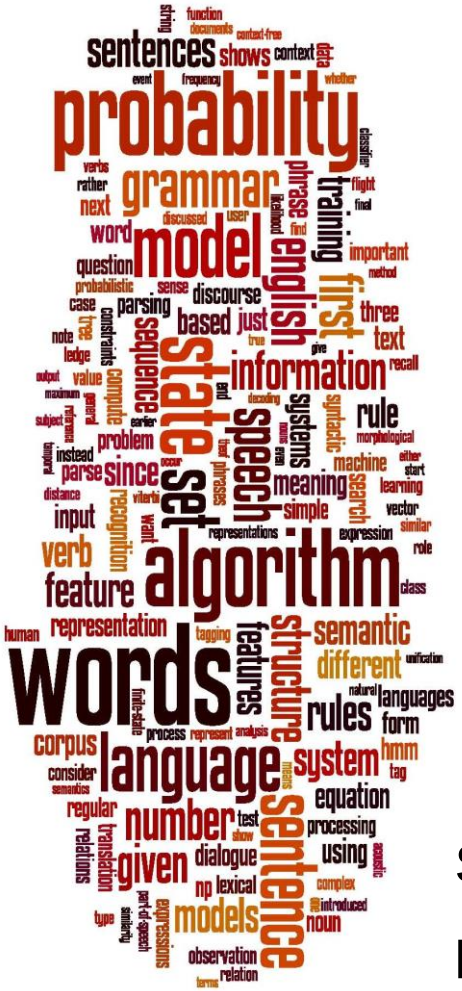


Naivni Bajesov Klasifikator

Klasifikacija Teksta

Predavač: Aleksandar Kovačević

Slajdovi preuzeti sa CS 124, Stanford

https://web.stanford.edu/class/cs124/

Da li je e-mail spam?

Subject: **Important notice!**
From: Stanford University <newsforum@stanford.edu>
Date: October 28, 2011 12:34:16 PM PDT
To: undisclosed-recipients:;

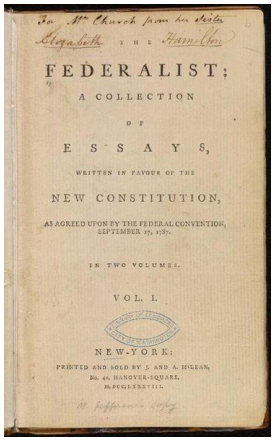
Greats News!

You can now access the latest news by using the link below to login to Stanford University News Forum.

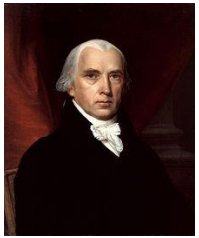
http://www.123contactform.com/contact-form-StanfordNew1-236335.html

Click on the above link to login for more information about this new exciting forum. You can also copy the above link to your browser bar and login for more information
about the new services.

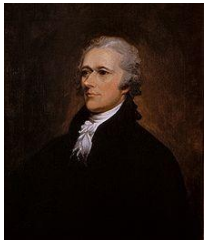
Ko je napisao "*The Federalist*" dokumente?



- 1787–8: anonimni tekstovi sa ciljem da ubede New York da ratifikuje Ustav S.A.D – pravi autori su Jay, Madison i Hamilton.
- Za 12 dokumenta su se vodile rasprave o tome ko je tačno od trojice navedenih pravi autor
- 1963: problem je rešen od strane Mosteller i Wallace upotrebom Bajesovskih metoda



James Madison



Alexander Hamilton

Da li je autor muško ili žensko?

1. By 1925 present--day Vietnam was divided into three parts under French colonial rule. The southern region embracing Saigon and the Mekong delta was the colony of Cochin--China; the central area with its imperial capital at Hue was the protectorate of Annam…
2. Clara never failed to be astonished by the extraordinary felicity of her own name. She found it hard to trust herself to the mercy of fate, which had managed over the years to convert her greatest shame into one of her greatest assets…

S. Argamon, M. Koppel, J. Fine, A. R. Shimoni, 2003. "Gender, Genre, and Writing Style in Formal Written Texts," Text, volume 23, number 3, pp. 321–346

Da li je recenzija filma (knjige, prozivoda...) pozitvina ili negativna?

- unbelievably disappointing

- Full of zany characters and richly applied satire, and some great plot twists

- this is the greatest screwball comedy ever filmed

- It was pathetic. The worst part about it was the boxing scenes.

Šta je tema ovog rada?

Rad sa MEDLINE

?

MeSH - Hijerarhija Tema:

- Antogonists and Inhibitors
- Blood Supply
- Chemistry
- Drug Therapy
- Embryology
- Epidemiology
- …

Klasifikacija teksta

- Dodela kategorija, tema ili žanrova
- Detekcija spama
- Identifikacija autora
- Identifikacija pola ili godišta autora
- Identifikacija jezika
- Analiza (detekcija) sentimenta
- …

Klasifikacija Teksta: definicija

- *Ulaz*:
  - dokument d
  - skup klasa $C = \{c_1, c_2, \ldots, c_J\}$

- *Izlaz*: prediktovana klasa $c \in C$

Klasifikacione metode: Ručno kreirana pravila

- Pravila zasnovana na kombinacijama reči i drugih osobina
  - spam: Na primer, detektovana je e-mail adresa sa „crne liste“ ILI teksta sadrži reči “dollars” i “have been selected”.
- Tačnost može biti visoka
  - Ako se pravila pažljivo kreiraju uz pomoć eksperata
- Ali, kreiranje i održavanje ovih pravila je vremenski zahtevan proces

Klasifikacione metode: Nadgledano Mašinsko Učenje

- *Ulaz:*
  - dokument d
  - skup klasa $C = \{c_1, c_2, \ldots, c_J\}$
  - Obučavajući skup od m ručno-označenih dokumenata $(d_1,c_1),\ldots,(d_m,c_m)$
- *Izlaz:*
  - klasifikator $\gamma: d \rightarrow c$

Klasifikacione metode: Nadgledano Mašinsko Učenje

- Može se koristiti bilo koji klasifikator
  - Naivni Bajes (*Naïve Bayes*)
  - Logistička regresija
  - Mašine Potpornih Vektora
  - K-Najbližih Komšija
  - ...

