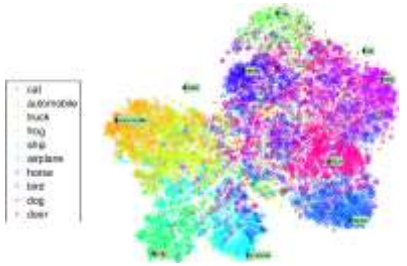
Strojno učenje

- Model (hipoteza, teorija) je podana v različnih oblikah:
  - pravila, odločitvena drevesa,
  - relacije,
  - enačba, sistemi enačb,
  - verjetnostne porazdelitve,
  - umetne nevronske mreže,
  - (prečiščeni) podatki – npr. izbrani tipični primeri rešenih problemov
  - itd.









Odkrivanje znanja iz podatkov

- 1. Razumevanje problemskega področja:
  - metodologije,
  - cilji,
  - kriteriji uspešnosti.
- 2. Razumevanje podatkov:
  - spoznavanje,
  - preverjanje kvalitete,
  - iskanje izjem.
- 3. Priprava podatkov:
  - zbiranje,
  - vrednotenje, vizualizacija,
  - poenotenje,
  - čiščenje, filtriranje,
  - transformiranje (diskretizacija, preslikave...).
- 4. Modeliranje, strojno učenje:
  - izbor ustrezne metode za strojno učenje,
  - gradnja in interno vrednotenje modelov,
  - ponavljanje postopkov.
- 5. Vrednotenje rezultatov:
  - vrednotenje glede na različne kriterije,
  - sprejem najboljših modelov,
  - ocena celotnega procesa in odločitev o naslednjem koraku.
- 6. Uporaba:
  - kdo in kdaj bo uporabljal rezultate,
  - problem prenosljivosti modela (na nove podatke),
  - praktična uporaba znanja.

Podatkovno rudarjenje (data mining)

- napovedovanje: klasifikacija in regresija in povezovalna pravila
- rudarjenje besedil (text mining)
- rudarjenje slik (image mining)
- rudarjenje grafov (graph mining)
- rudarjenje videa
- rudarjenje glasbe
- …

