

NAÏVE BAYES KLASIFIKATOR

JELENA JOVANOVIĆ

Email: jeljov@gmail.com

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

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

- 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. 6.5%

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
$P(E)$	0.065
(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 ?

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

$$P(H|E) = 0.50 * 0.025 / 0.065 = 0.19$$

NB KLASIFIKATOR

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

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

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:

$$f = \operatorname{argmax}_{c \text{ iz } C} P(c|o) \quad (2)$$

Primenom Bayes-ovog pravila, dobijamo:

$$f = \operatorname{argmax}_{c \text{ iz } C} P(o|c) * P(c) \quad (3)$$

NB KLASIFIKATOR

$$f = \operatorname{argmax}_{c \text{ iz } C} P(o|c) * P(c) \quad (3)$$

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, \dots, x_n)
- umesto $P(o|c)$ imaćemo $P(x_1, x_2, x_3, \dots, x_n|c)$
- da bi izračunali $P(x_1, x_2, x_3, \dots, x_n|c)$ 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 $P(x_1, x_2, \dots, x_n | c)$, a time i ceo problem klasifikacije

NB KLASIFIKATOR

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

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

Time dolazimo do opšte jednačine NB algoritma:

$$f = \operatorname{argmax}_c \text{ iz } c \ P(c) * \prod_{i=1, n} P(x_i | c)$$

NB KLASIFIKATOR

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

$P(c)$ = 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 (nominal, numerički)

OSObine NB KLASIFIKATORA

- Veoma brz i efikasan
- Najčešće daje dobre rezultate
 - često se pokazuje kao bolji ili bar podjednako dobar kao drugi, sofisticiraniji modeli
- Nije memorijski zahtevan
- Ima vrlo mali afinitet ka preteranom podudaranju sa podacima za trening (overfitting)
- Pogodan i kada imamo malu količinu podataka za trening

OSOBI NE NB KLASIFIKATORA

- “Otporan” na nevažne atribute
 - atributi koji su podjednako distribuirani kroz skup podataka za trening, pa nemaju veći uticaj na izbor klase
- Namenjen primarno za rad sa nominalnim atributima; u slučaju numeričkih atributa:
 - koristiti raspodelu verovatnoća atributa (tipično Normalna raspodela) za procenu verovatnoće svake od mogućih vrednosti atributa
 - uraditi diskretizaciju vrednosti atributa