Klasifikacija Teksta i Naivni Bajes

Naivni Bajes Intuicija

- Jednostavan (“naivan”) klasfikacioni metod zasnovan na Bajesovoj teoremi
- Oslanja se na jako jednostavnu reprezentaciju dokumenata
  - Vreća reči (*Bag of words*)

Bag-of-words reprezentacija

$\gamma($ I love this movie! It's sweet, but with satirical humor. The dialogue is great and the adventure scenes are fun… It manages to be whimsical and romantic while laughing at the conventions of the fairy tale genre. I would recommend it to just about anyone. I've seen it several times, and I'm always happy to see it again whenever I have a friend who hasn't seen it yet. $)=c$

Bag-of-words reprezentacija

$\gamma($ I **love** this movie! It's **sweet**, but with **satirical** humor. The dialogue is **great** and the adventure scenes are **fun**… It manages to be **whimsical** and **romantic** while **laughing** at the conventions of the fairy tale genre. I would **recommend** it to just about anyone. I've seen it **several** times, and I'm always **happy** to see it **again** whenever I have a friend who hasn't seen it yet. $)=C$

Dan Jurafsky

Bag-of-words reprezentacija: koristimo podskup svih reči – termin koji se koristi za ovaj podskup je rečnik

$\gamma($ x love xxxxxxxxxxxxxxxx sweet xxxxxxx satirical xxxxxxxxxx xxxxxxxxxxx great xxxxxxx xxxxxxxxxxxxxxxxxxx fun xxxx xxxxxxxxxxxxx whimsical xxxx romantic xxxx laughing xxxxxxxxxxxxxxxxxxxxxxxxxxxxxx xxxxxxxxxxxxxx recommend xxxxx xxxxxxxxxxxxxxxxxxxxxxxxxxxxxx xx several xxxxxxxxxxxxxxxx xxxxx happy xxxxxxxxx again xxxxxxxxxxxxxxxxxxxxxxxxxxxxxx xxxxxxxxxxxxxxxx $)=C$

Bag-of-words reprezentacija

$\gamma($

great	2
love	2
recommend	1
laugh	1
happy	1
...	...

$)=c$

Bag-of-words reprezentacija za klasifikaciju dokumenata

Test dokument

parser
language
label
translation
…

?

Machine Learning	NLP	Garbage Collection	Planning	GUI
learning	parser	garbage	planning	…
training	tag	collection	temporal	
algorithm	training	memory	reasoning	
shrinkage	translation	optimization	plan	
network…	language…	region…	language…	

Naivni Bajesov Klasifikator – Formalnija Definicija

Bajesova teorema primenjena na dokumente i klase

- Za dokument *d* i klasu *c*

$$P(c \mid d) = \frac{P(d \mid c)P(c)}{P(d)}$$

Bayes-ова теорема

Bayes-ова теорема:

информације које нам доносе подаци

(вероватноћа да се догодио А ако знамо да се догодио С)

априорна вероватноћа-

претходно знање

(оно што знамо о догађајима А и С без скупа података)

$$P(C \mid A) = \frac{P(A \mid C)P(C)}{P(A)}$$

апостериорна вероватноћа

комбинација претходног знања и доказа из података

Пример Bayes-ове теореме

Дато:

Лекар зна да менингитис узрокује укочен врат у 50% случајева (информације које нам доносе подаци)

Априорна вероватноћа да пацијент има менингитис је 1/50,000 (претходно знање)

Априорна вероватноћа да пацијент има укочен врат је 1/20 (претходно знање)

Која је вероватноћа да пацијент који има укочен врат има менингитис?

$$P(M \mid S) = \frac{P(S \mid M)P(M)}{P(S)} = \frac{0.5 \times 1/50000}{1/20} = 0.0002$$

Naivni Bajesov klasifikator

$$c_{MAP} = \underset{c \in C}{\operatorname{argmax}} P(c \mid d)$$

MAP - "maksimalna aposterirorna verovatnoća"= najverovatnija klasa

$$= \underset{c \in C}{\operatorname{argmax}} \frac{P(d \mid c)P(c)}{P(d)}$$

Bajesova teorema

$$= \underset{c \in C}{\operatorname{argmax}} P(d \mid c)P(c)$$

Izbacujemo imenilac

Naivni Bajesov klasifikator

$$c_{MAP} = \underset{c \in C}{\operatorname{argmax}} P(d \mid c) P(c)$$

$$= \underset{c \in C}{\operatorname{argmax}} P(x_1, x_2, \ldots, x_n \mid c) P(c)$$

Dokument d reprezentovan pomoću osobina (*features*) $x_1..x_n$

Naivni Bajesov klasifikator

$$c_{MAP} = \underset{c \in C}{\operatorname{argmax}} P(x_1, x_2, \ldots, x_n \mid c) P(c)$$

$P(x_1..x_n|C)$ - Može samo da se proceni ako imamo jako jako puno primera u obučavajućem skupu.

