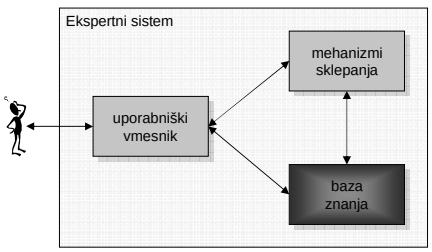
Strojno učenje

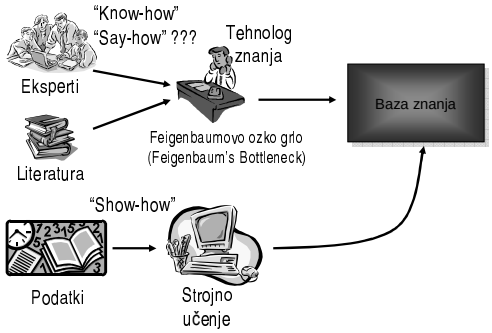
Marko Bohanec
Institut Jožef Stefan, Ljubljana in
Univerza v Novi Gorici

Kazalo

Arhitektura ES



Strojno učenje



povzeto po: Bojan Cestnik; Strojno učenje

Marko Bohanec

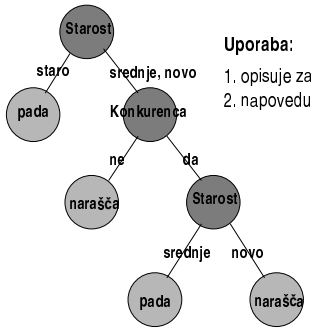
Primer

Dobiček	Starost	Konkurenca	Vrsta
pada	staro	ne	SW
pada	srednje	da	SW
narašča	srednje	ne	HW
pada	staro	ne	HW
narašča	novo	ne	HW
narašča	novo	ne	SW
narašča	srednje	ne	SW
narašča	novo	da	SW
pada	srednje	da	HW
pada	staro	da	SW

povzeto po: Bojan Cestnik; Strojno učenje

Marko Bohanec

Primer: Odločitveno drevo



Uporaba:

1. opisuje zakonitosti v obstoječih podatkih
2. napoveduje (klasificira) nove primere

povzeto po: Bojan Cestnik; Strojno učenje

Marko Bohanec

Metode strojnega učenja

- Statistične metode
  - Bayesov klasifikator
  - *k*-najbližjih sosedov (*k-Nearest Neighbors, k-NN*)
  - diskriminantna analiza
- Simbolično induktivno učenje
  - odločitvena drevesa (*Decision Trees*)
  - odločitvena pravila (*Decision Rules*)
  - učenje konceptov (*Concept Learning*)
  - indukcija logičnih programov (*ILP: Inductive Logic Programming*)
- Umetne nevronske mreže
  - večnivojske usmerjene NM
  - Kohonenove NM
  - Hopfieldove NM
  - Bayesove

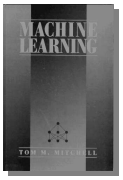
Zahteve pri strojnem učenju

- Zanesljivost delovanja
  - velika klasifikacijska točnost
- Transparentnost naučenega znanja
  - eksplicitna simbolična predstavitev, razumljiva ekspertom
- Sposobnost pojasnjevanja
  - argumentiranje in podpora ekspertnim odločitvam
- Odpornost na "šum" v podatkih
  - delovanje ob manjkajočih, nepopolnih ali nenatančnih podatkih
  - problemi iz realnega sveta

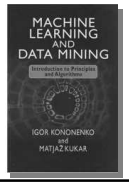
Nekatera področja uporabe

- Medicina
  - Diagnostika in prognostika
- Industrija
  - Kontrola kvalitete
  - Procesna kontrola
- Upravljanje in odločanje
  - Analiza podatkov o poslovanju
- "Podatkovno rudarjenje" (*Data Mining*)

Viri



Tom M. Mitchell: *Machine Learning.*
McGraw-Hill, 1997.



Igor Kononenko: *Strojno učenje, druga izdaja.*
Založba FE in FRI, 2005.

