Obrada prirodnih jezika

Elektrotehnički fakultet Univerziteta u Beogradu

Master akademske studije

modul Softversko inženjerstvo

2018/2019

Naivni bajesovski klasifikator

Vuk Batanović, Elektrotehnički fakultet Univerziteta u Beogradu

Thomas Bayes (1701 – 1761)

- Engleski statističar i filozof
- Autor teoreme o verovatnoći:

$$P(A|B) = \frac{P(A,B)}{P(B)} = \frac{P(B|A)P(A)}{P(B)}$$

Primena Bajesove teoreme u klasifikaciji

$$P(y|x) = \frac{P(y,x)}{P(x)} = \frac{P(x|y)P(y)}{P(x)}$$

- y je klasa, x je podatak
- $P(y)$ – prethodna/apriorna verovatnoća klase (engl. *prior probability*)
- $P(y|x)$ - aposteriorna verovatnoća klase (engl. *posterior probability*)
- $P(y,x)$ – zajednička verovatnoća klase i podatka (engl. *joint probability*)
- $P(x|y)$ – funkcija izvesnosti (engl. *likelihood function*)

Primena Bajesove teoreme u klasifikaciji

$$P(y|x) = \frac{P(y,x)}{P(x)} = \frac{P(x|y)P(y)}{P(x)}$$

- Klasifikaciona odluka – podatak svrstati u onu klasu koja je za njega najverovatnija tj. ima najveću aposteriornu verovatnoću
- Ovakvo odlučivanje se naziva *Maximum a posteriori - MAP*

$$y_{MAP} = argmax_y\, P(y|x)$$

$$= argmax_y \frac{P(x|y)P(y)}{P(x)}$$

$$= argmax_y\, P(x|y)P(y)$$

Primena Bajesove teoreme u klasifikaciji

- Da bi se izvršila klasifikacija potrebni su $P(y)$ i $P(x|y)$
- $P(y)$ – zastupljenost klase y u skupu podataka za obučavanje:

$$P(y) = \frac{Count(y)}{Count(x)}$$

- Podatak x se predstavlja pomoću nekih njegovih odlika, kojih ima n:

$$P(x|y) = P(x_1, x_2, x_3, \dots, x_n|y)$$

- Koristeći pravilo ulančavanja uslovnih verovatnoća sledi:

$$\begin{aligned} P(x_1, x_2, x_3, \dots, x_n|y) &= P(x_1|y)P(x_2, x_3, \dots, x_n|y, x_1) \\ &= P(x_1|y)P(x_2|y, x_1)P(x_3, \dots, x_n|y, x_1, x_2) \\ &= P(x_1|y)P(x_2|y, x_1) \cdots P(x_n|y, x_1, x_2, \dots, x_{n-1}) \end{aligned}$$

Naivna pretpostavka o statističkoj nezavisnosti odlika

- Pretpostavka: odlike koje se koriste pri odlučivanju nisu statistički zavisne jedna od druge – poznavanje vrednosti jedne odlike ništa ne govori o vrednosti neke druge
- Ako je tako, onda važi:

$$
\begin{aligned}
P(x|y) &= P(x_1, x_2, x_3, \dots, x_n|y) \\
&= P(x_1|y)P(x_2|y) \cdots P(x_n|y) \\
&= \prod_{i=1}^{n} P(x_i|y)
\end{aligned}
$$

- Time uslovna verovatnoća svake odlike zavisi samo od klase podatka
- Raspodela te verovatnoće - multinomijalna, Bernulijeva, Gausova,…

> “Essentially, all models are wrong, but some are useful.”

britanski statističar George Box

- Iako je praktično uvek pogrešna, pretpostavka o međusobnoj statističkoj nezavisnosti odlika znatno olakšava modeliranje
- U praksi, naivni bajesovski modeli često ostvaruju dobre rezultate – za pravilnu klasifikaciju nije neophodno poznavanje tačnih vrednosti $P(y|x)$ za sve y, već samo njihovog međusobnog redosleda

Formula za određivanje pripadnosti klasi

$$y_{MAP} = argmax_y \, P(x|y)P(y)$$

$$= argmax_y P(y) \prod_{i=1}^{n} P(x_i|y)$$

- Broj odlika je često veoma veliki, a verovatnoće su u opsegu $[0,1]$
- Da bi se izbegao *floating point underflow* umesto množenja verovatnoća vrši se sabiranje njihovih logaritama:

$$y_{MAP} = argmax_y \left(\log P(y) + \sum_{i=1}^{n} \log P(x_i|y) \right)$$

Metod najveće izvesnosti

- Verovatnoća $P(x_t|y)$ – odnos između zbirne vrednosti odlike x_t u podacima klase y i zbirne vrednosti svih odlika u podacima te klase:

$$P(x_t|y) = \frac{Count(x_t, y)}{\sum_{i=1}^{n} Count(x_i, y)}$$

- Ovakav metod određivanja verovatnoća $P(x_t|y)$ se naziva metod najveće izvesnosti (engl. *Maximum Likelihood Estimation – MLE*) jer se njime maksimizuje uslovna verovatnoća/funkcija izvesnosti podataka nad kojima se model obučava