P(c) - Koliko se često javlja klasa c?

Možemo samo da izbrojimo tj. izračunamo relativne frekvencije u korpusu.

Multinomialni Naivni Bajes – pretpostavka o nezavisnoti osobina (atributa)

$$P(x_1, x_2, \ldots, x_n \mid c)$$

- **Pretpostavka Bag-of-Words reprezentacije**: Pretpostavljamo da pozicija reči u tekstu nije važna
- **Uslovna nezavisnost**: Pretpostavljamo da su verovatnoće osobina tj. $P(x_i \mid c_j)$ nezavisne u odnosu na klasu *c.* Na taj način verovatnoću $P(x_1,..,x_n \mid C)$ računamo kao:

$$P(x_1, \ldots, x_n \mid c) = P(x_1 \mid c) \bullet P(x_2 \mid c) \bullet P(x_3 \mid c) \bullet \ldots \bullet P(x_n \mid c)$$

Multinomialni Naivni Bajes – pretpostavka o nezavisnoti osobina (atributa) $P(x_1, x_2, \ldots, x_n \mid c)$

- **Napomena:**
- **Termin multinominalni odnosi se na način izračunavanja verovatnoća $P(x_i \mid c)$.**
- **Pretpostavlja se da vrednosti $P(x_i \mid c)$ prate multinominalnu distribuciju.**
- **Postoje NB klasifikatori kod kojih se pretpostavju druge distribucije (binominalna, normalna itd.).**
- **Primere takvih klasifikatora radićemo na predmetu mašinsko učenje i SIAP.**

$$P(x_1, \ldots, x_n \mid c) = P(x_1 \mid c) \bullet P(x_2 \mid c) \bullet P(x_3 \mid c) \bullet \ldots \bullet P(x_n \mid c)$$

Multinomialni Naivni Bajes klasifikator

$$c_{MAP} = \underset{c \in C}{\operatorname{argmax}} P(x_1, x_2, \ldots, x_n \mid c) P(c)$$

$$c_{NB} = \underset{c \in C}{\operatorname{argmax}} P(c_j) \prod_{x \in X} P(x \mid c)$$

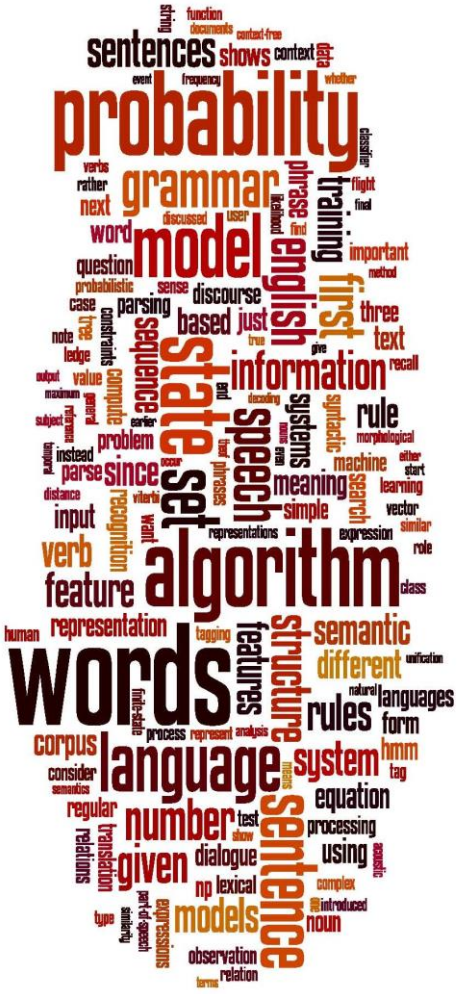
Primena NB klasifikatora na klasifikaciju teksta

positions ← sve pozicije reči u test dokumentu

$$c_{NB} = \underset{c_j \in C}{\operatorname{argmax}} P(c_j) \prod_{i \in positions} P(x_i \mid c_j)$$

	Doc	Words	Class
Training	1	Chinese Beijing Chinese	c
	2	Chinese Chinese Shanghai	c
	3	Chinese Macao	c
	4	Tokyo Japan Chinese	j
Test	5	Chinese Chinese Chinese Tokyo Japan	?

Za ovaj primer positions = {0,1,2,3,4}



Naivni Bajesov Klasifikator - Obučavanje

Obučavanje NB klasifikatora

- Koristimo metod maksimalne verovatnosti (*maximum likelihood estimates*)
  - konkretno koristimo frekvencije dobijene iz korpusa

$$\hat{P}(c_j) = \frac{doccount(C = c_j)}{N_{doc}}$$

Broj dokumenta klase c_j u korpusu podeljen sa brojem svih dokumenata u korpusu N_{doc}

$$\hat{P}(w_i \mid c_j) = \frac{count(w_i, c_j)}{\sum_{w \in V} count(w, c_j)}$$

V - predstavlja rečnik odnosno skup svih reči koje koristimo za reprezentaciju dokumenata.

