

Algoritmi za klasifikacijo:

osnovne metode

Vsebina

- Ni pravi klasifikator: **0R**
- Enostavni klasifikator: **1R**
- **Naivni Bayes**

Klasifikacija

- Naloga: zgraditi model ali *klasifikator* z uporabo že klasificiranih primerov (poznan razred) ter ga uporabiti za klasifikacijo "novih" primerov (neznan razred).
- Nadzorovano učenje: vrednost razreda primerov, ki se uporabljajo za izgradnjo modela, je poznana.
- Klasifikator je lahko: množica pravil, odločitveno drevo, nevronska mreža ...
- Tipične aplikacije: odobritve kreditov, neposredni marketing, odkrivanje prevar, diagnoze v medicini ...

Enostavni algoritmi



- Enostavni algoritmi pogosto zelo dobro delujejo!
- Mnogo primerov enostavnih struktur:
 - Klasifikator "večinskega razreda"
 - Le en atribut "opravi vse delo"
 - Vsi atributi prispevajo v enaki meri in neodvisno
 - Utežena linearna kombinacija
 - Na podlagi razdalje med primeri
 - Preprosta logična pravila
- Uspeh algoritma je odvisen od podatkov

Učenje enostavnih pravil

- **0R:** napove večinski razred
-
- **1R:** generira 1-nivojsko odločitveno drevo
 - oz. pravila, ki vsa testirajo le določen atribut
 - Osnovna različica 1R algoritma:
 - Vsaka vrednost atributa določa eno vejo v drevesu
 - V vsaki veji drevesa napovemo večinski razred
 - Napaka: delež primerov, ki ne pripadajo večinskemu razredu v določeni veji drevesa
 - Izberemo atribut z najmanjšo napako

(predpostavlja, da so vsi atributi nominalni)

Psevdokoda algoritma 1R

Za vsak atribut:

Za vsako vrednost atributa naredi pravilo na sledeči način:

preštej koliko pogosto nastopa vsaka vrednost razreda,

poišči najpogostejšo vrednost razreda,

naredi pravilo, ki priredi to vrednost razreda vrednosti atributa,

izračunaj napako pravila.

Izračunaj skupno napako 1-nivojskega drevesa za atribut

Izberi drevo atributa z najmanjšo skupno napako

- Pozor: manjkajoče vrednosti atributov so obravnavane kot ločene vrednosti

Primer: 1R na "weather" podatkih

Outlook	Temp	Humidity	Windy	Play
Sunny	Hot	High	False	No
Sunny	Hot	High	True	No
Overcast	Hot	High	False	Yes
Rainy	Mild	High	False	Yes
Rainy	Cool	Normal	False	Yes
Rainy	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Sunny	Mild	High	False	No
Sunny	Cool	Normal	False	Yes
Rainy	Mild	Normal	False	Yes
Sunny	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Rainy	Mild	High	True	No

Atribut	Pravila	Napake	Skupna napaka
Outlook*	Sunny → No	2/5	4/14
	Overcast → Yes	0/4	
	Rainy → Yes	2/5	
Temp	Hot → No*	2/4	5/14
	Mild → Yes	2/6	
	Cool → Yes	1/4	
Humidity*	High → No	3/7	4/14
	Normal → Yes	1/7	
Windy	False → Yes	2/8	5/14
	True → No*	3/6	


* označuje izenačen izid

Obravnava numeričnih atributov

- Z uporabo "razredne" diskretizacije
- Razpon vrednosti atributa razdelimo na intervale:
 - Uredimo vrednosti atributa po velikosti;
 - Postavimo meje intervalov, kjer se spremeni vrednost razreda:
 - razen tam, kjer se vrednost atributa ne spremeni;
 - Na ta način minimiziramo napako;
- Primer: atribut *temperature* – podatki "weather"

Outlook	Temperature	Humidity	Windy	Play
Sunny	85	85	False	No
Sunny	80	90	True	No
Overcast	83	86	False	Yes
Rainy	75	80	False	Yes
...