Metod najveće izvesnosti

- Šta se dešava kada se odlika x_t nikad ne javlja u podacima klase y u skupu za obučavanje ?
  - $P(x_t|y) = 0$
  - Model odbacuje klasu y na osnovu samo jedne odlike
- Ovakvo rezonovanje je problematično jer odsustvo odlike x_t u podacima klase y iz skupa za obučavanje može da bude slučajno
  - Naročito ako je broj odlika veliki, a skup za obučavanje mali – problem proređenosti podataka (engl. *data sparsity*)
  - Preterana prilagođenost modela (engl. *overfitting*)
- Korekcija – umesto $P(x_t|y) = 0$ staviti da je $P(x_t|y)$ jednako nekoj maloj nenultoj vrednosti

Laplasovo poravnanje

- Laplasovo poravnanje (engl. *Laplace smoothing*) ili aditivno poravnanje – brojanje zbirne vrednosti odlika u svakoj klasi ne počinje od nule već od neke predefinisane vrednosti α
  - Obično se uzima $\alpha = 1$
- Verovatnoća $P(x_t|y)$ se stoga računa kao:

$$P(x_t|y) = \frac{Count(x_t, y) + 1}{\sum_{i=1}^{n}(Count(x_i, y) + 1)} = \frac{Count(x_t, y) + 1}{n + \sum_{i=1}^{n} Count(x_i, y)}$$

- Laplasovo poravnanje se može smatrati oblikom regularizacije modela
- Deo verovatnoća javljanja odlika viđenih u klasi y u skupu za obučavanje se preraspodeljuje na odlike koje u toj klasi nisu (još uvek) opažene

Poziciona nezavisnost odlika

- (Multinomijalni) naivni bajesovski model implicitno pretpostavlja da ne postoji zavisnost između uslovne verovatnoće odlika i njihove pozicije u podacima
  - Primer: raspored reči u tekstu pri klasifikaciji dokumenata
- Ukoliko su podaci tako strukturirani da postoji redosled između odlika koje se razmatraju, taj redosled se u modelu ignoriše
  - U modeliranju tekstualnih podataka ovaj pristup je poznat kao vreća reči (engl. *bag-of-words*), pri čemu su individualne reči odlike koje se uzimaju u obzir
- Ovakva pretpostavka omogućava da se broj parametara modela drastično smanji

Multinomijalni naivni bajesovski klasifikator

- Do sada opisan model se naziva multinomijalnim bajesovskim klasifikatorom (engl. *Multinomial Naïve Bayes*)
- $Count(x_t, y)$ govori o zbirnoj vrednosti odlike x_t u podacima klase y
- Ponekad se odlike binarizuju – tada $Count(x_t, y)$ govori u koliko podataka klase y se odlika x_t javlja, a ne koja je njena zbirna vrednost u tim podacima
- U računanju $P(x|y)$ učestvuju samo one odlike koje se javljaju (imaju nenultu vrednost) u podatku x

Primer klasifikacije teksta – multinomijalni naivni bajesovski klasifikator

	doc	Reči u dokumentu	Klasa
Skup za obučavanje	1	Chinese Beijing Chinese	C
	2	Chinese Chinese Shanghai	C
	3	Beijing Macao	C
	4	Tokyo Japan Chinese	J
Skup za testiranje	5	Chinese Chinese Chinese Tokyo Japan	?

$$P(C) = \frac{3}{4}$$

$$P(J) = \frac{1}{4}$$

$$P(Chinese|C) = \frac{4+1}{6+8} = \frac{5}{14} \qquad P(Tokyo|C) = \frac{0+1}{6+8} = \frac{1}{14} \qquad P(Japan|C) = \frac{0+1}{6+8} = \frac{1}{14}$$

$$P(Chinese|J) = \frac{1+1}{6+3} = \frac{2}{9} \qquad P(Tokyo|J) = \frac{1+1}{6+3} = \frac{2}{9} \qquad P(Japan|J) = \frac{1+1}{6+3} = \frac{2}{9}$$

Primer klasifikacije teksta – multinomijalni naivni bajesovski klasifikator

	doc	Reči u dokumentu	Klasa
Skup za obučavanje	1	Chinese Beijing Chinese	C
	2	Chinese Chinese Shanghai	C
	3	Beijing Macao	C
	4	Tokyo Japan Chinese	J
Skup za testiranje	5	Chinese Chinese Chinese Tokyo Japan	?