Izračunavanje verovatnoća

$$\hat{P}(w_i \mid c_j) = \frac{count(w_i, c_j)}{\sum_{w \in V} count(w, c_j)}$$

broj pojavljivanja reči w_i u dokumentima koji imaju klasu c_j podeljen sa brojem pojavljivanja svih reči iz rečnika V u svim dokumentima koji imaju klasu c_j

- Kreiramo mega-dokument za klasu *j* tako što sve dokumente ove klase spojimo u jedan dokument
  - Računamo frekvenciju reču w_i u tom mega-dokumentu

Problem sa prethodnim formulama

- Šta ako nijedan od dokumenata **pozitivne (*thumbs--up*) klase** u obučavajućem skupu nema reč ***fantastic***?

$$\hat{P}(\text{"fantastic"}|\text{positive}) = \frac{count(\text{"fantastic"}, \text{positive})}{\sum_{w \in V} count(w, \text{positive})} = 0$$

- Bez obzira na vrednosti drugih verovatnoća ako je jedna od verovatnoća 0 verovatnoća klase *c* je 0!

$$c_{MAP} = \text{argmax}_c \hat{P}(c) \prod_i \hat{P}(x_i|c)$$

Laplasovo (dodaj-1) poravnavanje za NB

Laplace (add–1) smoothing

$$\hat{P}(w_i \mid c) = \frac{count(w_i, c) + 1}{\sum_{w \in V} \left(count(w, c) + 1 \right)}$$

$$= \frac{count(w_i, c) + 1}{\left(\sum_{w \in V} count(w, c) \right) + |V|}$$

Multinominalni Naivni Bajes: obučavanje

- Iz obučavajućeg skupa odrediti *Rečink* (*Vocabulary*) – npr. naš rečnik čine 100 najfrekventjih reči u korpusu – naravno postoje i mnoge druge sofisticiranije metode za određivanje rečnika.

- Izračunati $P(c_j)$
  - Za svako c_j uraditi

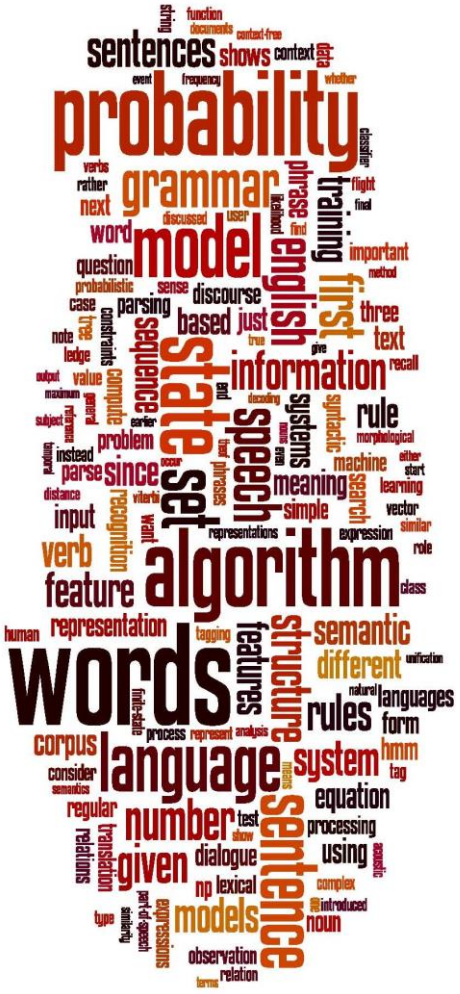
    $docs_j \leftarrow$ svi dokumenti klase c_j

$$P(c_j) \leftarrow \frac{|docs_j|}{|\text{total \# documents}|}$$

- Izračunati $P(w_k \mid c_j)$
  - $Text_j \leftarrow$ jedan dokument koji sadrži sve $docs_j$
  - Za svaku reč w_k u *Rečinku*

    $n_k \leftarrow$ # pojavljivanja w_k u $Text_j$

$$P(w_k \mid c_j) \leftarrow \frac{n_k + \alpha}{n + \alpha |Vocabulary|}$$



Naivni Bajes – relacija sa modelima jezika (*language models*)

Naivni Bajes i Modeli Jezika

- NB klasifikator može da koristi bilo koje osobine
  - URL, e-mail adrese, rečnike, emotikone...
- Ali ako koristimo B-O-W reprezentaciju kao na prethodnim slajdovima i važi da:
  - Kao osobine koristimo samo reči
  - Koristimo sve reči, a ne neki podskup (rečnik)
- Onda je
  - NB vrlo sličan sa modelima jezika.

Jedna klasa = model jezika zasnovan na unigramima (rečima)

- Svakoj reči dodeljujemo: P(word | c)
- Svakoj rečenici dodeljujemo: P(s|c)=Π P(word|c)

Klasa *pozitivno*

0.1	I
0.1	love
0.01	this
0.05	fun
0.1	film

I	love	this	fun	film
0.1	0.1	.05	0.01	0.1

P(s | pos) = 0.0000005

Naivni Bajes kao Model Jezika

Koja klasa dodeljuje veću verovatnoću s?

Model pozitivno	
0.1	I
0.1	love
0.01	this
0.05	fun
0.1	film

Model negativno	
0.2	I
0.001	love
0.01	this
0.005	fun
0.1	film

I	love	this	fun	film
0.1	0.1	0.01	0.05	0.1
0.2	0.001	0.01	0.005	0.1

$P(s|pos) > P(s|neg)$

NB klasifikator Primer

$$\hat{P}(c) = \frac{N_c}{N}$$

$$\hat{P}(w \mid c) = \frac{count(w,c)+1}{count(c)+|V|}$$

	Dok	Reč	Klasa
Ob. skup	1	Chinese Beijing Chinese	c
	2	Chinese Chinese Shanghai	c
	3	Chinese Macao	c
	4	Tokyo Japan Chinese	j
Test skup	5	Chinese Chinese Chinese Tokyo Japan	?

