

Klasifikacija – Naivni Bajes

NIKOLA MILIKIĆ

EMAIL: nikola.milicic@fon.bg.ac.rs

URL: <http://nikola.milicic.info>

Bajesovo pravilo

$$P(H|E) = \frac{P(E|H) * P(H)}{P(E)}$$

- H – hipoteza (hypothesis)
- E – opažaj (evidence) vezan za hipotezu H, tj. podaci na osnovu kojih bi trebalo da potvrdimo ili odbacimo hipotezu H
- P(H) – verovatnoća hipoteze H (prior probability)
- P(E) – verovatnoća opažaja tj stanja na koje ukazuju prikupljeni podaci
- P(E | H) – (uslovna) verovatnoća opažaja E ukoliko važi hipoteza H
- P(H | E) – (uslovna) verovatnoća hipoteze H ukoliko imamo opažaj E

Bajesovo pravilo - primer

Pretpostavite sledeće:

- jednog jutra ste se probudili sa povišenom temperaturom
- prethodnog dana ste čuli da je u gradu počela da se širi virusna infekcija, ali da je verovatnoća zaraze mala, svega 2.5%
- takođe ste čuli da je u 50% slučajeva virusna infekcija praćena povišenom temperaturom
- u vašem slučaju, povišena temperatura se javlja svega par puta u godini, tako možemo reći da je verovatnoća da imate povišenu temp. 5%

Pitanje: kolika je verovatnoća da, pošto imate povišenu temp., da imate i virusnu infekciju?

Bajesovo pravilo - primer

Teorija	Primer
Hipoteza (H)	Imate virusnu infekciju
$P(H)$	0.025
Opažaj (evidence - E)	Imate povišenu temperaturu
$P(E)$	0.05
(uslovna) verovatnoća opažaja E ukoliko važi hipoteza H: $P(E H)$	Verovatnoća da je virusna infekcija praćena povišenom temperaturom 0.50
(uslovna) verovatnoća hipoteze H ukoliko imamo opažaj E: $P(H E)$	Verovatnoća da pošto imate povišenu temp., da imate i virusnu infekciju ?

$$P(H|E) = P(E|H) * P(H) / P(E)$$

$$P(H|E) = 0.50 * 0.025 / 0.05 = 0.25$$

Naivni Bajes klasifikator

- Klasifikator zasnovan na Bajesovom pravilu
- Uvodi dve “naivne” pretpostavke nad atributima:
 - svi atributi su a priori podjednako važni
 - svi atributi su statistički nezavisni (vrednost jednog atributa nam ne govori ništa o vrednosti drugog atributa)
- Ove pretpostavke najčešće nisu tačne, ali i pored toga, u praksi ovaj klasifikator daje dobre rezultate

Naivni Bajes klasifikator

Pretpostavka nezavisnosti atributa značajno pojednostavljuje računanje uslovnih verovatnoća

$$\begin{aligned} P(E|H) &= P(E_1, E_2, \dots, E_n | H) \\ &= P(E_1 | H) * P(E_2 | H) * \dots * P(E_n | H) \end{aligned}$$

E_i – opažaj (raspoloživi podaci) vezan za atribut i

$$P(H|E) = \frac{P(E_1 | H) * P(E_2 | H) * \dots * P(E_n | H) * P(H)}{P(E)}$$

Primer – Predviđanje da li će se predstava odigrati

ToPlayOtNotToPlay.arff

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

Kako sunčano vreme (outlook=sunny) utiče na ishod?

Pretpostavimo da znamo da je napolju sunčano

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		

U tom slučaju, postoji 60% šanse za Play = no, odnosno $P(\text{Play=no} \mid \text{outlook=sunny})=0.6$

Kako (ostale) vrednosti atributa Outlook utiču na ishod?

Outlook	Play	
	yes	no
sunny	2	3
overcast	4	0
rainy	3	2
TOTAL	9	5



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

Za svaku vrednost atributa, odrediti učestanost pojavljivanja za svaki od ishoda (yes/no) ...