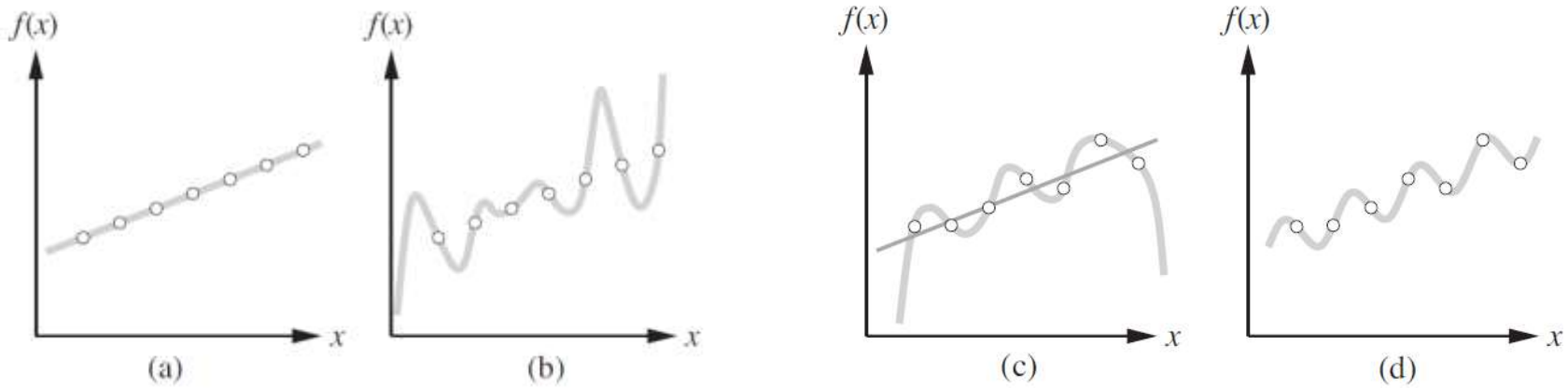
Pregled metod strojnega učenja

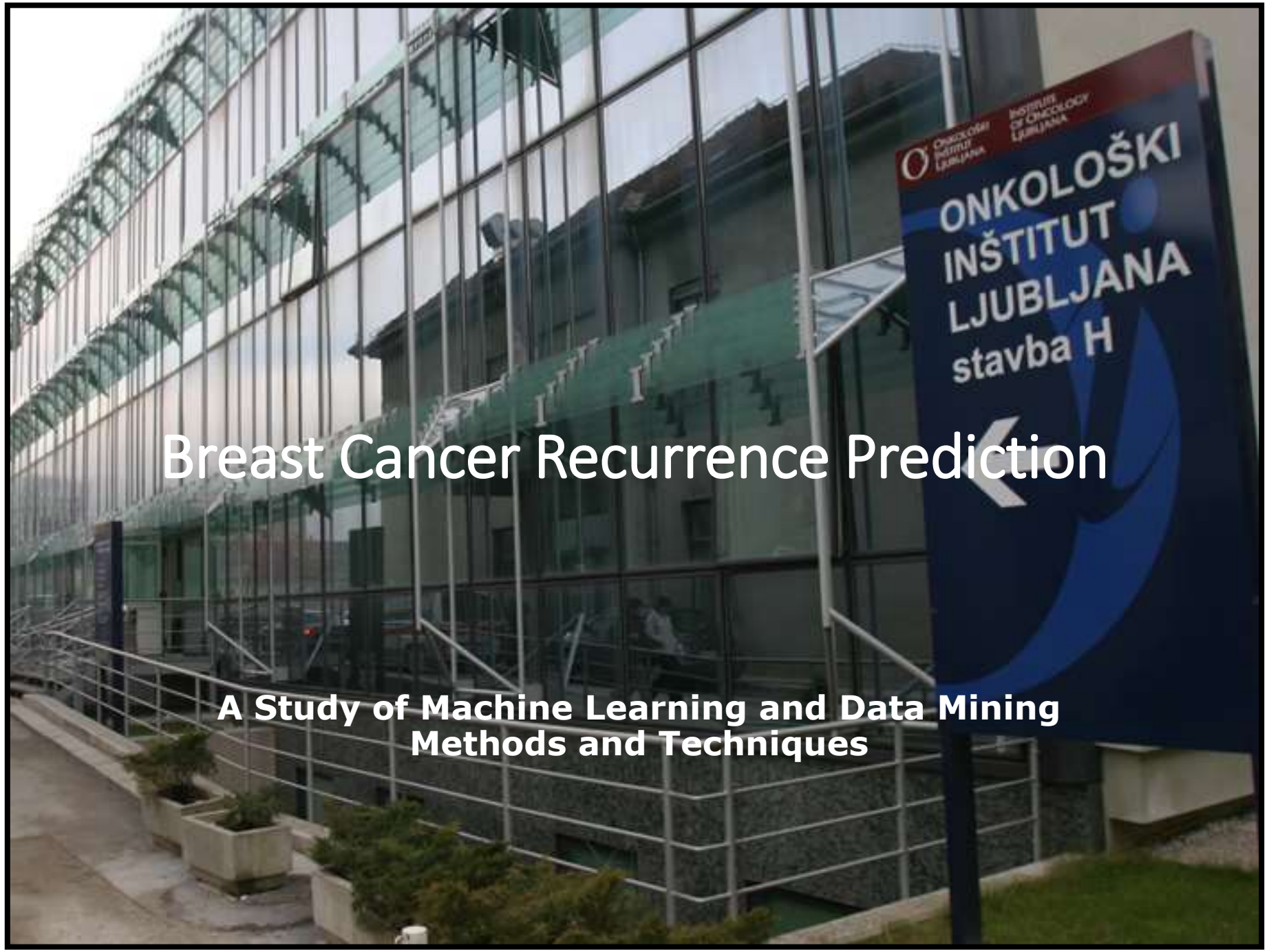
- nadzorovano učenje (supervised learning)
  - uvrščanje = klasifikacija (classification)
  - regresija (regression)
- nenadzorovano učenje (unsupervised learning)
  - razvrščanje = gručenje (clustering)
  - povezovalna pravila (association rules)
  - asociativne nevronske mreže
  - matrična faktorizacija (matrix factorization)
- učenje relacij (relational learning)
  - induktivno logično programiranje (inductive logic programming)
  - učenje sistemov (diferencialnih) enačb
- spodbujevano učenje (reinforcement learning)

