

Uvod u obradu prirodnog jezika

6.1. Zadaci klasifikacije teksta

Branko Žitko

prevedeno od: Dan Jurafsky, Chris Manning

Je li ovo SPAM?

Subject: **Važna obavijest!**

From: PMF Split info@pmfst.hr

Date: 16. Svibanj, 2013 12:34:56

To: undisclosed-recipients

Sjajne vijesti!

Možete pristupiti najnovijim vijestima koristeći donji link za prijavu na forum Prirodoslovno-matematičkog fakulteta

<http://www.kontakt-forum.hr/forum/form-pmf-split.html>

Kliknite na gornji link da dobijete više informacija o ovom novom forumu. Također možete kopirati gornji link i prenijeti ga u Web preglednik i prijaviti se kako bi saznali više o ovoj novoj usluzi.

© Prirodoslovno-matematički fakultet

Pozitivna ili negativna kritika filma?

- nevjerojatno razočarenje 👎
- pun otkačenih likova i bogato primijenjena satira, s nekim velikim zapletima radnje 👍
- ovo je najveća ekscentrična komedija ikad snimljena 👍
- ovo je jadno, najgori dio je definitivno scena boksa 📵

Koja je tema ovog članka?

MEDLINE članak

MeSH - hijerarhija kategorija subjekta

- kemija
- krvotok
- terapija lijekovima
- embriologija
- epidemiologija
- ...



Klasifikacija teksta

- Pridruživanje kategorije, naslova ili žanra nekoj temi
- Detekcija spam-a
- Identifikacija autora
- Identifikacija dobi/spola
- Identifikacija jezika
- Analiza sentimenta
- ...

Klasifikacija teksta: definicija

- Ulaz:
 - dokument d
 - fiksni skup klasa $\mathcal{C} = \{c_1, c_2, \dots, c_K\}$
- Izlaz:
 - predviđena klasa $c \in \mathcal{C}$

Metode klasifikacije: ručno pisana pravila

- Pravila temeljena na kombinacijama riječi i drugih osobina
 - spam: crna-lista-adresa I LI ("Š" I "izabrani ste")
- Preciznost može biti velika
 - ako su pravila brižno pisana od strane eksperta
- Ali izgradnja i održavanje pravila je skupo

Metode klasifikacije: nadzirano strojno učenje

- Ulaz:
 - dokument d
 - fiksni skup klasa $C = \{c_1, c_2, \dots, c_K\}$
 - skup za trening N ručno označenih dokumenata $(d_1, c_1), \dots, (d_N, c_N)$
- Izlaz:
 - naučeni klasifikator $\gamma: d \rightarrow c$

Metode klasifikacije: nadzirano strojno učenje

- Bilo koja vrsta klasifikatora
 - Naivni Bayes
 - Logistička regresija
 - Stroj s potpornim vektorima
 - k-najbližih susjeda
 - ...

Uvod u obradu prirodnog jezika

6.2. Naivni Bayes (naive bayes)

Branko Žitko

prevedeno od: Dan Jurafsky, Chris Manning

Ideja naivnog Bayesa

- Jednostavna (naivna) metoda klasifikacije temeljena na Bayesovom pravilu
- Oslanja se na jednostavnu reprezentaciju teksta
 - vreća riječi (bag of words)

Reprezentacija vreće riječi

Y(

Volim ovaj film! Sladak je, ali sa satiričnim humorom. Dijalog je super i pustolovne scene su zabavne... Uspijeva biti hirovit i romantičan, iako ismijava konvencije žanra bajke. Ja bih ga preporučio bilo kome. Vidio sam ga nekoliko puta i uvijek se radujem vidjeti ga ponovno kadgod imam prijatelja koji ga još nije vidio.