$$P(C|d5) \propto P(C)P(Chinese|C)^3 P(Tokyo|C)P(Japan|C) = \frac{3}{4} \times \left(\frac{5}{14}\right)^3 \times \frac{1}{14} \times \frac{1}{14} \approx 0.000174$$

$$P(J|d5) \propto P(J)P(Chinese|J)^3 P(Tokyo|J)P(Japan|J) = \frac{1}{4} \times \left(\frac{2}{9}\right)^3 \times \frac{2}{9} \times \frac{2}{9} \approx 0.000135$$

Multivarijacioni Bernulijev naivni bajesovski klasifikator

- U čestoj upotrebi u klasifikaciji teksta je i multivarijacioni Bernulijev bajesovski klasifikator (engl. *Multivariate Bernoulli Naïve Bayes*)
- Odlike su binarnog/indikatorskog tipa
- $Count(x_t, y)$ govori u koliko podataka klase y se odlika x_t javlja
- U računanju $P(x|y)$ učestvuju sve odlike koje se koriste u modelu, čak i one koje se ne javljaju (imaju nultu vrednost) u podatku x
- Pretpostavka o pozicionoj nezavisnosti odlika ovde nije neophodna, jer Bernulijev model po definiciji uzima u obzir samo prisustvo/odsustvo odlika

Multivarijacioni Bernulijev naivni bajesovski klasifikator

- Verovatnoća $P(x_t|y)$ – procenat podataka klase y u kojima je odlika x_t prisutna:

$$P(x_t|y) = \frac{Count(x_t, y)}{Count(y)}$$

- Laplasovo poravnanje:

$$P(x_t|y) = \frac{Count(x_t, y) + 1}{K + Count(y)} = \frac{Count(x_t, y) + 1}{2 + Count(y)}$$

- K je broj mogućih ishoda tj. vrednosti odlike x_t
- $K = 2$ kad je odlika x_t binarna

Primer klasifikacije teksta – multivarijacioni Bernulijev naivni bajesovski klasifikator

	doc	Reči u dokumentu	Klasa
Skup za obučavanje	1	Chinese Beijing Chinese	C
	2	Chinese Chinese Shanghai	C
	3	Beijing Macao	C
	4	Tokyo Japan Chinese	J
Skup za testiranje	5	Chinese Chinese Chinese Tokyo Japan	?

$$P(C) = \frac{3}{4}$$

$$P(J) = \frac{1}{4}$$

$$P(Chinese|C) = \frac{2+1}{2+3} = \frac{3}{5} \qquad P(Tokyo|C) = \frac{0+1}{2+3} = \frac{1}{5} \qquad P(Japan|C) = \frac{0+1}{2+3} = \frac{1}{5}$$

$$P(Chinese|J) = \frac{1+1}{2+1} = \frac{2}{3} \qquad P(Tokyo|J) = \frac{1+1}{2+1} = \frac{2}{3} \qquad P(Japan|J) = \frac{1+1}{2+1} = \frac{2}{3}$$

Primer klasifikacije teksta – multivarijacioni Bernulijev naivni bajesovski klasifikator

	doc	Reči u dokumentu	Klasa
Skup za obučavanje	1	Chinese Beijing Chinese	C
	2	Chinese Chinese Shanghai	C
	3	Beijing Macao	C
	4	Tokyo Japan Chinese	J
Skup za testiranje	5	Chinese Chinese Chinese Tokyo Japan	?

$$P(C) = \frac{3}{4}$$

$$P(J) = \frac{1}{4}$$

$$P(Beijing|C) = \frac{2+1}{2+3} = \frac{3}{5} \qquad P(Shanghai|C) = \frac{1+1}{2+3} = \frac{2}{5} \qquad P(Macao|C) = \frac{1+1}{2+3} = \frac{2}{5}$$

$$P(Beijing|J) = \frac{0+1}{2+1} = \frac{1}{3} \qquad P(Shanghai|J) = \frac{0+1}{2+1} = \frac{1}{3} \qquad P(Macao|J) = \frac{0+1}{2+1} = \frac{1}{3}$$

Primer klasifikacije teksta – multivarijacioni Bernulijev naivni bajesovski klasifikator

	doc	Reči u dokumentu	Klasa
Skup za obučavanje	1	Chinese Beijing Chinese	C
	2	Chinese Chinese Shanghai	C
	3	Beijing Macao	C
	4	Tokyo Japan Chinese	J
Skup za testiranje	5	Chinese Chinese Chinese Tokyo Japan	?