Nadzorovano učenje: atributna predstavitev učnih primerov

- **podana**: množica **učnih primerov**
  $(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)$**,**
  kjer je vsak y_j vrednost neznane funkcije $y = f(x)$
- **naloga**: najdi funkcijo h, ki je najboljši približek funkciji f

- x_j so **atributi** ali **značilke** (vrednost ali vektor) ali **neodvisne spremenljivke**
- *y* je **ciljna** (target) ali **odvisna spremenljivka**
- funkcijo h imenujemo **hipoteza, teorija, model**
- primeri hipotez skozi dve množici točk:

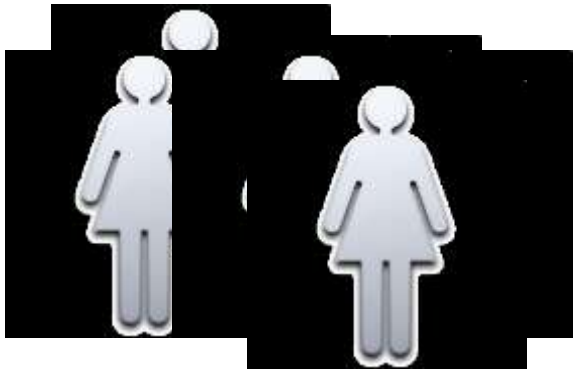




Breast Cancer Recurrence Prediction

A Study of Machine Learning and Data Mining Methods and Techniques

The Data



Provided by the Institute of Oncology, Ljubljana

Post-surgery data for about 1000 breast cancer patients.

+

Recurrence and time of recurrence.

The Data

	class1	class2	menop	stage	grade	hType	PgR	inv	nLymph	cTh	hTh	famHist	LVI	ER	maxNode	posRatio	age
300	11.82	0	1	2	2	1	0	0	1	1	0	3	0	1	2	3	2
301	4.89	1	0	1	2	1	0	0	2	1	0	0	0	2	1	4	3
302	14.63	0	1	1	4	2	0	0	0	0	0	1	0	1	1	1	3
303	21.83	0	0	1	4	2	1	0	1	0	0	9	0	4	1	2	2
304	19.87	0	0	1	2	1	0	0	0	0	0	0	0	1	2	1	2
305	7.54	0	1	2	3	1	9	2	1	0	1	1	0	3	3	3	4
306	15.15	0	0	1	4	2	1	0	0	0	0	2	0	4	1	1	2
307	0.30	1	0	2	2	1	0	0	3	0	0	9	0	1	1	4	2
308	12.49	0	1	2	2	3	1	0	0	0	0	0	0	4	1	1	5
309	1.77	1	0	2	3	1	1	2	2	1	0	9	1	3	3	3	2

Each patient is described with 17 values:
- 15 patient’s features
- 2 values, which describe the outcome

1 instance = 1 patient

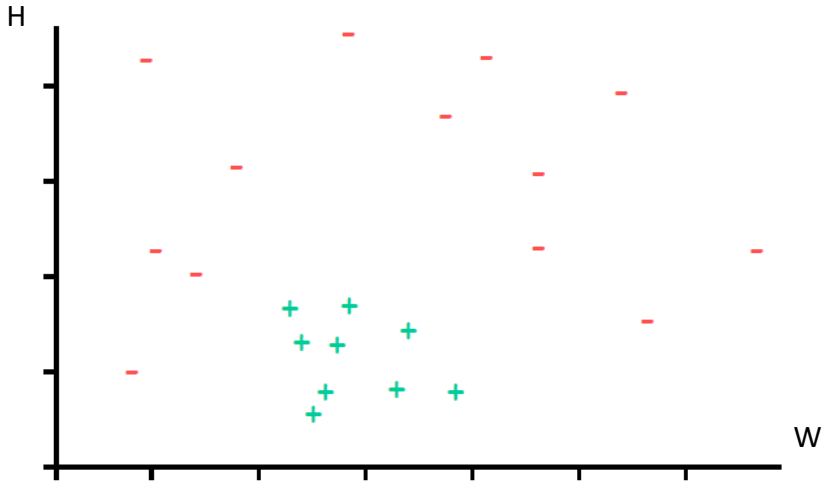
	class1	class2	menop	stage	grade	hType	PgR	inv	nLymph	cTh	hTh	famHist	LVI	ER	maxNode	posRatio	age
300	11.82	0	1	2	2	1	0	0	1	1	0	3	0	1	2	3	2
301	4.89	1	0	1	2	1	0	0	2	1	0	0	0	2	1	4	3
302	14.63	0	1	1	4	2	0	0	0	0	0	1	0	1	1	1	3
303	21.83	0	0	1	4	2	1	0	1	0	0	9	0	4	1	2	2
304	19.87	0	0	1	2	1	0	0	0	0	0	0	0	1	2	1	2
305	7.54	0	1	2	3	1	9	2	1	0	1	1	0	3	3	3	4
306	15.15	0	0	1	4	2	1	0	0	0	0	2	0	4	1	1	2
307	0.30	1	0	2	2	1	0	0	3	0	0	9	0	1	1	4	2
308	12.49	0	1	2	2	3	1	0	0	0	0	0	0	4	1	1	5
309	1.77	1	0	2	3	1	1	2	2	1	0	9	1	3	3	3	2