To je potrebno za računanje uslovnih verovatnoća ($P(E_i|H)$)

Kako vrednosti svih atributa utiču na ishod?

Outlook	Play	
	yes	no
sunny	2	3
overcast	4	0
rainy	3	2
TOTAL	9	5

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



Ponoviti isti postupak za sve ostale attribute

Outlook	Play		Temp.	Play		Humid.	Play		Windy	Play		TOTAL	Play
	yes	no		yes	no		yes	no		yes	no		
sunny	2	3	hot	2	2	high	3	4	false	6	2	yes	9
overcast	4	0	mild	4	2	normal	6	1	true	3	3	no	5
rainy	3	2	cool	3	1								
TOTAL	9	5	TOTAL	9	5	TOTAL	9	5	TOTAL	9	5	TOTAL	14

Pretvaramo učestanost pojavljanja u procene verovatnoća

Outlook	Play		Temp.	Play		Humid.	Play		Windy	Play		Play	
	yes	no		yes	no		yes	no		yes	no		
sunny	2	3	hot	2	2	high	3	4	false	6	2	yes	9
overcast	4	0	mild	4	2	normal	6	1	true	3	3	no	5
rainy	3	2	cool	3	1								
TOTAL	9	5	TOTAL	9	5	TOTAL	9	5	TOTAL	9	5	TOTAL	14

Pretvoriti učestanost pojavljanja u procene verovatnoća ($P(E_i | H)$ i $P(H)$)

Outlook	Play		Temp.	Play		Humid.	Play		Windy	Play		Play	
	yes	no		yes	no		yes	no		yes	no		
sunny	0.22	0.60	hot	0.22	0.40	high	0.33	0.80	false	0.67	0.40	yes	0.64
overcast	0.44	0.00	mild	0.44	0.40	normal	0.67	0.20	true	0.33	0.60	no	0.36
rainy	0.33	0.40	cool	0.33	0.20								

2 pojavljanja **Play = no**, gde je **Outlook = rainy**
5 pojavljanja **Play = no**

Verovatnoće ishoda za date vremenske uslove

Za date vremenske uslove (tj vrednosti atributa):

Outlook = sunny

Temperature = cool

Humidity = high

Windy = true

Odrediti verovatnoću oba moguća ishoda Play varijable (Yes, No)

$$P(H|E) = \frac{P(E|H) * P(H)}{P(E)}$$

$$P(E) = L(\text{Play} = \text{true}|E) + L(\text{Play} = \text{false}|E)$$

$$P(\text{Play} = \text{yes}): 0.0053 / (0.0053 + 0.0206) = \mathbf{20.5\%}$$

$$P(\text{Play} = \text{no}): 0.0206 / (0.0053 + 0.0206) = \mathbf{79.5\%}$$

Primena Laplace estimator-a

Učestanosti pojavljivanja vrednosti atributa iz originalnog dataset-a

Outlook	Play		Temp.	Play		Humid.	Play		Windy	Play		Play	
	yes	no		yes	no		yes	no		yes	no		
sunny	2	3	hot	2	2	high	3	4	false	6	2	yes	9
overcast	4	0	mild	4	2	normal	6	1	true	3	3	no	5
rainy	3	1	cool	3	1								
TOTAL	9	5	TOTAL	9	5	TOTAL	9	5	TOTAL	9	5	TOTAL	14

Laplace estimator:
uvećati sve učestanosti za 1

Nakon dodavanja 1 svakom broju (Laplace estimator)

Outlook	Play		Temp.	Play		Humid.	Play		Windy	Play		Play	
	yes	no		yes	no		yes	no		yes	no		
sunny	3	4	hot	3	3	high	4	5	false	7	3	yes	12
overcast	5	1	mild	5	3	normal	7	2	true	4	4	no	8
rainy	4	3	cool	4	2								
TOTAL	12	8	TOTAL	12	8	TOTAL	11	7	TOTAL	11	7	TOTAL	20