Verovatnoće klasa:

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

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

Uslovne verovatnoće:

P(Chinese|*c*) = (5+1) / (8+6) = 6/14 = 3/7

P(Tokyo|*c*) = (0+1) / (8+6) = 1/14

P(Japan|*c*) = (0+1) / (8+6) = 1/14

P(Chinese|*j*) = (1+1) / (3+6) = 2/9

P(Tokyo|*j*) = (1+1) / (3+6) = 2/9

P(Japan|*j*) = (1+1) / (3+6) = 2/9

Određivanje klase:

$P(c|d5) \propto 3/4 * (3/7)^3 * 1/14 * 1/14 \approx 0.0003$

$P(j|d5) \propto 1/4 * (2/9)^3 * 2/9 * 2/9 \approx 0.0001$

NB – spam filter

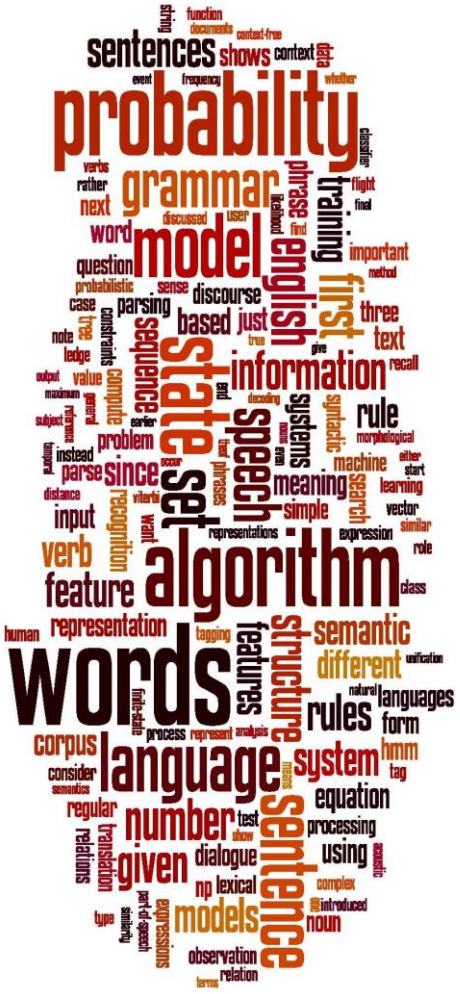
- Neke od mogućih osobina – kokretno softver SpamAssasin:
  - Pominjanja: "Generic Viagra", "Online Pharmacy " i slično...
  - Pominjanja: millions of (dollar) ((dollar) NN,NNN,NNN.NN)
  - Fraze: impress ... girl
  - From: ima puno brojeva
  - Subject e-maila je sav velikim slovima all capitals
  - HTML ima malo teksta u odnosu na slike
  - "One hundred percent guaranteed"
  - Recenica u kojoj se tvrdi da mozete da se ojdavite sa liste
  - "Prestigious Non--Accredited Universities"
  - http://spamassassin.apache.org/tests_3_3_x.html

Naivni Bajes nije baš tako naivan

- Vrlo brz, ne treba mu puno memorije
- Robusan na bezznačajne osobine

  Osobine koje su jako slične za sve primere ne menjaju verovatnoće
- Jako dobar kod problema gde postoji puno osobina jednakog kvaliteta

  Stabla odlučivanja imaju problem *fragmentacije* u takvim slučajevima
- Ako pretpostavka o nezavisnosti stvarno važi NB je optimalan model
- Čak iako ne važi praksa je pokazala da je NB dobar klasifikator
- U trenutnom stajnu ML oblasti NB je pozdan model i dobar osnonovni model



NB klasifikator
Evaluacija

Preciznost
Odziv
F-mera

Matrica Konfuzije (Confusion Matrix)

- Odlučimo koja je klasa za nas pozitivna, koja negativna i formiramo matricu.

	starno poz.	stvarno neg.
predikcija poz.	tp	fp
predikcija neg.	fn	tn

Preciznost i Odziv
Precision and recall

- **Preciznost (*Precision*)** : % predikcija poz. koji su stvarno poz tj. tp/(tp+fp)
- **Odziv (*Recall*)**: % stvarno poz od predikcija poz. tj. tp/(tp+fn)
- Ovo su preciznost i odziv za pozitivnu klasu, negativna klasa također ima preciznost i odziv koji se računaju na isti način, ali se ona posmatra kao pozitivna

	starno poz.	stvarno neg.
predikcija poz.	tp	fp
predikcija neg.	fn	tn

F-mera

- Mera koja procenjuje koji je balans između P i R:

$$F = \frac{2PR}{\mathrm{P+R}}$$

Ako imamo više od dve klase: Supovi binarnih klasifikatora

- Problemi kod kojih imamo multi-klasnu klasifikaciju
  - Kasifikujemo dokument koji može da pripada 0, 1, ili >1 klasa.
  - Kad kažemo 0 klasa misli se na situaciju npr. kad dozvoljavamo recimo da dokument bude ni pozitivan ni negativan, ali pripada nekoj klasi koju obično obeležavamo sa *Ostalo.*
- Za svaku klasu $c \in C$
  - Obučiti klasifikator γ_c tako da može da razdvoji klasu c od ostalih klasa $c' \in C$
- Za test dokument d,
  - Primenti svaki model γ_c
  - d pripada svakoj klasi c za koju γ_c vrati true
  - Ako želimo jasnu klasfikaciju tj. samo jedno c možemo npr. da koristimo γ_c koji vraćaju pouzdanaost u true vrednost. Konačna klasa bila bi ona c čiji je model vratio najveću pouzdanost.