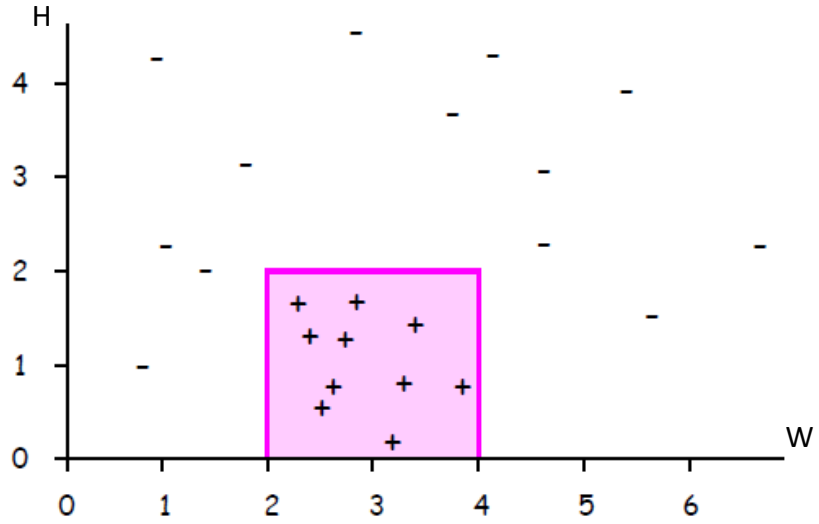
- Menopause?
- Tumor stage
- Tumor grade
- Histological type
- Progesterone receptor lvl.
- Invasive tumor type
- Number of positive lymph nodes

- Hormonal therapy?
- Chemotherapy?
- Family medical history
- Lymphovascular invasion?
- Estrogen receptor lvl.
- Size of max. removed node
- Ratio of positive lymph nodes
- Age group

Primer: gobe

- razpoznavanje užitnih gob
- atributa (x): W (width) in H (height)
- razred (y): strupena (-), užitna (+)

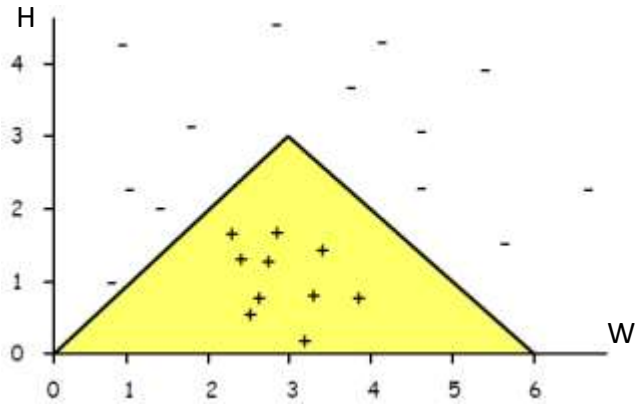




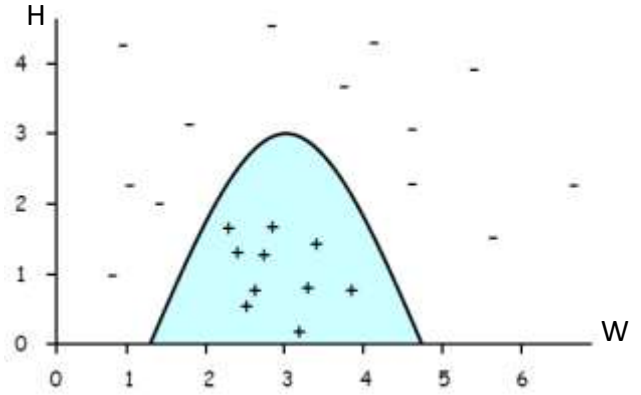

```
IF W>2 and W<4 and H<2
THEN "edible" ELSE "poisonous"
```