Kazalo

- Uvod
- Metode strojnega učenja
  - Atributno učenje
  - Učenje odločitvenih dreves
  - Statistične metode
  - Umetne nevronske mreže
  - HINT: učenje hierarhičnih modelov (DEX)

Atributno učenje

Podano:
- tabela "rešenih" primerov,
- opisanih z vrednostmi atributov A_1 do A_N
- in razredom C

C	A_1	A_2	…	A_N
C_1	$V_{1,1}$	$V_{1,2}$	…	$V_{1,N}$
C_2	$V_{2,1}$	$V_{2,2}$	…	$V_{2,N}$
…				
C_M	$V_{M,1}$	$V_{M,2}$	…	$V_{M,N}$

Naloga: poiskati pravilo za razred C glede na vrednosti A_1 do A_N

Odločitvena drevesa

- Vozlišča predstavljajo pogoje (teste)
  - praviloma: preverjanje vrednosti atributa
  - pogosto: binarna vejitev (pogoji tipa $A = v$ ali $A \geq v$)
- Listi predstavljajo razrede
- Drevo omogoča klasifikacijo novih primerov:
  - začni pri korenu
  - potuj navzdol v skladu z rezultati testov
  - odgovor (razred) je v listu

Marko Bohanec

Učenje odločitvenih dreves

***TDIDT**: Top-Down Induction of Decision Trees*

Oseba	Starost	Spol	Dohodki	Stranka
Ana Kranjc	32	Ž	10.000	da
Micka Kovač	53	Ž	1.000.000	da
Meta Novak	27	Ž	20.000	ne
Jana Bevc	55	Ž	20.000	da
Peter Dolenc	26	M	100.000	da
Janez Gorenc	50	M	200.000	da

primer povzet po: Sadio D2 erosk ij: Decision Trees

Marko Bohanec

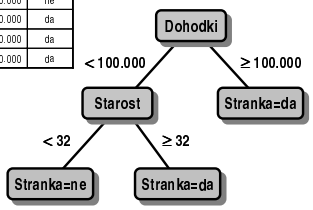
Klasifikacija in regresija

- **Klasifikacija**:
  - razred C je diskretna spremenljivka
  - pravilo (Oseba, Starost, Spol, Dohodki) $\Rightarrow$ Stranka
- **Regresija**:
  - "razred" C je zvezna spremenljivka
  - pravilo (Oseba, Starost, Spol, Stranka) $\Rightarrow$ Dohodki

Marko Bohanec

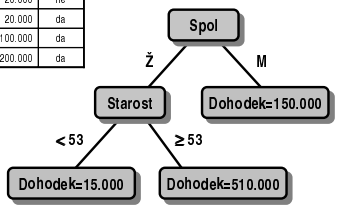
Klasifikacijsko odločitveno drevo

Oseba	Starost	Spol	Dohodki	Stranka
Ana Kranjc	32	Ž	10.000	da
Micka Kovač	53	Ž	1.000.000	da
Meta Novak	27	Ž	20.000	ne
Jana Bevc	55	Ž	20.000	da
Peter Dolenc	26	M	100.000	da
Janez Gorenc	50	M	200.000	da



Regresijsko odločitveno drevo

Oseba	Starost	Spol	Dohodki	Stranka
Ana Kranjc	32	Ž	10.000	da
Micka Kovač	53	Ž	1.000.000	da
Meta Novak	27	Ž	20.000	ne
Jana Bevc	55	Ž	20.000	da
Peter Dolenc	26	M	100.000	da
Janez Gorenc	50	M	200.000	da



primer povzet po: Sadio D2 erovski, Decision Trees

Učenje odločitvenih dreves

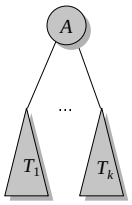
KLJUČNI KONCEPTI

1. Gradnja drevesa
   - algoritem
   - izbiranje atributov
2. Preverjanje kakovosti drevesa
   - učna in testna množica
   - klasifikacijska točnost
3. Rezanje drevesa
   - rezanje naprej
   - rezanje nazaj

Gradnja klasifikacijskega drevesa

ALGORITEM

- Če vsi učni primeri pripadajo istemu razredu C_i potem je rezultat list C
- Sicer
  - Izberi *najboljši* atribut A (ali *najboljšo* delitev po A)
  - Razdeli učno množico glede na vrednosti A
  - Rekurzivno zgradi poddrevesa $T_1..T_k$ za vsako podmnožico
  - Rezultat je drevo z vozliščem A in poddrevesi $T_1..T_k$

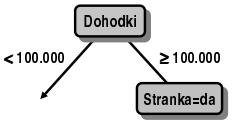


Primer

Oseba	Starost	Spol	Dohodki	Stranka
Ana Kranjc	32	Ž	10.000	da
Micka Kovač	53	Ž	1.000.000	da
Meta Novak	27	Ž	20.000	ne
Jana Bevc	55	Ž	20.000	da
Peter Dolenc	26	M	100.000	da
Janez Gorenc	50	M	200.000	da

Vsi primeri v istem razredu?