64 65 68 69 70 71 72 72 75 75 80 81 83 85

Yes | No | Yes Yes Yes | No No  Yes Yes Yes | No | Yes Yes | No

Problem prekomernega prileganja

- Ta način diskretizacije je (zelo) občutljiv na šum:
 - primer z napačno vrednostjo razreda bo zelo verjetno povzročil tvorbo novega intervala pri diskretizaciji;
- Ekstremni primer:
atribut *time stamp* bo diskretiziran z napako nič;
- Preprosta rešitev:
določimo mejo = najmanjše dovoljeno število primerov v posameznem intervalu

Primer "razredne" diskretizacije z mejo

- Primer (min. št. primerov = 3):

64	65	68	69	70	71	72	72	75	75	80	81	83	85
Yes	Ⓢ No	Ⓢ Yes	Yes	Yes	No	No	Yes	Yes	Yes	No	Ⓢ Yes	Yes	Ⓢ No

- Končni rezultat diskretizacije za atribut *temperature*

64	65	68	69	70	71	72	72	75	75	80	81	83	85
Yes	No	Yes	Yes	Yes	No	No	Yes	Yes	Yes	No	Yes	Yes	No

1R + diskretizacija z mejo

- Primer pravil za "weather-numeric" podatke:

Atribut	Pravila	Napake	Skupna napaka
Outlook	Sunny → No	2/5	4/14
	Overcast → Yes	0/4	
	Rainy → Yes	2/5	
Temperature	$\leq 70.5 \rightarrow \text{Yes}$	1/5	5/14
	$> 70.5 \text{ and } \leq 77.5 \rightarrow \text{Yes}$	2/5	
	$> 77.5 \rightarrow \text{No}^*$	2/4	
Humidity	$\leq 82.5 \rightarrow \text{Yes}$	1/7	3/14
	$> 82.5 \text{ and } \leq 95.5 \rightarrow \text{No}$	2/6	
	$> 95.5 \rightarrow \text{Yes}$	0/1	
Windy	False → Yes	2/8	5/14
	True → No*	3/6	

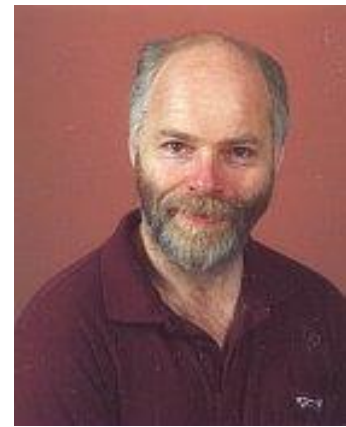
* označuje izenačen izid

Zgodovina algoritma 1R

- 1R je prvi opisal R.C. Holte v svojem članku (1993):
 - Eksperiment na 16 bazah podatkov (uporaba *prečnega preverjanja* za oceno modelov);
 - Meja minimalnega števila primerov v intervalu za numerične attribute nastavljena na 6 (rezultat isprobavanja večih vrednosti meje);
 - Preprosta pravila algoritma 1R so se odrezala ne dosti slabše od zahtevnejših odločitvenih dreves;
- Zaključek: splača se najprej poskusiti preproste metode!

Very Simple Classification Rules Perform Well on Most Commonly Used Datasets

Robert C. Holte, Computer Science Department, University of Ottawa



Bayes-ovo (statistično) modeliranje



- "Nasprotno" od 1R: upoštevajo se vsi atributi
- Dve predpostavki – atributi naj bodo:
 - *Enako pomembni* (za napoved razreda),
 - *Statistično neodvisni* (pri podanem razredu):
 - če poznamo vrednost nekega atributa, ne moremo sklepati na vrednosti ostalih atributov (pri poznani vrednosti razreda)
- Predpostavka "neodvisnosti" skoraj nikoli ne drži!
- Ampak ... Bayes-ov model v praksi dobro deluje