)=C



Reprezentacija vreće riječi

Y(

Volim ovaj film! **Sladak** je, ali sa **satiričnim** humorom. Dijalog je **super** i pustolovne scene su **zabavne...** Uspijeva biti **hirovit** i **romantičan**, iako **ismijava** konvencije žanra bajke. Ja bih ga **preporučio** bilo kome. Vidio sam ga **nekoliko** puta i uvijek se radujem vidjeti ga **ponovno** kadgod imam prijatelja koji ga još nije vidio.

)=C



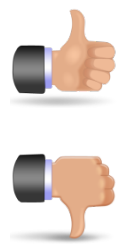
Reprezentacija vreće riječi

Y(

Volim ----- Sladak -----
satiričnim -----
super -----
zabavne ----- hirovit -
romantičan ----- ismijava -----

preporučio -----
nekoliko -----
----- ponovno -----

)=C





Reprezentacija vreće riječi

$Y($

volim	2
sladak	2
preporučio	1
ismijava	1
super	1
...	...

$) = C$

Uvod u obradu prirodnog jezika

6.3. Formalizacija naivnog Bayesovog klasifikatora

Branko Žitko

prevedeno od: Dan Jurafsky, Chris Manning

Bayesovo pravilo primijenjeno na dokumente i klase

- Za dokument d i klasu c

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

Naivni Bayesov klasifikator

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

MAP
=
Maximum a posteriori
=
najvjerojatnija klasa

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

Bayesovo
pravilo

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

Izbacivanje
nazivnika

Naivni Bayesov klasifikator

izglednost

priori

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

$$c_{MAP} = \underset{c \in \mathcal{C}}{\operatorname{argmax}} P(f_1, f_2, \dots, f_n|c)P(c)$$

Dokument d prikazan
kao skup osobina
 f_1, f_2, \dots, f_n

Naivni Bayesov klasifikator

izglednost

prior

$$c_{MAP} = \underset{c \in C}{argmax} P(f_1, f_2, \dots, f_n | c) P(c)$$

- Koliko često se klasa c pojavljuje
 - možemo izračunati relativne frekvencije u korpusu
- Kako odrediti izglednost od d i osobina f_1, f_2, \dots, f_n
 - procjena svih mogućih kombinacija osobina bi zahtijevala veliki broj parametara i ogromni skupove za treniranje
 - npr. svi mogući skupovi riječi i pozicija tih riječi
- Stoga koriste se dvije pojednostavljajuće pretpostavke

$$P(f_1, f_2, \dots, f_n | c)$$

- **Pretpostavka vreće riječi**

- pozicija nije važna
- stoga f_1, f_2, \dots, f_n kodira identitet riječi, a ne njen položaj

- **Uvjetna nezavisnost**

- vjerojatnost osobina $P(f_i | c_j)$ su nezavisne za danu klasu c

$$P(f_1, f_2, \dots, f_n | c) = P(f_1 | c) \times P(f_2 | c) \times \dots \times P(f_n | c)$$

Naivni Bayesov klasifikator

$$c_{MAP} = \underset{c \in \mathcal{C}}{\operatorname{argmax}} P(f_1, f_2, \dots, f_n | c) P(c)$$

$$c_{NB} = \underset{c \in \mathcal{C}}{\operatorname{argmax}} P(c) \prod_{f \in F} P(f | c)$$

Primjena naivnog Bayesovog klasifikatora

pozicije \leftarrow sve pozicije riječi u testnom dokumentu

$$c_{NB} = \underset{c \in C}{\operatorname{argmax}} P(c) \prod_{i \in \textit{pozicije}} P(w_i | c)$$

$$c_{NB} = \underset{c \in C}{\operatorname{argmax}} \left(\log P(c) + \sum_{i \in \textit{pozicije}} \log P(w_i | c) \right)$$

Uvod u obradu prirodnog jezika

6.4. Učenje naivnog Bayesa

Branko Žitko

prevedeno od: Dan Jurafsky, Chris Manning

Učenje naivnog Bayesovog modela

- Kako naučiti vjerojatnosti $P(c)$ i $P(f_i|c)$?
- Prvi pokušaj: procjena maksimalne izglednosti (MLE)
 - jednostavno koristi frekvencije podataka

