NAÏVE BAYES KLASIFIKATOR

JELENA JOVANOVIĆ

Email: jelena.jovanovic@fon.bg.ac.rs

Web: http://jelenajovanovic.net

ZAŠTO BAŠ NAÏVE BAYES?

Naïve Bayes (NB) se navodi kao algoritam koji bi trebalo među prvima razmotriti pri rešavanju zadataka klasifikacije

Razlozi:

- Jednostavan je*
- Ima dobre performanse
- Vrlo je skalabilan
- Može se prilagoditi za gotovo bilo koji problem klasifikacije

*Occam's Razor: "Other things being equal, simple theories are preferable to complex ones"

PODSEĆANJE: BAYES-OVO 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

BAYES-OVO 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?

BAYES-OVO 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 ?

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

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

NB KLASIFIKATOR

Ako je ***c*** klasa, a ***o*** objekat, prema Bayes-ovom pravilu, verovatnoća da je dati objekat ***o*** klase ***c*** je:

$$P(c|o) = P(o|c) * P(c) / P(o) \qquad (1)$$

NB KLASIFIKATOR

Zadatak: za dati skup klasa ***C*** i objekat ***o***, potrebno je pronaći onu klasu ***c*** iz skupa ***C*** koja ima najveću uslovnu verovatnoću za objekat ***o***; formalno iskazano:

$$f = argmax_{c \in C} P(c|o) \qquad (2)$$

Primenom Bayes-ovog pravila, dobijamo:

$$f = argmax_{c \in C} P(o|c) * P(c) \qquad (3)$$

NB KLASIFIKATOR

Potrebno je odrediti verovatnoće P(c) i P(o|c)

P(c) se može *proceniti* relativno jednostavno: brojanjem pojavljivanja klase ***c*** u skupu za trening ***O***

P(o|c) - verovatnoća da u klasi ***c*** ‘zateknemo’ objekat ***o*** – nije tako jednostavno odrediti i tu uvodimo pretpostavku koja NB algoritam čine “naivnim”

NB KLASIFIKATOR

Kako odrediti P(o|c)?

- objekat ***o*** predstavljamo kao skup atributa (features) koji ga opisuju (x_1, x_2, ...,x_n)
- umesto P(o|c) imaćemo P(x_1, x_2, x_3, ...x_n|c)
- da bi izračunali P(x_1, x_2, x_3, ...x_n|c) uvodimo naivnu pretpostavku:
  - atributi koji opisuju objekat ***o*** su međusobno nezavisni tj. objekat ***o*** možemo posmatrati kao prost skup atributa

NB KLASIFIKATOR

Uvedena pretpostavka

- (✖) nije uvek validna
- (✖) može dovesti do značajnog gubitka informacija koje iz podataka možemo da izvučemo
- (✔) omogućuju značajno jednostavnije računanje $P(x_1, x_2,...,x_n|c)$, a time i ceo zadatak klasifikacije

NB KLASIFIKATOR

Na osnovu uvedenih pretpostavki, $P(x_1, x_2,...,x_n|c)$ možemo da predstavimo kao proizvod individualnih uslovnih verovatnoća

$$P(x_1, x_2,...,x_n|c) = P(x_1|c) * P(x_2|c) * \dots * P(x_n|c)$$

Time dolazimo do opšte jednačine NB algoritma:

$$f = argmax_{c \in C} P(c) * \prod_{i=1}^{n} P(x_i|c)$$

NB KLASIFIKATOR

Procena verovatnoća se vrši na osnovu skupa za trening, i zasniva na sledećem:

$P(c)$ = br. instanci klase ***c*** / ukupan br. instanci u skupu za trening

$P(x_i|c)$ se određuje na osnovu raspodele atributa x_i u objektima klase ***c***; sam proračun zavisi od tipa atributa (nominalni, numerički)

PRIMER RAČUNANJA VEROVATNOĆA ZA NOMINALNE ATRIBUTE

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

Likelihood Table				
Outlook	Yes	No		
Sunny	3	2	=(5/14)	0.36
Rainy	2	3	=(5/14)	0.36
Overcast	4	0	=(4/14)	0.28
	=(9/14)	=(5/14)		
	0.64	0.36		

P (Sunny | Yes) = 3/9 = 0.33

P (Sunny) = 5/14 = 0.36

P (Yes) = 9/14 = 0.64

Izvori slika:
https://www.saedsayad.com/naive_bayesian.htm
https://morioh.com/p/6d50f0574db6

Računanje verovatnoća u slučaju numeričkih atributa

U slučaju numeričkih atributa:

- ako je poznata, koristi se raspodela vrednosti atributa za procenu verovatnoće mogućih vrednosti atributa
  - tipična primena u slučaju atributa koji imaju normalnu raspodelu ili se mogu transformisati tako da budu približno normalne rasp.
- U suprotnom, uraditi diskretizaciju vrednosti atributa

PRIMER RAČUNANJA VEROVATNOĆA ZA NUMERIČKE ATRIBUTE SA NORM. RASPODELOM

		Humidity	*Mean*	*StDev*
Play	yes	86 96 80 65 70 80 70 90 75	79.1	10.2
Golf	no	85 90 70 95 91	86.2	9.7

$$P(\text{humidity} = 74 \mid \text{play} = \text{yes}) = \frac{1}{\sqrt{2\pi}(10.2)} e^{-\frac{(74-79.1)^2}{2(10.2)^2}} = 0.0344$$

$$P(\text{humidity} = 74 \mid \text{play} = \text{no}) = \frac{1}{\sqrt{2\pi}(9.7)} e^{-\frac{(74-86.2)^2}{2(9.7)^2}} = 0.0187$$

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

$$\sigma = \left[\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \mu)^2\right]^{0.5}$$ Standard deviation

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

Izvor: https://www.saedsayad.com/naive_bayesian.htm

Osobine NB klasifikatora

- Brz i najčešće daje dobre rezultate - često se pokazuje kao podjednako dobar ili čak i bolje neko sofisticiraniji modeli
- Skalabilan i po pitanju broja observacija i po pitanju broja atributa
- Nije memorijski zahtevan
- Ima vrlo mali afinitet ka preteranom podudaranju sa podacima za trening (overfitting)
- Jednostavan za interpretaciju

Osobine NB klasifikatora

- Pretpostavka nezavisnosti atributa: često nerealna i može dovesti do slabijih performansi, posebno u slučaju visokokorelisanih atributa
- Problem nulte frekvencije: problem koji se javlja u slučaju kada određene vrednosti atributa nisu prisutne u trening setu, a mogu se pojaviti u test setu, odnosno primeni modela