Verovatnoće ishoda za date vremenske uslove pri korišćenju Laplace estimatora

Outlook	Play		Temp.	Play		Humid.	Play		Windy	Play		Play	
	yes	no		yes	no		yes	no		yes	no		
sunny	0.25	0.50	hot	0.25	0.38	high	0.36	0.71	false	0.64	0.43	yes	0.64
overcast	0.42	0.13	mild	0.42	0.38	normal	0.64	0.29	true	0.36	0.57	no	0.36
rainy	0.33	0.38	cool	0.33	0.25								

Vremenski uslovi:

Outlook = overcast, Temperature = cool, Humidity = high, Windy = true

Bez korišćenja Laplace estimator-a:

Play = no: 79.5%

Play = yes: 20.5%

Korišćenjem Laplace estimator-a:

Play = no: 72.0%

Play = yes: 28.0%

Efekat Laplace estimator-a opada povećanjem dataset-a

Predikciona pravila – verovatnoće svih kombinacija

	Play			Play			Play			Play			Play
Outlook	yes	no	Temp.	yes	no	Humid.	yes	no	Windy	yes	no		
sunny	0.25	0.50	hot	0.25	0.38	high	0.36	0.71	false	0.64	0.43	yes	0.64
overcast	0.42	0.13	mild	0.42	0.38	normal	0.64	0.29	true	0.36	0.57	no	0.36
rainy	0.33	0.38	cool	0.33	0.25								



Inst	Outlook	Temp.	Humid.	Windy	Play	Outlook	Temp.	Humid.	Windy	Play	Like.	Prob.
	overcast	cool	high	false	no	0.13	0.25	0.71	0.43	0.36	0.0034	14.2%
	overcast	cool	high	false	yes	0.42	0.33	0.36	0.64	0.64	0.0207	85.8%
	overcast	cool	high	true	no	0.13	0.25	0.71	0.57	0.36	0.0046	27.8%
	overcast	cool	high	true	yes	0.42	0.33	0.36	0.64	0.64	0.0118	72.2%
	overcast	cool	normal	false	yes	0.42	0.33	0.64	0.64	0.64	0.0362	96.4%
	overcast	cool	normal	true	no	0.13	0.25	0.29	0.57	0.36	0.0018	8.1%
7	overcast	cool	normal	true	yes	0.42	0.33	0.64	0.36	0.64	0.0207	91.9%
	overcast	hot	high	false	no	0.13	0.38	0.71	0.43	0.36	0.0051	24.9%
3	overcast	hot	high	false	yes	0.42	0.25	0.36	0.64	0.64	0.0155	75.1%

Izračunati verovatnoće za sve kombinacije vremenskih uslova

Predikciona pravila – verovatnoće svih kombinacija

Inst	Outlook	Temp.	Humid.	Windy	Play	Prob.
	overcast	cool	normal	false	yes	96.4%
	overcast	mild	normal	false	yes	95.7%
13	overcast	hot	normal	false	yes	93.0%
7	overcast	cool	normal	true	yes	91.9%
	overcast	mild	normal	true	yes	90.4%
5	rainy	cool	normal	false	yes	87.6%
	overcast	cool	high	false	yes	85.8%
10	rainy	mild	normal	false	yes	85.5%
	overcast	hot	normal	true	yes	85.0%
2	sunny	hot	high	true	no	83.7%
	overcast	mild	high	false	yes	83.4%
9	sunny	cool	normal	false	yes	79.9%
	rainy	hot	normal	false	yes	77.9%
	sunny	mild	normal	false	yes	76.8%
	sunny	mild	high	true	no	75.5%
3	overcast	hot	high	false	yes	75.1%
	rainy	cool	normal	true	yes	75.1%
	rainy	hot	high	true	no	74.3%