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

Broj dokumenata klase c

Broj svih dokumenata

$$\hat{P}(w_i|c) = \frac{broj(w_i, c)}{\sum_{w \in V} broj(w, c)}$$

Broj pojavljivanja riječi w_i u svim dokumentima klase c

Broj pojavljivanja riječi w u svim dokumentima klase c

V - sve riječi iz svih dokumenata

Procjena parametara

- Koliko puta se riječ w_i pojavljuje među svim riječima u dokumentu klase c

$$\hat{P}(w_i|c) = \frac{broj(w_i, c)}{\sum_{w \in V} broj(w, c)}$$

- Kreira se mega-dokument za klasu c tako što se povežu svi dokumenti klase c
 - koristi se frekvencija riječi w iz mega-dokumenta

Problem kod maksimalne izglednosti

- Što ako nemamo niti jedan dokument za treniranje s riječju "fantastično" koja je klasificirana za klasu pozitivno?

$$\hat{P}(\text{"fantastično"}|\text{poz}) = \frac{\text{broj}(\text{"fantastično"}, \text{poz})}{\sum_{w \in V} \text{broj}(w, \text{poz})} = 0$$

- Nulte vjerojatnosti se ne mogu izbjeći

$$c_{NB} = \underset{c \in \{\text{poz}, \text{neg}\}}{\operatorname{argmax}} P(\text{poz}) \prod_{i \in \text{pozicije}} P(w_i | c)$$

Laplace (dodaj 1) izgladivanje za naivnog Bayesa

$$\begin{aligned}\hat{P}(w_i|c) &= \frac{broj(w_i, c) + 1}{\sum_{w \in V} (broj(w, c) + 1)} \\ &= \frac{broj(w_i, c) + 1}{\sum_{w \in V} broj(w, c) + |V|}\end{aligned}$$

$$\begin{aligned}\hat{P}(w_u|c) &= \frac{\textit{broj}(w_u, c) + 1}{\sum_{w \in V} \textit{broj}(w, c) + |V| + 1} \\ &= \frac{1}{\textit{broj}(w, c) + |V| + 1}\end{aligned}$$

- Nepoznate riječi se mogu ignorirati
 - Ako riječ w_u iz testnog skupa nije u rječniku V onda se w_u ignorira

Naivni Bayes: Učenje

UČENJE(D, C)

za svaku klasu $c \in C$

$N_{doc} \leftarrow$ broj dokumenata iz D

$N_c \leftarrow$ broj dokumenata iz D klase c

$prior[c] \leftarrow \frac{N_c}{N_{doc}}$

$V \leftarrow$ riječnik dokumenata iz D

$megadoc[c]$ proširi s d za $d \in D$ koji su klase c

za svaku riječ $w \in V$

$broj[w, c] \leftarrow$ broj pojavljivanja od w u $megadoc(c)$

$izvjesnost[w, c] \leftarrow \frac{broj[w, c] + 1}{\sum_{w' \in V} (broj[w', c] + 1)}$

vrați $prior, izvjesnost, V$

Naivni Bayes: Testiranje

TESTIRANJE (*testdoc*, *prior*, *izvjesnost*, C , V)

za svaku klasu $c \in C$

$posterior[c] \leftarrow prior[c]$

za svaku poziciju i iz *testdoc*

$w \leftarrow testdoc[i]$

ako $w \in V$

$posterior[c] = posterior[c] * izvjesnost[w, c]$

vрати $\operatorname{argmax}_{c \in C} posterior[c]$

Uvod u obradu prirodnog jezika

6.5. Odnos s modelom jezika

Branko Žitko

prevedeno od: Dan Jurafsky, Chris Manning

Naivni Bayes i model jezika

- Naivni Bayesov klasifikator može koristiti bilo koje osobine
 - URL, email adresa, rječnici, svojstva mreže
- Ali ako
 - koristimo **samo** riječi kao osobine
 - koristimo **sve** riječi iz teksta (ne iz podskupa teksta)
- onda
 - Naivni Bayes ima velike sličnosti s modelom jezika