***Najboljša* delitev?**



Oseba	Starost	Spol	Dohodki	Stranka
Ana Kranjc	32	Ž	10.000	da
Meta Novak	27	Ž	20.000	ne
Jana Bevc	55	Ž	20.000	da

Oseba	Starost	Spol	Dohodki	Stranka
Micka Kovač	53	Ž	1.000.000	da
Peter Dolenc	26	M	100.000	da
Janez Gorenc	50	M	200.000	da

Izbiranje atributov (delitev)

ZAHTEVA: Delitev na čimbolj "čiste" podmnožice

MERE "NEČISTOČE"

za dva razreda, $p(C_1)=p_1$, $p(C_2)=p_2$

- Entropija E: $-p_1\log_2 p_1 - p_2\log_2 p_2$
- Napaka prevladujočega razreda: $1-\max(p_1,p_2)$
- Indeks Gini: $1-(p_1^2+p_2^2)$

INFORMACIJSKI PRISPEVEK: Koliko pridobimo ob delitvi?

- $\text{Gain}(S,A) = E(S) - \sum_v |S_v|/|S| \; E(S_v)$
- maksimiziramo Gain

Primer

(1 od 3)

Vreme	Temp	Vlaga	Veter	Tenis
sončno	vroče	visoka	ne	ne
sončno	vroče	visoka	da	ne
oblačno	vroče	visoka	ne	da
dež	zmerno	visoka	ne	da
dež	hladno	norm	ne	da
dež	hladno	norm	da	ne
oblačno	hladno	norm	da	da
sončno	zmerno	visoka	ne	ne
sončno	hladno	norm	ne	da
dež	zmerno	norm	ne	da
sončno	zmerno	norm	da	da
oblačno	zmerno	visoka	da	da
oblačno	vroče	norm	ne	da
dež	zmerno	visoka	da	ne

Vreme? sončno [2+, 3–] E=0,97
oblačno [4+, 0–] E=0
dež [3+, 2–] E=0,97

Vlaga? visoka [3+, 4–] E=0,99
norm [6+, 1–] E=0,59

Veter? ne [6+, 2–] E=0,81
da [3+, 3–] E=1,00

Marko Bohanec

Primer

(2 od 3)

Veter? ne [6+, 2–] E=0,81
da [3+, 3–] E=1,00

Gain(S,A)= $E(S) - \sum_v |S_v|/|S| \; E(S_v)$

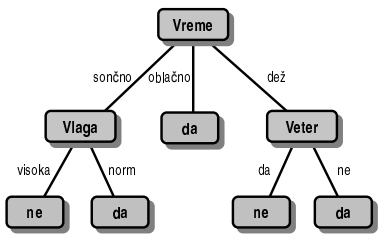
$E(S) = E(9+, 5–) = -(9/14)\log_2(9/14) - (5/14)\log_2(5/14) = 0{,}94$
$E(S_{\text{Veter=ne}}) = 0{,}81$
$E(S_{\text{Veter=da}}) = 1{,}00$
Gain(S,Veter) = 0,94–(8/14)0,81–(6/14)1,00 = **0,048**

Gain(S,Vreme) = **0,246 (max)**
Gain(S,Vlaga) = **0,151**
Gain(S,Temp) = **0,029**

Marko Bohanec

Primer

(3 od 3)



Marko Bohanec

Mere kvalitete odločitvenih dreves

Klasifikacijska točnost:
Kako točno drevo klasificira nove primere?
Kakšna je točnost v primerjavi z *apriorno* ("*naivni klasifikator*")?

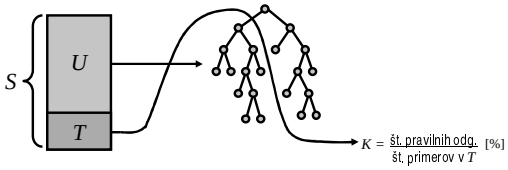
Razumljivost:
Ali ekspert razume drevo in njegovo vsebino?
Ali ga lahko interpretira, utemelji?