Primer: gobe

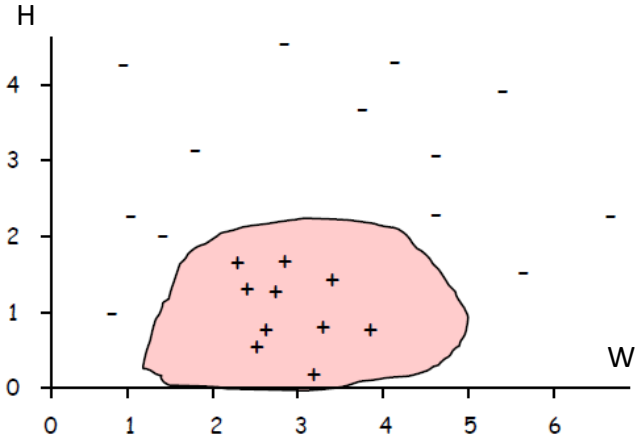
- ali pa …



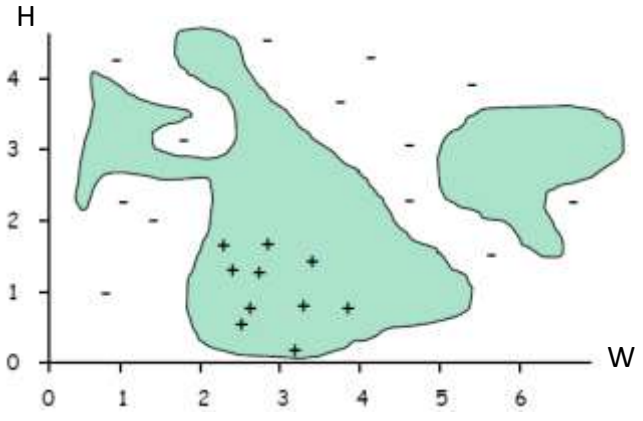
IF H > W THEN "poisonous"
ELSE IF H > 6 – W THEN "poisonous" ELSE "edible"



IF H < 3 – $(W-3)^2$ THEN "edible"
ELSE "poisonous"



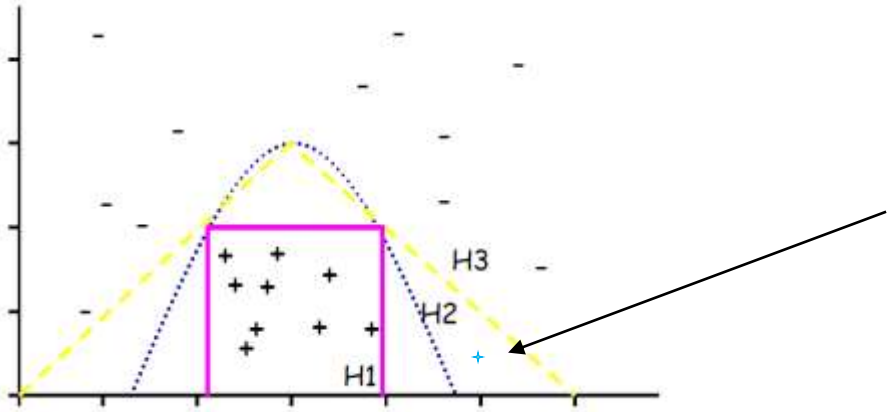
?



?