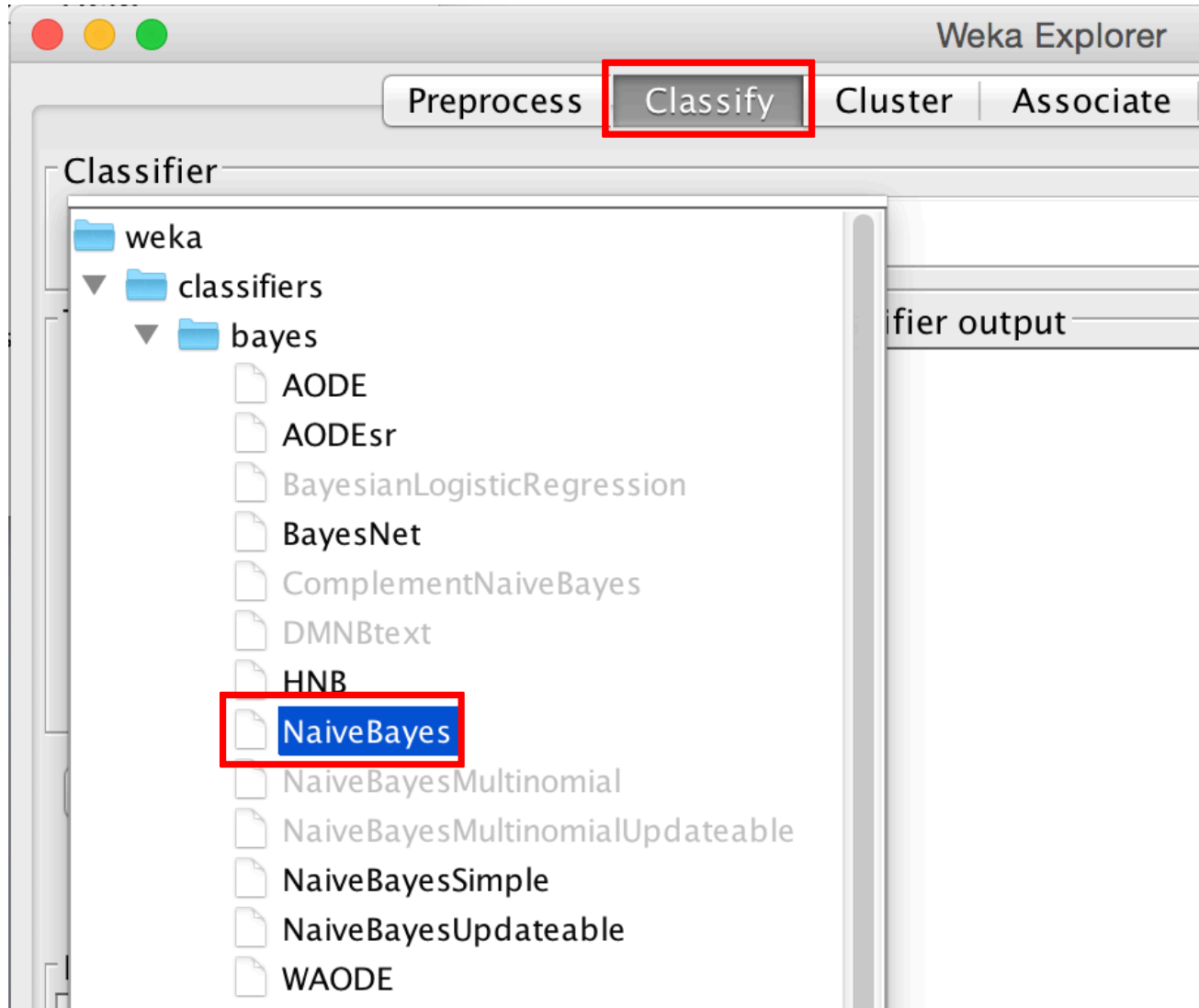
Inst	Outlook	Temp.	Humid.	Windy	Play	Prob.
	overcast	cool	high	true	yes	72.2%
	sunny	cool	high	true	no	72.0%
	rainy	mild	normal	true	yes	71.6%
1	sunny	hot	high	false	no	68.8%
12	overcast	mild	high	true	yes	68.4%
	sunny	hot	normal	false	yes	66.5%
14	rainy	mild	high	true	no	63.5%
	sunny	cool	normal	true	yes	63.0%
	rainy	cool	high	false	yes	61.7%
	rainy	hot	normal	true	yes	60.2%
	rainy	cool	high	true	no	59.1%
11	sunny	mild	normal	true	yes	58.6%
4	rainy	mild	high	false	yes	57.3%
8	sunny	mild	high	false	no	57.0%
	overcast	hot	high	true	yes	56.4%
	rainy	hot	high	false	no	55.4%
	sunny	hot	normal	true	no	54.0%
	sunny	cool	high	false	no	52.4%