Velikost:
Povezano z razumljivostjo: zaželena *čim manjša* drevesa!

Marko Bohanec

Klasifikacijska točnost

Postopek gradnje in preverjanja klasifikatorja:
- množico primerov S razdelimo na:
  - ***učno*** množico U (npr. 70%) in
  - ***testno*** množico T (30%)
- zgradimo drevo upoštevajoč samo U
- na T preverimo ***točnost klasifikacije*** = delež pravilnih klasifikacij

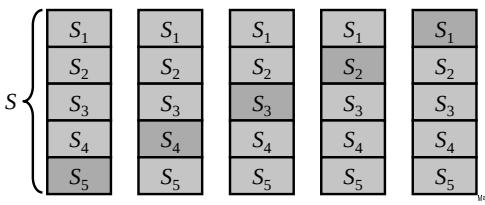


Marko Bohanec

Prečno preverjanje

Slov: Prečno preverjanje, npr. "10-kratno prečno preverjanje"
Angl: "Cross validation", e.g., "10-fold cross validation"

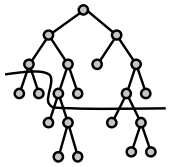
Postopek preverjanja klasifikatorja:
- celotno množico primerov S naključno razdelimo na n čim bolj enakih delov
- n krat ponovimo: učimo na n–1 delih, testiramo na preostalem enem delu
- celotna točnost klasifikacije je povprečje n takšnih klasifikacij



Marko Bohanec

Rezanje dreves

Spodnji deli drevesa (okrog listov) so manj zanesljivi:
- manjše število učnih primerov
- prevelika prilagoditev podatkom (*overfitting*)



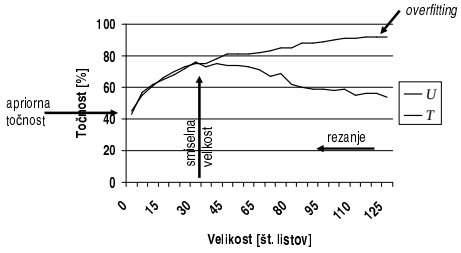
Rezanje (*pruning*):
- **Naprej**: predčasno ustavimo gradnjo
- **Nazaj**: drevo zgradimo do konca, nato režemo manj zanesljive dele (bolje!)

Pridobimo:
- Manjše drevo – večja preglednost in razumljivost
- Večja točnost na testni množici primerov

Marko Bohanec

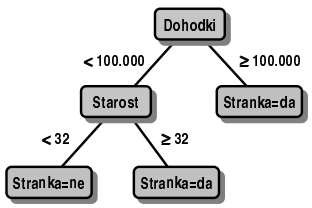
Rezanje dreves

Klasifikacijska točnost v odvisnosti od velikosti drevesa



Marko Bohanec

Od dreves k pravilom



PRAVILA
if Dohodki ≥ 100.000
  then Stranka=da
if Dohodki < 100.000 **and**
  Starost ≥ 32
  then Stranka=da
if Dohodki < 100.000 **and**
  Starost < 32
  then Stranka=ne

ODLOČITVENI SEZNAM
if Dohodki ≥ 100.000
  then Stranka=da
else if Starost ≥ 32
  then Stranka=da
  else Stranka=ne

Marko Bohanec

Nekaj učnih algoritmov in programov

KLASIFIKACIJA

- ID3 (Quinlan 1979)
- CART (Breiman et al. 1984)
- Assistant (Cestnik et al. 1987)
- C4.5 (Quinlan 1993)
- C5.0, See5 (RuleQuest)

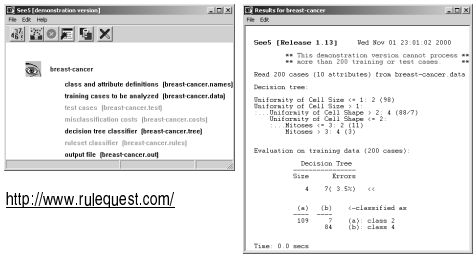
REGRESIJA

- CART (Breiman et al. 1984)
- RETIS (Karalič 1992)
- M5 (Quinlan 1993)
- Cubist (RuleQuest)