Primer: gobe

- prostor hipotez vsebuje več hipotez
- vse prikazane hipoteze so **konsistentne** z učno množico
- dobra hipoteza je dovolj splošna (**general**), kar pomeni, da pravilno napoveduje vrednost y za nove (še nevidene) primere

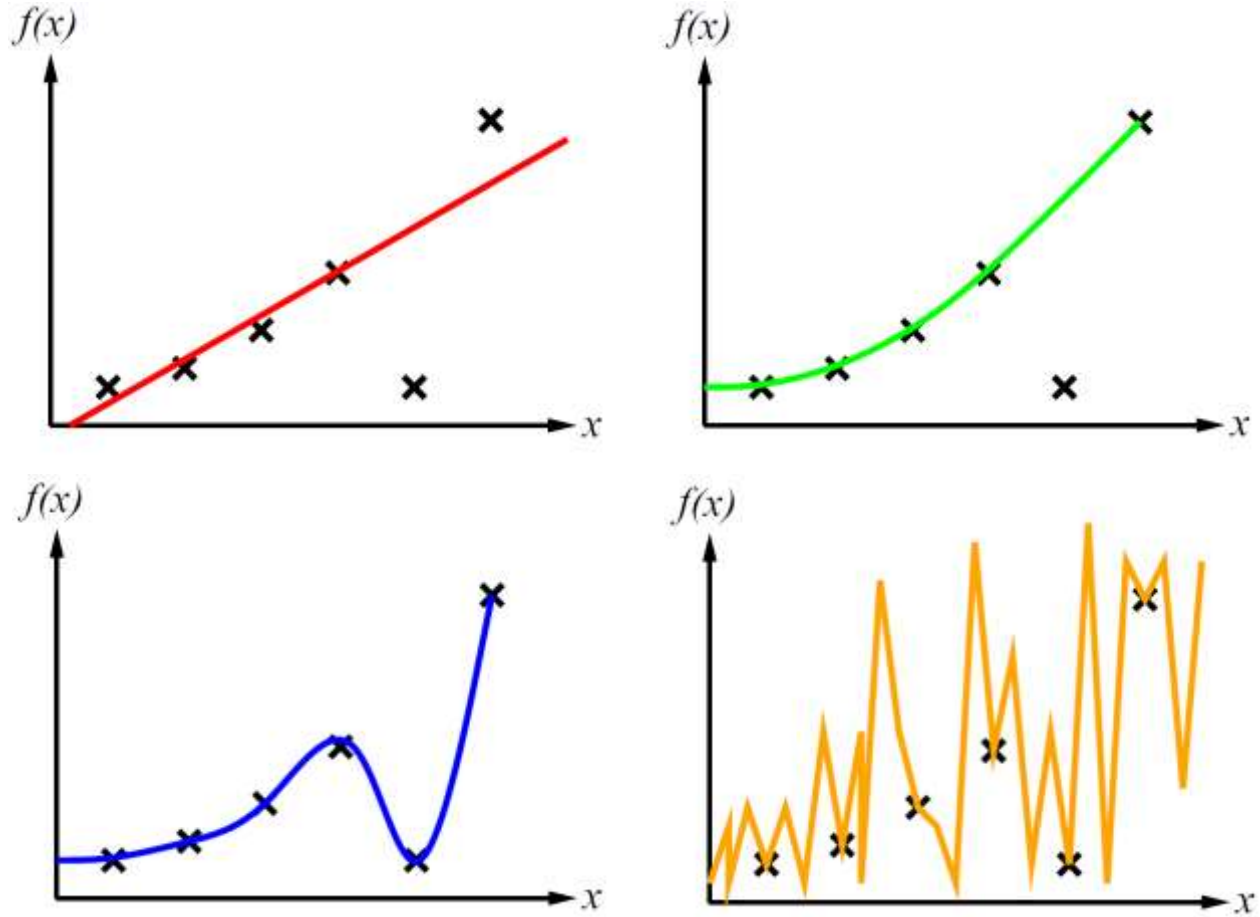


kam klasificirati ta primer?
(glede na H1 in H2 je -, glede na H3 je +)

- kako izbrati primerno hipotezo? Princip **Ockhamove britve** (*Ockham's razor*) (William o Ockham, 1320, angleški filozof):
  - najbolj verjetna hipoteza je najbolj preprosta hipoteza
  - *Given two explanations of the data, all other things being equal, the simpler explanation is preferrable.*

Primer

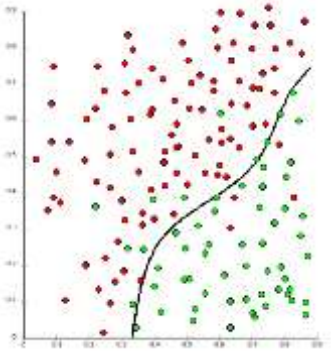
- podoben problem je tudi pri drugačnih primerih (iskanje funkcije, ki opisuje podane točke)