Evaluacija:
Jako poznat skup: Reuters--21578 Data Set

- Najčešće korišćen skup, 21,578 dokumenata
- 9603 obučavajući skup, 3299 test skup (ModApte/Lewis split)
- 118 klasa
  - Članak (dokument) može da bude u više od jedne kategorije
  - Učimo 118 binarnih klasifikatora

- Samo oko 10 od 118 kategorija ima veći broj članaka
  - Earn (2877, 1087)
  - Acquisitions (1650, 179)
  - Money-fx (538, 179)
  - Grain (433, 149)
  - Crude (389, 189)
  - Trade (369,119)
  - Interest (347, 131)
  - Ship (197, 89)
  - Wheat (212, 71)
  - Corn (182, 56)

Primer dokumenta iz Reuters--21578

<REUTERS TOPICS="YES" LEWISSPLIT="TRAIN" CGISPLIT="TRAINING-SET" OLDID="12981" NEWID="798">

<DATE> 2-MAR-1987 16:51:43.42</DATE>

<TOPICS><D>livestock</D><D>hog</D></TOPICS>

<TITLE>AMERICAN PORK CONGRESS KICKS OFF TOMORROW</TITLE>

<DATELINE> CHICAGO, March 2 - </DATELINE><BODY>The American Pork Congress kicks off tomorrow, March 3, in Indianapolis with 160 of the nations pork producers from 44 member states determining industry positions on a number of issues, according to the National Pork Producers Council, NPPC.

Delegates to the three day Congress will be considering 26 resolutions concerning various issues, including the future direction of farm policy and the tax law as it applies to the agriculture sector. The delegates will also debate whether to endorse concepts of a national PRV (pseudorabies virus) control and eradication program, the NPPC said.

A large trade show, in conjunction with the congress, will feature the latest in technology in all areas of the industry, the NPPC added. Reuter

&#3;</BODY></TEXT></REUTERS>

Matrica Konfuzije

- Za svaki par klasa <c_1,c_2> koliko dokumenata klase c_1 su pogrešno klasifikovani kao c_2?
  - $c_{3,2}$: 90 dokumenata klase *wheat* je pogrešno klasifikovano u klasu *poultry*

Dokumenti u test skup	Predikcija UK	Predikcija poultry	Predikcija wheat	Predikcija coffee	Predikcija interest	Predikcija trade
Stvarno UK	95	1	13	0	1	0
Stvarno poultry	0	1	0	0	0	0
Stvarno wheat	10	90	0	1	0	0
Stvarno coffee	0	0	0	34	3	7
Stvarno interest	-	1	2	13	26	5
Stvarno trade	0	0	2	14	5	10

Evalucione mere za jednu klasu

Odziv:

Deo dokumenata koji stvano imaju klasu *i* koji su tačno klasifikovani:

$$\frac{c_{ii}}{\sum_j c_{ij}}$$

, *i* je indeks vrste, a *j* kolone u tabeli sa prethodnog slajda – delimo element sa glavne dijagonale sa zbirom elemenata cele vrste.

Evalucione mere za jednu klasu

Preciznost:

Deo dokumenata kojima je klasifikator doelio klasu *i,* a koji su stvarno u klasi *i*:

$$\frac{c_{ii}}{\sum_j c_{ji}}$$

, *i* je indeks vrste, a *j* kolone u tabeli sa prethodnog slajda – delimo element sa glavne dijagonale sa zbirom elemenata cele kolone.

Evaluciona mera na za sve klase zajedno

Tačnost: (1 −error rate)

Deo svih dokumenata koji su tačno klasifikovani, u odnosu a sve dokumente u korpusu:

$$\frac{\sum_i c_{ii}}{\sum_j \sum_i c_{ij}}$$

, *i* je indeks vrste, a *j* kolone u tabeli sa prethodnog slajda – delimo zbir svih elemenata sa glavne dijagonale sa zbirom elemenata cele kolone.

Mikro i Makro proseci

- Ako imamo više klasa, na koji način kombinujemo mere na nivou jedne klase u jednu meru?
- **Makro prosek**: Izračunamo meru za svaku klasu i onda uzmemo prosek (npr. prosek F-mera)
- **Mikro prosek**: Formiramo matricu konfuzije i onda iz nje računamo mere na nivou cele matrice. Npr. za Odziv uradimo zbir svih tp u celoj matrici i podelimo sa svim (tp+fn)

Mikro i Makro proseci: Primer

Klasa 1

	Truth: yes	Truth: no
Classifier: yes	10	10
Classifier: no	10	970

Klasa 2

	Truth: yes	Truth: no
Classifier: yes	90	10
Classifier: no	10	890

Mikro Prosek Tabela

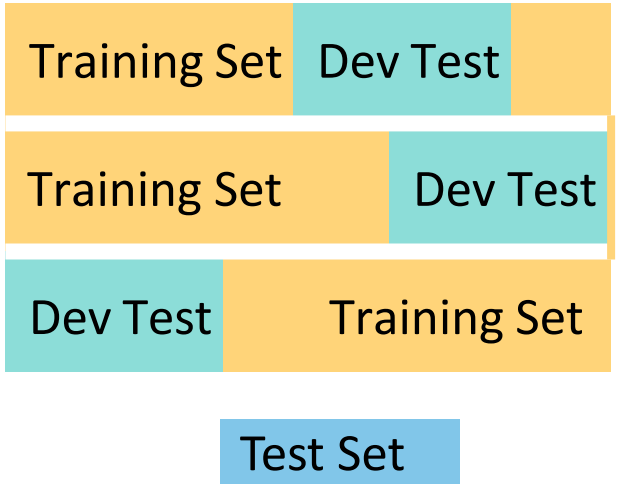
	Truth: yes	Truth: no
Classifier: yes	100	20
Classifier: no	20	1860