- Weka (Waikato University, NZ)
- Orange (Univerza v Ljubljani, FRI)

Marko Bohanec

See5 (RuleQuest)

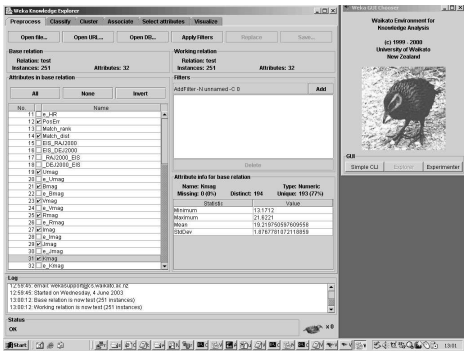


http://www.rulequest.com/

Marko Bohanec

Weka

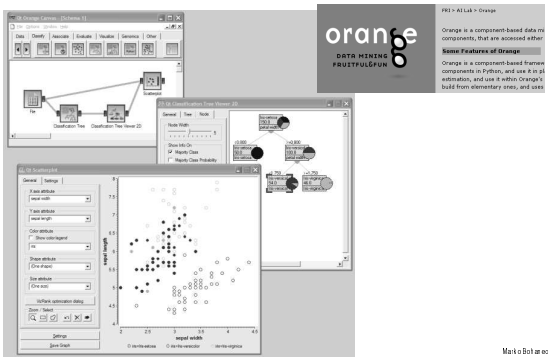
http://www.cs.waikato.ac.nz/~ml/weka/



Marko Bohanec

Orange

http://magix.fri.uni-lj.si/orange/



Marko Bohanec

Kazalo

- Uvod
- Metode strojnega učenja
  - Atributno učenje
  - Učenje odločitvenih dreves
  - Statistične metode
    - Bayesovo pravilo
    - *k*-NN: *k* najbližjih sosedov
  - Umetne nevronske mreže
  - HINT: učenje hierarhičnih modelov (DEX)

Marko Bohanec

Bayesovo pravilo

- *Bayesovo pravilo*: verjetnost razreda C pri pogoju $V_1..V_n$

$$P(C|V_1..V_n) = P(C)\frac{P(V_1..V_n|C)}{P(V_1..V_n)}$$

- *"Naivno" Bayesovo pravilo*: predpostavka neodvisnosti $V_1..V_n$

$$P(C|V_1..V_n) = P(C)\prod_{i=1}^{n}\frac{P(C|V_i)}{P(C)}$$

Marko Bohanec

Primer

Vreme	Temp	Vlaga	Veter	Tenis
sončno	vroče	visoka	ne	ne
sončno	vroče	visoka	da	ne
oblačno	vroče	visoka	ne	da
dež	zmerno	visoka	ne	da
dež	hladno	norm	ne	da
dež	hladno	norm	da	ne
oblačno	hladno	norm	da	da
sončno	zmerno	visoka	ne	ne
sončno	hladno	norm	ne	da
dež	zmerno	norm	ne	da
sončno	zmerno	norm	da	da
oblačno	zmerno	visoka	da	da
oblačno	vroče	norm	ne	da
dež	zmerno	visoka	da	ne

Vreme	Temp	Vlaga	Veter	Tenis
sončno	hladno	norm	da	?

$$P(C|V_1..V_n) = P(C)\prod_{i=1}^{n}\frac{P(C|V_i)}{P(C)}$$

$$P(da|son,hla,norm,da) =$$

$$= P(da)\frac{P(da|son)}{P(da)}\frac{P(da|hla)}{P(da)}\frac{P(da|norm)}{P(da)}\frac{P(da|da)}{P(da)}$$

$$= \frac{9}{14}\left(\frac{2}{9}\frac{4}{9}\right)\left(\frac{3}{9}\frac{4}{9}\right)\left(\frac{6}{9}\frac{1}{9}\right)\left(\frac{3}{9}\frac{4}{9}\right) = \mathbf{0,48}$$

$\Rightarrow$ **Tenis=da**

$$P(ne|son,hla,norm,da) =$$

$$= P(ne)\frac{P(ne|son)}{P(ne)}\frac{P(ne|hla)}{P(ne)}\frac{P(ne|norm)}{P(ne)}\frac{P(ne|da)}{P(ne)}$$

$$= \frac{5}{14}\left(\frac{3}{5}\frac{4}{5}\right)\left(\frac{1}{4}\frac{1}{5}\right)\left(\frac{1}{7}\frac{1}{5}\right)\left(\frac{3}{6}\frac{4}{5}\right) = \mathbf{0,24}$$