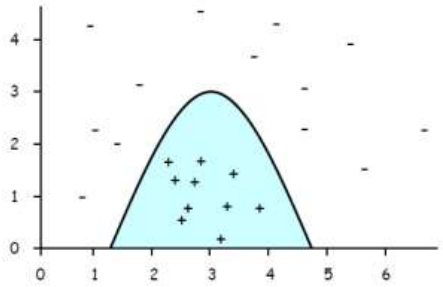


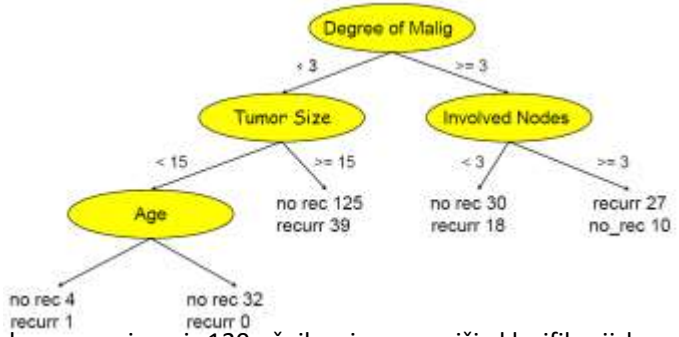
Vrste problemov

- **klasifikacija** in **regresija**

- **klasifikacija:**
  - y pripada **končnem naboru vrednosti** (je diskretna spremenljivka)
  - npr. $y \in \{užitna, strupena\}$, $y \in \{sonce, oblačno, dež\}$, $y \in \{zdrav, bolan\}$
  - y imenujemo **razred** *(angl. class)*
  - primeri:
    - napovedovanje vremena iz podatkov prejšnjih let
    - diagnosticiranje novih pacientov na osnovi znanih diagnoz za stare paciente
    - klasifikacija neželene elektronske pošte
    - napovedovanje vračila kredita



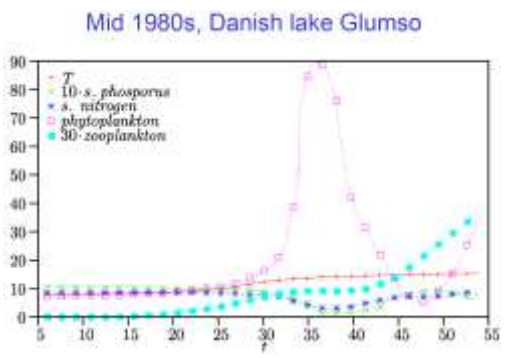




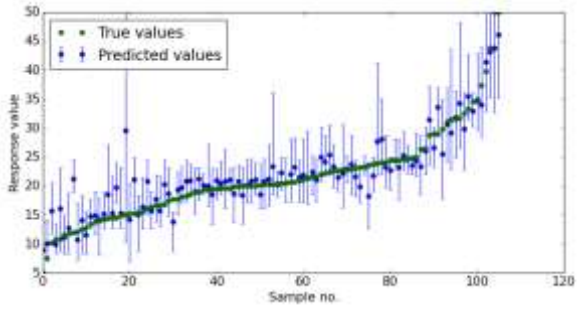
drevo, zgrajeno iz 139 učnih primerov; višja klasifikacijska točnost kot zdravniška

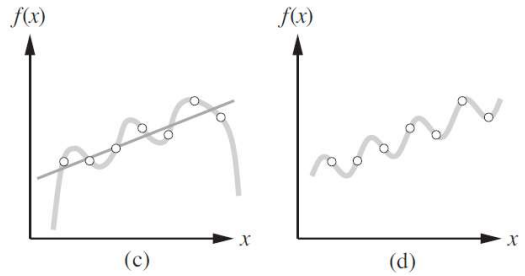
Vrste problemov

- **regresija:**
  - y je število (običajno $y \in \mathbb{R}$, je zvezna spremenljivka)
  - npr. y je temperatura,
  - y imenujemo **označba** (angl. *label*) ali **regresijska spremenljivka**
  - primeri:
    - napovedovanje razmnoževanja alg
    - medicinska prognostika
    - napovedovanje količine padavin
    - napovedovanje koncentracije ozona
    - napovedovanje gibanja cen delnic