$$P(C|d5) \propto P(C) \times P(Chinese|C) \times P(Tokyo|C) \times P(Japan|C)$$

$$\times \big(1 - P(Beijing|C)\big) \times \big(1 - P(Shanghai|C)\big) \times \big(1 - P(Macao|C)\big) =$$

$$\frac{3}{4} \times \frac{3}{5} \times \frac{1}{5} \times \frac{1}{5} \times \frac{2}{5} \times \frac{3}{5} \times \frac{3}{5} \approx 0.0026$$

Primer klasifikacije teksta – multivarijacioni Bernulijev naivni bajesovski klasifikator

	doc	Reči u dokumentu	Klasa
Skup za obučavanje	1	Chinese Beijing Chinese	C
	2	Chinese Chinese Shanghai	C
	3	Beijing Macao	C
	4	Tokyo Japan Chinese	J
Skup za testiranje	5	Chinese Chinese Chinese Tokyo Japan	?

$$P(J|d5) \propto P(J) \times P(Chinese|J) \times P(Tokyo|J) \times P(Japan|J)$$

$$\times \left(1 - P(Beijing|J)\right) \times \left(1 - P(Shanghai|J)\right) \times \left(1 - P(Macao|J)\right) =$$

$$\frac{1}{4} \times \frac{2}{3} \times \frac{2}{3} \times \frac{2}{3} \times \frac{2}{3} \times \frac{2}{3} \times \frac{2}{3} \approx 0.022$$

Razlike između multinomijalnog i Bernulijevog modela u klasifikaciji teksta

	Multinomijalni model	Bernulijev model
$Count(x_t, y)$	Broje se (sabiraju) vrednosti odlike x_t u podacima klase y	Broje se javljanja odlike x_t u podacima klase y
x_t	Nije binarnog tipa (ali se može binarizovati)	Binarnog tipa
$P(x_t\|y)$	Odnos zbirne vrednosti odlike x_t i zbirne vrednosti svih odlika u podacima klase y	Procenat podataka klase y u kojima je odlika x_t prisutna
$P(x\|y)$	Uključuje samo odlike koje se javljaju (imaju nenultu vrednost) u podatku x	Uključuje sve odlike koje se koriste u modelu
Laplasovo poravnanje	$\frac{Count(x_t, y) + 1}{n + \sum_{i=1}^{n} Count(x_i, y)}$	$\frac{Count(x_t, y) + 1}{K + Count(y)}$

Razlike između multinomijalnog i Bernulijevog modela u klasifikaciji teksta

- Pretpostavka o pozicionoj nezavisnosti odlika
  - Multinomijalni model – potrebna
  - Bernulijev model – nepotrebna
- Binarizovan multinomijalni model nije jednak Bernulijevom!

Generativni modeli

- Modeliraju zajedničku verovatnoću klase i podataka $P(y,x)$
- Na taj način se modelira proces generisanja podataka određenog tipa
- Naivni bajesovski klasifikator je generativni model – aposteriorna verovatnoća klase se dobija iz zajedničke verovatnoće preko Bajesove teoreme:

$$P(y|x) = \frac{P(y,x)}{P(x)}$$

- U generativne modele spadaju:
  - skriveni Markovljevi lanci (engl. HMM – *Hidden Markov Models*)
  - Bajesovske mreže (engl. *Bayesian networks*)
  - latentna Dirišleova alokacija (engl. LDA – *Latent Dirichlet Allocation*)

Prednosti naivnog bajesovskog klasifikatora

- Jednostavnost
- Brzina učenja i klasifikacije – dovoljan jedan prolaz kroz podatke
- Nije osetljiv na irelevantne odlike
- Dobro se ponaša kada postoji veći broj podjednako važnih odlika
- Interpretabilnost – veća težina pridružena određenoj odlici znači da je ona važnija pri odlučivanju
- Direktno je primenjiv na višeklasnu klasifikaciju
- Model se lako može ažurirati novopristiglim podacima

Mane naivnog bajesovskog klasifikatora

- (Netačna) pretpostavka o statističkoj nezavisnosti odlika
  - Kada su dve odlike korelisane dolazi do duplog brojanja – daje iskrivljene procene o važnosti odlika
- Vrednosti aposteriornih verovatnoća koje model daje su često znatno iskrivljene u korist najverovatnije klase

Upotreba naivnog bajesovskog klasifikatora

- Često se koristi kao *baseline* zbog svoje jednostavnosti i brzine
- Često postiže bolje performanse od složenijih modela kada je dostupno malo podataka za obučavanje
- Zbog brzine je kompetitivan i kada je dostupno puno podataka
- Raširen algoritam u klasifikaciji tekstualnih podataka
  - Multinomijalna varijanta (sa binarizovanim odlikama) obično najbolja
  - Primene - detekcija spama, analiza sentimenta, klasifikacija teksta po temama, analiza autorstva,…
- Koristi se i u medicinskoj dijagnostici, analizi genomskih podataka,…