- Makro preciznost za *Yes*: (0.5 + 0.9)/2 = 0.7
- Mikro preciznost za *Yes*: 100/120 = 0.83
- Mikro vrednostima dominiraju klase koje su najfrekventnije

Unakrsna Validacija i Validacioni skup

Obučavajući skup | Validacioni skup | Test skup

- Validacioni skup – koristimo ga tokom razvoja modela za podešavanje parametara i selekciju osobina.
- Test skup – koristimo ga za krajnju procenu kvaliteta modela
  - može da bude varljiv (previše lak ili previše težak test skup)
- Unakrsna validacija (objašjeno na sledećem slajdu) – može da se koristi i tokom razvoja ili za krajnju procenu. Nikako za oba u isto vreme. Ako je koristimo u toku razvoja, moramo da imamo odvojen test skup.



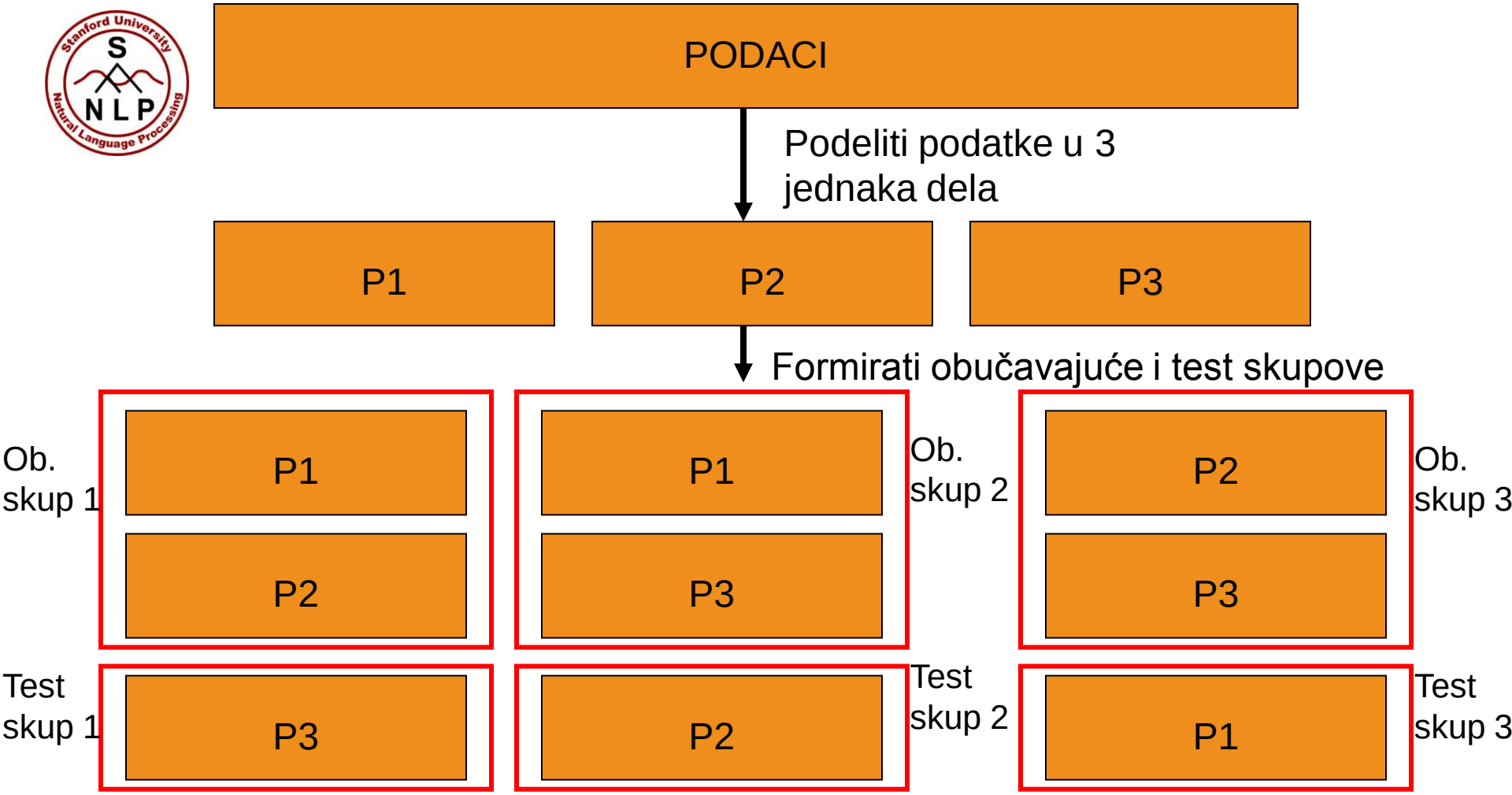
K-tostruka unakrsna validacija (*k-fold cross-validation*)

1. Podeliti obučavajući skup u *k* jednakih delova (*folds*).
2. Formirati sve moguće kombinacije delova na ovaj način:
   obučiti model na (k_1+...+k_{n-1}), testirati na k_n
   obučiti model na (k_1+...+k_{n-2}+k_n), testirati na k_{n-1}
   ....

Izračunati prosek performansi svih modela.

Obično se podela vrši na slučajan način.
Moguće je i koristi stratifikovano uzorkovanje kada želimo da odnos klasa koji je ubučavajućem skupu bude isti u svakom delu.
Npr. u celom skupu ima 2/3 pozitivne i 1/3 neg klase – tako onda uzorkujemo svaki deo.