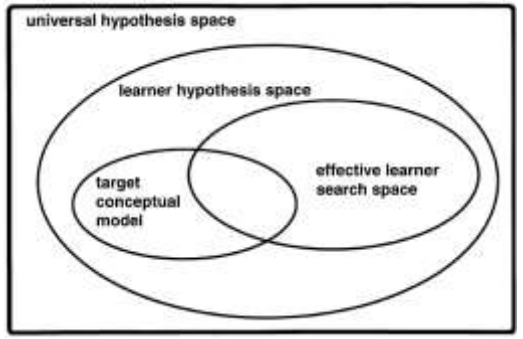


zakonitosti razmnoževanja alg





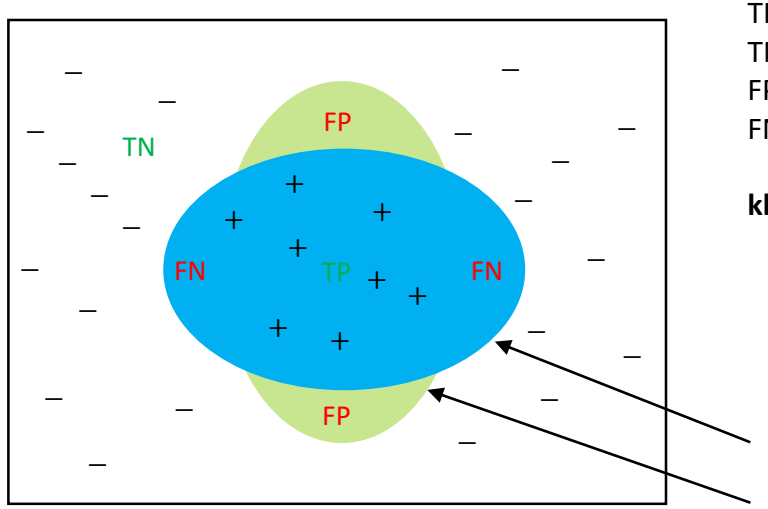
Prostor hipotez



- denimo, da imamo
  - binarno klasifikacijo
  - n binarnih atributov
- sledi:
  - → 2^n različnih učnih primerov
    → 2^{2^n} hipotez (denimo, da lahko hipotezo opišemo s tabelo napovedi za vse primere)
- primer:
  - za 10 atributov izbiramo med 10^{308} možnimi hipotezami
  - za 20 atributov izbiramo med $10^{300.000}$ možnimi hipotezami
  - v resnici: hipotez je že več, izračunavajo lahko isto funkcijo
- potrebujemo:
  - zavedanje o pristranosti hipotez
  - algoritme za gradnjo "dobrih" hipotez
  - metode za ocenjevanje hipotez / ocenjevanje učenja

Evalviranje hipotez

- pomembni kriteriji:
  - **konsistentnost** hipotez s primeri
  - **razumljivost** (interpretability, comprehensibility) hipotez
  - **točnost** hipotez:
    - točnost na učnih podatkih? (pristranost hipotez?)
    - točnost na novih podatkih?
    - točnost na testnih podatkih?
- ocenjevanje uspešnosti pri klasifikaciji:



TP – pravilno pozitivno klasificirani primeri (angl. *true positive*)
TN – pravilno negativno klasificirani primeri (angl. *true negative*)
FP – napačno pozitivno klasificirani primeri (angl. *false positive*)
FN – napačno negativno klasificirani primeri (angl. *false negative*)

klasifikacijska točnost (angl. classification accuracy):

$$CA = \frac{TP + TN}{TP + TN + FP + FN} = \frac{TP + TN}{N}$$

Primeri uporabe strojnega učenja:

Primeri uporabe strojnega učenja:

Primeri uporabe strojnega učenja:

Primeri uporabe strojnega učenja:

Primeri uporabe strojnega učenja:

Pregled metod strojnega učenja

- klasifikacija:
  - Odločitvena drevesa
  - naivni Bayesov klasifikator
  - Klasifikator z najbližjimi sosedi
  - Diskriminantne funkcije
  - metoda podpornih vektorjev (SVM)
  - Naključni gozdovi
  - Umetne nevronske mreže
  - Globoke nevronske mreže
- regresija:
  - Regresijska drevesa
  - Linearna regresija
  - Lokalno utežena regresija
  - Regresijske funkcije
  - Metoda podpornih vektorjev
  - Naključni gozdovi
  - Umetne nevronske mreže
  - Globoke nevronske mreže