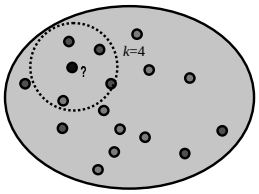
Marko Bohanec

K-najbližjih sosedov (*k*-NN)

- Pri klasifikaciji primera $X = V_1..V_n$ v učni množici poiščemo *k najbližjih sosedov*
- Primer X klasificiramo v večinski razred C

Definirati je potrebno:

- k
- mero razdalje med primeri



Marko Bohanec

Primer

Vreme	Temp	Vlaga	Veter	Tenis	Razd
sončno	vroče	visoka	ne	ne	4
sončno	vroče	visoka	da	ne	3
oblačno	vroče	visoka	ne	da	5
dež	zmerno	visoka	ne	da	5
dež	hladno	norm	ne	da	3
dež	hladno	norm	da	ne	2
oblačno	hladno	norm	da	da	1
sončno	zmerno	visoka	ne	ne	3
sončno	hladno	norm	ne	da	1
dež	zmerno	norm	ne	da	4
sončno	zmerno	norm	da	da	1
oblačno	zmerno	visoka	da	da	3
oblačno	vroče	norm	ne	da	4
dež	zmerno	visoka	da	ne	4

Vreme	Temp	Vlaga	Veter	Tenis
sončno	hladno	norm	da	?

$k = 4$, razdalja "Manhattan"

$\Rightarrow$ **Tenis=da** (3:1)

Marko Bohanec

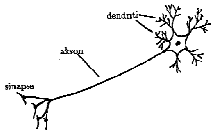
Kazalo

- Uvod
- Metode strojnega učenja
  - Atributno učenje
  - Učenje odločitvenih dreves
  - Statistične metode
    - Bayesovo pravilo
    - *k*-NN: *k* najbližjih sosedov
  - Umetne nevronske mreže
  - HINT: učenje hierarhičnih modelov (DEX)

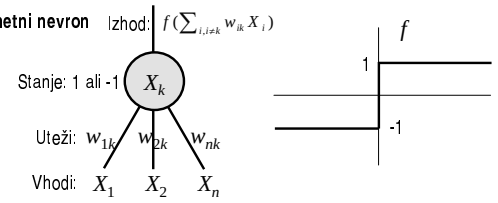
Marko Bohanec

Umetne nevronske mreže

Naravni nevron

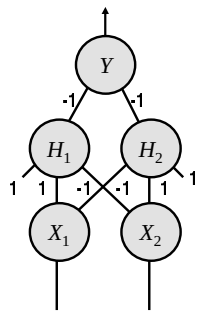


Umetni nevron Izhod: $f(\sum_{i,i\neq k} w_{ik} X_i)$



Marko Bohanec

Primer: XOR (ekskluzivni "ali")

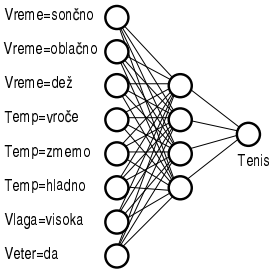


X_1	X_2	Y
1	1	-1
1	-1	1
-1	1	1
-1	-1	-1

Marko Bohanec

Primer

Vreme	Temp	Vlaga	Veter	Tenis
sončno	vroče	visoka	ne	ne
sončno	vroče	visoka	da	ne
oblačno	vroče	visoka	ne	da
dež	zmerno	visoka	ne	da
dež	hladno	norm	ne	da
dež	hladno	norm	da	ne
oblačno	hladno	norm	da	da
sončno	zmerno	visoka	ne	ne
sončno	hladno	norm	ne	da
dež	zmerno	norm	ne	da
sončno	zmerno	norm	da	da
oblačno	zmerno	visoka	da	da
oblačno	vroče	norm	ne	da
dež	zmerno	visoka	da	ne