Trostruka unakrsna validacija (*3-fold cross-validation*)



NB za klasifikaciju teksta
Problemi u praksi

Nedostatak obučavajućeg skupa?

Treba koristiti Ručno kreirana pravila

Ako sadrži reči "wheat" ili "grain", a ne sadrži "whole" ili "bread" onda je kategorija *grain*

- Moramo pažiljivo da kreiramo pravila
  - Treba nam bar neki manji označen skup da bi mogli da naštelujemo pravila
  - Kreiranje pravila je vremenski zahtevan posao

Jako mali obučavajući skup?

- Korisiti Naivni Bajesov model
  - NB ima veliki *bijas* i malu *varijansu*
  - Ova dva pojma ću grubo objasniti na predavanju, a u detalje ih učite na predmetu Mašinsko Učenje i SIAP
- Povećati obučavajući skup
  - „Ubediti“ ljude da vam označe podatke
- Probati metode polu-nadgledanog učenja:
  - Co-training itd. – više na predmetu MU

Relativno veliki obučavajući skup?

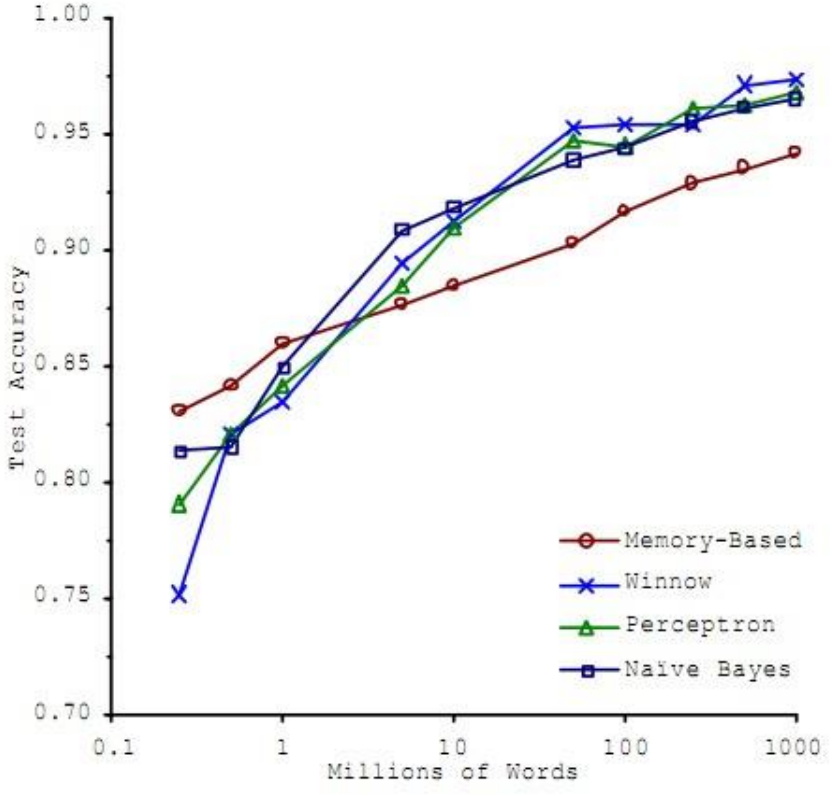
- Savršena postavka za kompleksnije modele
  - SVM
  - Regularized Logistic Regression
  - Oba modela ćemo objasniti kasnije tokom kursa
- Možete koristiti i stabla odlučivanja
  - Interpretabilan model
  - Lako je objasniti zašto je neki dokument dobio neku klasu

Jako veliki obučavajući skup?

- Možemo dobiti jako veliku tačnost!
- Cena je naravno brzina:
  - SVM (sporo obučavanje) ili kNN (spora primena na test skupu)
- Naivni Bajes je ovde odlična opcija jer je brz!
- Naravno *Deep Learning* je uvek odlična opcija kada imate puno podataka.

Tačnost kao funkcija veličine obučavajućeg skupa

- Sa velikim ob. skupom
  - Skoro da nije važno koji model koristite



Brill and Banko on spelling correction

Sistemi koji se koriste u praksi: često kombinuju ML modele i ručno kreirana pravila

- Automatska klasifikacija
- Analiza grešaka
- Ručno kerirati pravila da reše tipične greške tj. slučajeve koji su koji su teški za ML

U praksi imamo problem sa jako malim vrednostima (*underflow*) Rešenje: logaritmovanje

- Množenje puno verovatnoća može da rezultuje jako malim vrednostima.
- Pošto važi log(*xy*) = log(*x*) + log(*y*)
  - Bolje je sabrati logaritme verovatnoća nego množiti verovatnoće.
- Log je monotona funkcija pa se rezultati modela ne menjaju

$$c_{NB} = \underset{c_j \in C}{\operatorname{argmax}} \log P(c_j) + \sum_{i \in positions} \log P(x_i \mid c_j)$$

Kako doštelovati (tweak) performanse modela

- Osobine specifične za domen – nije isto kad klasifikujemo tvitove ili blogove, nije isto kad radimo sa naučim radovima ili *Blic* vestima
- Nekada moramo da normalizujemo neke delove:
  - Umesto 1234, 258 imamo BROJ, BROJ, slično za hemijske formule, …
- Nekada pomaže *Upweighting*: Neke reči brojimo kao da se pojavljuju dva puta, npr:
  - naslov (Cohen & Singer 1996)
  - prva rečenica svakog pasusa (Murata, 1999)
  - rečenice koje sadrže reči iz naslova (Ko *et al,* 2002)