Pravila koja predviđaju klasu za sve kombinacije vrednosti atributa

Neće sve predikcije biti korektne

Inst	Outlook	Temp.	Humid.	Windy	Play	Prob.	Actual
1	sunny	hot	high	false	no	72.6%	no
2	sunny	hot	high	true	no	86.1%	no
3	overcast	hot	high	false	yes	71.6%	yes
4	rainy	mild	high	false	yes	52.8%	yes
5	rainy	cool	normal	false	yes	85.5%	yes
6	rainy	cool	normal	true	yes	75.1%	no
7	overcast	cool	normal	true	yes	90.4%	yes
8	sunny	mild	high	false	no	61.4%	no
9	sunny	cool	normal	false	yes	76.8%	yes
10	rainy	mild	normal	false	yes	83.0%	yes
11	sunny	mild	normal	true	yes	54.2%	yes
12	overcast	mild	high	true	yes	64.3%	yes
13	overcast	hot	normal	false	yes	91.7%	yes
14	rainy	mild	high	true	no	67.6%	no

Naivni Bajes u Weka-i



Predikcija na test dataset-u

ToPlayOtNotToPlay.arff dataset

The image shows two windows from the Weka software interface. The left window, titled "Test options", has three radio buttons: "Use training set", "Supplied test set" (which is selected and highlighted with a red box), and "Cross-validation". The "Supplied test set" option has a "Set..." button next to it. Below these are "Percentage split" (set to 66%) and a "More options..." button. At the bottom are "Start" and "Stop" buttons. The right window, titled "Classifier evaluation options", has several checked checkboxes: "Output model", "Output per-class stats", "Output confusion matrix", and "Store predictions for visualization". There are also unchecked checkboxes for "Output entropy evaluation measures" and "Error plot point size proportional to margin". At the bottom, there is a "Choose" button and a text field containing "PlainText -distribution".

Test options

- Use training set
- Supplied test set **Set...**
- Cross-validation Folds 10
- Percentage split % 66
- More options...

(Nom) Play

Start Stop

Classifier evaluation options

- Output model
- Output per-class stats
- Output entropy evaluation measures
- Output confusion matrix
- Store predictions for visualization
- Error plot point size proportional to margin

Output predictions **Choose** PlainText -distribution

Prikaz rezultata klasifikacije

Outlook	Play		Temp.	Play		Humid.	Play		Windy	Play		TOTAL	
	yes	no		yes	no		yes	no		yes	no		
sunny	3	4	hot	3	3	high	4	5	false	7	3	yes	12
overcast	5	1	mild	5	3	normal	7	2					
rainy	4	3	cool	4	2								
TOTAL	12	8	TOTAL	12	8	TOTAL	11	7					

Classifier output

Attribute	no (0.38)	yes (0.63)