Marko Bohanec

Lastnosti umetnih nevronskih mrež

- Matematična osnova
  linearna algebra
- Paralelizem
  nevroni hkrati izvajajo operacije
- Večsmerno izvajanje
  vhod je lahko tudi izhod in obratno
- Robustnost
  majhna občutljivost na okvare nevronov in sinaps
- Učenje
  s spontanim spreminjanjem uteži
  razmeroma počasno
- Razlaga odločitve
  slaba, praktično nemogoča

Marko Bohanec

Kazalo

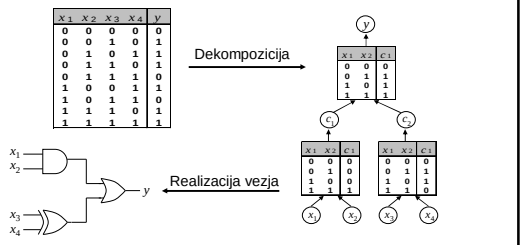
- Uvod
- Metode strojnega učenja
  - Atributno učenje
  - Učenje odločitvenih dreves
  - Statistične metode
    - Bayesovo pravilo
    - *k*-NN: *k* najbližjih sosedov
  - Umetne nevronske mreže
  - HINT: učenje hierarhičnih modelov (DEX)

Marko Bohanec

HINT: Učenje hierarhičnih modelov

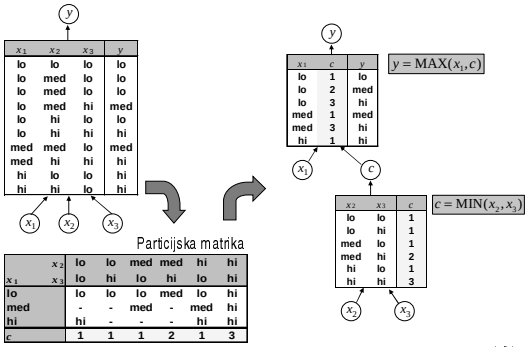
Cilj: učenje hierarhičnih modelov tipa DEX

Algoritem temelji na Ashenhurst-Curtisovi dekompoziciji logičnih funkcij



Marko Bohanec

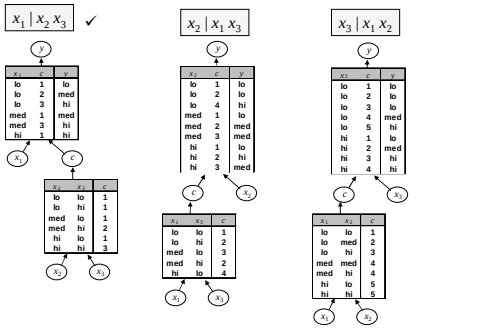
HINT: Osnovni korak



x_1 \ x_2 x_3	lo lo	lo hi	med lo	med hi	hi lo	hi hi
lo	lo	lo	lo	med	lo	hi
med	-	-	med	-	med	hi
hi	hi	-	-	-	hi	hi
c	1	1	1	2	1	3

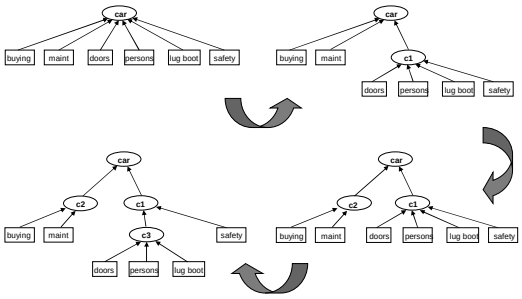
Marko Bohanec

HINT: Izbira najboljše delitve



Marko Bohanec

HINT: Oblikovanje hierarhije konceptov



Marko Bohanec

Strojno učenje: Povzetek

- Rešujejo problem "ozkega grla" (*Feigenbaum's Bottleneck*)
- Učenje je "avtomatizirano" (skupaj z ekspertom), ne "avtomatsko"
- Številne metode: statistične, simbolične, nevronske mreže
- Pomembni koncepti:
  - klasifikacija, regresija
  - klasifikacijska točnost
  - razumljivost, velikost modelov
  - učna in testna množica
  - prilagoditev podatkom (*overfitting*)
- Odločitvena drevesa: razširjena, pogosta uporaba; koncept rezanja
- Statistične metode: dobre, a ne gradijo opisov konceptov
- Nevronske mreže: zanimive lastnosti, zahtevno učenje, problem razumljivosti
- HINT: učenje konceptov po načelu razgradnje funkcij

Marko Bohanec