Primer: "weather" podatki

Outlook			Temperature			Humidity			Windy			Play	
Yes No			Yes No			Yes No			Yes No			Yes No	
Sunny	2	3	Hot	2	2	High	3	4	False	6	2	9	5
Overcast	4	0	Mild	4	2	Normal	6	1	True	3	3		
Rainy	3	2	Cool	3	1								
Sunny	2/9	3/5	Hot	2/9	2/5	High	3/9	4/5	False	6/9	2/5	9/14	5/14
Overcast	4/9	0/5	Mild	4/9	2/5	Normal	6/9	1/5	True	3/9	3/5		
Rainy	3/9	2/5	Cool	3/9	1/5								

Tabela **frekvenc** (zgornji del) in **verjetnosti** (spodnji del)

Outlook	Temp	Humidity	Windy	Play
Sunny	Hot	High	False	No
Sunny	Hot	High	True	No
Overcast	Hot	High	False	Yes
Rainy	Mild	High	False	Yes
Rainy	Cool	Normal	False	Yes
Rainy	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Sunny	Mild	High	False	No
Sunny	Cool	Normal	False	Yes
Rainy	Mild	Normal	False	Yes
Sunny	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Rainy	Mild	High	True	No

Primer: "weather" podatki

Outlook			Temperature			Humidity			Windy			Play	
Yes No			Yes No			Yes No			Yes No			Yes	No
Sunny	2	3	Hot	2	2	High	3	4	False	6	2	9	5
Overcast	4	0	Mild	4	2	Normal	6	1	True	3	3		
Rainy	3	2	Cool	3	1								
Sunny	2/9	3/5	Hot	2/9	2/5	High	3/9	4/5	False	6/9	2/5	9/14	5/14
Overcast	4/9	0/5	Mild	4/9	2/5	Normal	6/9	1/5	True	3/9	3/5		
Rainy	3/9	2/5	Cool	3/9	1/5								

Napovedovanje
"novega" dneva:

Outlook	Temp.	Humidity	Windy	Play
Sunny	Cool	High	True	?

Verjetje (likelihood) vrednosti razreda:

$$\text{Za "yes"} = 2/9 \times 3/9 \times 3/9 \times 3/9 \times 9/14 = 0,0053$$

$$\text{Za "no"} = 3/5 \times 1/5 \times 4/5 \times 3/5 \times 5/14 = 0,0206$$

Pretvorba v verjetnosti z uporabo normalizacije:

$$P(\text{"yes"}) = 0,0053 / (0,0053 + 0,0206) = 0,205$$

$$P(\text{"no"}) = 0,0206 / (0,0053 + 0,0206) = 0,795$$

Bayes-ovo pravilo

- Verjetnost dohodka H pri poznanih dejstvih E :

$$\Pr[H | E] = \frac{\Pr[E | H] \Pr[H]}{\Pr[E]}$$

- ***A priori*** verjetnost hipoteze H : $\Pr[H]$
 - Verjetnost dogodka H *preden* poznamo dejstva E
- ***A posteriori*** verjetnost hipoteze H : $\Pr[H | E]$
 - Verjetnost dogodka H *po tem*, ko že poznamo dejstva E

povzeto po: Bayes “Essay towards solving a problem in the doctrine of chances” (1763)

Thomas Bayes

Rojen: 1702 v Londonu, Anglija

Premiril: 1761 v Tunbridge Wellsu, Kent, Anglija



Klasifikacija z naivnim Bayes-om

- Klasifikacijsko učenje:
kakšna je verjetnost razreda pri poznanih verjetnostih vrednosti ostalih atributov primera?
 - Dejstva E = vrednosti atributov primera
 - Hipoteza H = vrednost razreda pri podanih vrednostih atributov
- Naivna predpostavka:
dejstva razdelimo na dele (attribute), ki so *neodvisni*

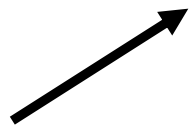
$$\Pr[H | E] = \frac{\Pr[E_1 | H] \Pr[E_2 | H] \dots \Pr[E_n | H] \Pr[H]}{\Pr[E]}$$

Primer "weather"

Outlook	Temp.	Humidity	Windy	Play
Sunny	Cool	High	True	?