Outlook		
sunny	4.0	3.0
overcast	1.0	5.0
rainy	3.0	4.0
[total]	8.0	12.0
Temp.		
hot	3.0	3.0
mild	3.0	5.0
cool	2.0	4.0
[total]	8.0	12.0
Humidity		
high	5.0	4.0
normal	2.0	7.0
[total]	7.0	11.0

Klasifikator automatski primenjuje Laplace estimator

Predikcije na test dataset-u

Classifier output

=== Predictions on test set ===

inst#	actual	predicted	error	distribution
1	2:no	1:yes	+	*0.897,0.103
2	1:yes	1:yes		*0.706,0.294
3	2:no	2:no		0.39,*0.61
4	2:no	2:no		0.348,*0.652
5	2:no	2:no		0.348,*0.652
6	2:no	2:no		0.153,*0.847
7	1:yes	1:yes		*0.754,0.246

Instanca 1 je obeležena kao pogrešno klasifikovana

Verovatnoće ishoda Play varijable za svaku instancu u datasetu

Osobine Naivnog Bajesa

- Namenjen primarno za rad sa nominalnim atributima
- U slučaju numeričkih atributa:
 - koristiti raspodelu verovatnoća atributa (tipično Normalna raspodela) za procenu verovatnoće vrednosti atributa u dataset-u
 - uraditi diskretizaciju vrednosti atributa

Baseline klasifikator



Baseline klasifikator

Dataset: diabetes.arff

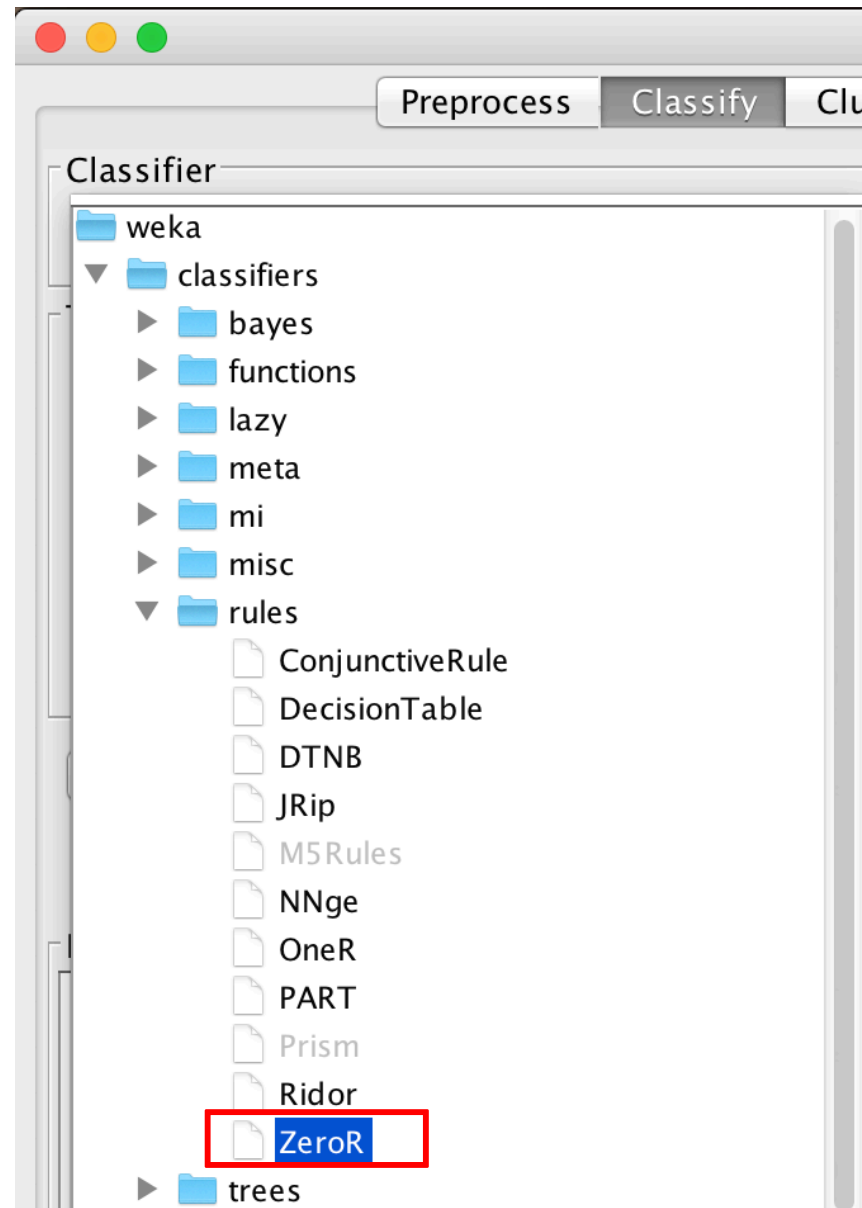
- Imamo ukupno 768 instanci: 500 negativnih i 268 pozitivnih
- A priori verovatnoća da će biti negativan, odnosno pozitivan ishod je