← *Dejstva E*

*Verjetnost
razreda "yes"*



$$\begin{aligned}\Pr[\text{yes} \mid E] &= \Pr[\text{Outlook} = \text{Sunny} \mid \text{yes}] \\ &\quad \times \Pr[\text{Temperature} = \text{Cool} \mid \text{yes}] \\ &\quad \times \Pr[\text{Humidity} = \text{High} \mid \text{yes}] \\ &\quad \times \Pr[\text{Windy} = \text{True} \mid \text{yes}] \\ &\quad \times \frac{\Pr[\text{yes}]}{\Pr[E]} \\ &= \frac{\frac{2}{9} \times \frac{3}{9} \times \frac{3}{9} \times \frac{3}{9} \times \frac{9}{14}}{\Pr[E]}\end{aligned}$$

Problem "frekvenca = nič"

- Kaj pa, če se določena vrednost atributa ne pojavi pri vseh vrednostih razreda?
(npr. "Humidity = high" za razred "yes")
 - Verjetnost bo tu enaka nič! $\Pr[Humidity = High \mid yes] = 0$
 - *A posteriori* verjetnost bo prav tako nič!
(ne glede na verjetnosti ostalih vrednosti!) $\Pr[yes \mid E] = 0$
- Rešitev: dodamo 1 vsem frekvencam v tabeli frekvenc
(*Laplace-ova ocena*)
- Rezultat: verjetnosti ne bodo nikoli nič!
(še več: stabilizira ocene verjetnosti)

*Popravljene ocene verjetnosti

- Včasih je primerneje dodati kako drugo konstanto, ki je različna od 1
- Primer: atribut *outlook* za razred *yes*

$$\frac{2 + \mu/3}{9 + \mu}$$

Sunny

$$\frac{4 + \mu/3}{9 + \mu}$$

Overcast

$$\frac{3 + \mu/3}{9 + \mu}$$

Rainy

- Ni nujno, da so uteži enake (njihova vsota pa mora biti vedno 1)

$$\frac{2 + \mu p_1}{9 + \mu}$$

$$\frac{4 + \mu p_2}{9 + \mu}$$

$$\frac{3 + \mu p_3}{9 + \mu}$$

Manjkajoče vrednosti

- Pri učenju:
manjkajočih vrednosti ne upoštevamo pri izračunu frekvenc v tabeli frekvenc
- Pri klasifikaciji:
atributa ne upoštevamo pri izračunu verjetij
- Primer:

Outlook	Temp.	Humidity	Windy	Play
?	Cool	High	True	?

Verjetje za razred "yes" = $3/9 \times 3/9 \times 3/9 \times 9/14 = 0,0238$

Verjetje za razred "no" = $1/5 \times 4/5 \times 3/5 \times 5/14 = 0,0343$

$P(\text{"yes"}) = 0,0238 / (0,0238 + 0,0343) = 41\%$

$P(\text{"no"}) = 0,0343 / (0,0238 + 0,0343) = 59\%$

Numerični atributi

- Običajna predpostavka: atributi imajo *normalno* ali *Gaussi-ovo* verjetnostno porazdelitev (glede na razred)
- Funkcija gostote verjetnosti za normalno porazdelitev je definirana z dvema parametroma:

- *Povprečje vzorca* μ

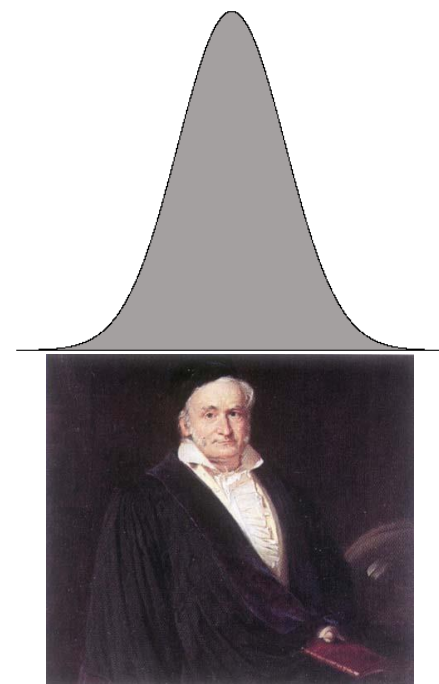
$$\mu = \frac{1}{n} \sum_{i=1}^n x_i$$

- *Standardna deviacija* σ

$$\sigma = \frac{1}{n-1} \sum_{i=1}^n (x_i - \mu)^2$$

- Funkcija gostote verjetnosti $f(x)$ je:

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



Karl Gauss, 1777-1855
znani nemški matematik

Primer "weather-numeric"

Outlook			Temperature		Humidity		Windy			Play	
<i>Yes No</i>			<i>Yes No</i>		<i>Yes No</i>		<i>Yes No</i>			<i>Yes No</i>	
Sunny	2	3	64, 68,	65, 71,	65, 70,	70, 85,	False	6	2	9	5
Overcast	4	0	69, 70,	72, 80,	70, 75,	90, 91,	True	3	3		
Rainy	3	2	72, ...	85, ...	80, ...	95, ...					
Sunny	2/9	3/5	$\mu = 73$	$\mu = 75$	$\mu = 79$	$\mu = 86$	False	6/9	2/5	9/14	5/14
Overcast	4/9	0/5	$\sigma = 6,2$	$\sigma = 7,9$	$\sigma = 10,2$	$\sigma = 9,7$	True	3/9	3/5		
Rainy	3/9	2/5									

- Primer izračuna gostote verjetnosti:

$$f(\text{Temperature} = 66 \mid \text{yes}) = \frac{1}{\sqrt{2\pi}(6,2)} e^{-\frac{(66-73)^2}{2*(6,2)^2}}$$

Klasifikacija "novega" dneva

- "Nov" dan:

Outlook	Temp.	Humidity	Windy	Play
Sunny	66	90	true	?

Verjetje za razred "yes" = $2/9 \times 0,0340 \times 0,0221 \times 3/9 \times 9/14 = 0,000036$

Verjetje za razred "no" = $3/5 \times 0,0291 \times 0,0380 \times 3/5 \times 5/14 = 0,000136$

$P(\text{"yes"}) = 0,000036 / (0,000036 + 0,000136) = 20,9\%$

$P(\text{"no"}) = 0,000136 / (0,000036 + 0,000136) = 79,1\%$

- Manjkajoče vrednosti v fazi učenja niso vključene v izračun povprečij in standardnih deviacij

Naivni Bayes: diskusija

- Naivni Bayes deluje presenetljivo dobro (čeprav vemo, da "neodvisnosti" atributov ni zadoščeno)
- Zakaj? Ker klasifikacija ne zahteva natančnih ocen verjetnosti *dokler največja verjetnost pripada "pravemu" razredu*
- Čeprav: dodajanje prevelikega števila redundantnih atributov lahko povzroči težave (npr. popolnoma korelirani atributi)
- Pozor: številni numerični atributi niso normalno porazdeljeni (→ *uporaba drugih verjetnostnih porazdelitev, t.i. **jedrne funkcije***)

Razširitve naivnega Bayes-a

- Izboljšave:
 - izbor "najboljših" atributov (npr. s požrešnim iskanjem);
 - pogosto deluje enako dobro, če ne boljše na podmnožici vseh atributov;
- Bayes-ove mreže

Povzetek

- **ZeroR** – ni pravi klasifikator, napove večinski razred
- **OneR** – temelji na pravilih in upošteva en sam atribut
- **Naivni Bayes** – vključuje vse attribute, uporablja Bayes-ovo pravilo za oceno verjetnosti razreda pri podanih vrednostih primera.
- Preproste metode pogosto dobro delujejo, ampak ...
 - ... zahtevnejše metode so lahko boljše (kot bomo videli kasneje)