$$\textit{Negativan} = \frac{500}{768} = 0.651$$

$$\textit{Pozitivan} = \frac{268}{768} = 0.349$$

- Baseline klasifikator klasifikuje svaku instancu u klasu sa najvećom a priori verovatnoćom
- U Weka-i implementacija osnovnog klasifikatora: **rules -> ZeroR**

Klasifikator u Weka-i: rules -> ZeroR



Baseline klasifikator

- Otvoriti dataset **diabetes.arff**
- Test option: Percentage split 66%
- Probati klasifikatore:
 - **rules -> ZeroR** **65%**
 - **trees -> J48** **76%**
 - **bayes -> NaiveBayes** **77%**
 - **lazy -> Ibk** **73%**
- Za svaki problem klasifikacije proveriti da li odabrani algoritam daje bolje rezultate od osnovnog klasifikatora

Primer 2 – Skup podataka “Jestive pečurke”

EdibleMushrooms.arff

- Skup podataka “Jestive pečurke” je nastao na osnovu knjige “National Audubon Society Field Guide to North American Mushrooms”
- Skup podataka obuhvata opise hipotetičkih uzoraka pečuraka koje pripadaju jednoj od 23 vrste pečuraka
- Postoji ukupno 8124 instanci sa 22 nominalna atributa koji opisuju karakteristike pečuraka; klasifikacioni atribut je onaj koji ukazuje da li je pečurka jestiva ili ne
- Cilj je da predvidimo da li je nepoznata pečurka jestiva ili ne

Podaci u datasetu su fiktivni i rezultate ne treba koristiti prilikom stvarnog određivanja jestivosti pečuraka!

Primer 3 – Skup podataka “Supermarket”

supermarket.arff

- Dataset opisuje podatke o prodaji artikala u malom novozelandskom supermarketu u jednom danu.
- Podaci su nominalni i odnose se na različita odeljenja (department) u prodavnici i različite grupe artikala (npr. “bread and cake” se odnosi na grupu pekarskih proizvoda).
- Ako atribut ima vrednost “t”, to znači da je korpa kupca sadržala makar jedan proizvod iz datog odeljenja, odnosno da je kupac kupio makar jedan proizvod iz navedene kategorije proizvoda.
- Klasa ima vrednost “low” ili “high” u zavisnosti od toga da li je kupovina bila manja ili veća od 100\$

Preporuke i zahvalnice

"Data Mining with Weka" and "More Data Mining with Weka": MOOCs from the University of Waikato. A self-paced session of "Data Mining with Weka" runs until 23 October June 2015.

- Link: <https://www.youtube.com/user/WekaMOOC/>

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

<http://goo.gl/cqdp3l>

Pitanja?

NIKOLA MILIKIĆ

EMAIL: nikola.milicic@fon.bg.ac.rs

URL: <http://nikola.milicic.info>