Svaka klasa je unigram

- Svakoj riječi w se pridružuje $P(w|c)$
- Svakoj rečenici s se pridružuje $P(s|c) = \prod P(w|c)$

Klasa = poz	
0.1	Ja
0.1	volim
0.01	ovaj
0.05	novi
0.1	film

$$P(s|\text{poz}) = 0.1 * 0.1 * 0.01 * 0.05 * 0.01 = 0.0000005$$

Svaka klasa je unigram

- Koja klasa pridružuje veću vjerojatnost rečenici s ?

Klasa = poz		Klasa = neg	
0.1	Ja	0.2	Ja
0.1	volim	0.001	volim
0.01	ovaj	0.01	ovaj
0.05	novi	0.005	novi
0.1	film	0.1	film

$$P(s|\text{poz}) = 0.1 * 0.1 * 0.01 * 0.05 * 0.01 = 0.0000005$$

$$P(s|\text{neg}) = 0.2 * 0.001 * 0.01 * 0.005 * 0.1 = 0.00000001$$

$$P(s|\text{poz}) > P(s|\text{neg})$$

Uvod u obradu prirodnog jezika

6.6. Multinominalni naivni Bayes: Radni primjer

Branko Žitko

prevedeno od: Dan Jurafsky, Chris Manning

Naivni Bayes i model jezika

	Dokument	Riječi	Klasa
Treniranje	d ₁	Italija Rim Italija	IT
	d ₂	Italija Italija Firenca	IT
	d ₃	Italija Ankona	IT
	d ₄	Pariz Francuska Italija	FR
Test	d ₅	Italija Italija Italija Pariz Francuska	?

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

$V = \{\text{Italija, Rim, Firenca, Ankona, Francuska, Pariz}\}$

$$\hat{P}(w_i|c) = \frac{broj(w_i, c) + 1}{\sum_{w \in V} broj(w, c) + |V|}$$

Prior

Uvjetna vjerojatnost

Izbor klase

Naivni Bayes i model jezika

	Dokument	Riječi	Klasa
Treniranje	d ₁	Italija Rim Italija	IT
	d ₂	Italija Italija Firenca	IT
	d ₃	Italija Ankona	IT
	d ₄	Pariz Francuska Italija	FR
Test	d ₅	Italija Italija Italija Pariz Francuska	?

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

$V = \{\text{Italija, Rim, Firenca, Ankona, Francuska, Pariz}\}$

$$\hat{P}(w_i|c) = \frac{\text{broj}(w_i, c) + 1}{\sum_{w \in V} \text{broj}(w, c) + |V|}$$

Prior

$$P(\text{IT}) = \frac{3}{4}$$

$$P(\text{FR}) = \frac{1}{4}$$

Uvjetna vjerojatnost

$$P(\text{Italija}|\text{IT}) = \frac{5+1}{8+6} = \frac{6}{14} = \frac{3}{7}$$

$$P(\text{Pariz}|\text{IT}) = \frac{0+1}{8+6} = \frac{1}{14}$$

$$P(\text{Francuska}|\text{IT}) = \frac{0+1}{8+6} = \frac{1}{14}$$

$$P(\text{Italija}|\text{FR}) = \frac{1+1}{3+6} = \frac{2}{9}$$

$$P(\text{Pariz}|\text{FR}) = \frac{1+1}{3+6} = \frac{2}{9}$$

$$P(\text{Francuska}|\text{FR}) = \frac{1+1}{3+6} = \frac{2}{9}$$

Izbor klase

$$P(\text{IT}|\text{d}_5) \propto \frac{3}{4} \cdot \left(\frac{3}{7}\right)^3 \cdot \frac{1}{14} \cdot \frac{1}{14} \approx 0.0003$$

$$P(\text{FR}|\text{d}_5) \propto \frac{1}{4} \cdot \left(\frac{2}{9}\right)^3 \cdot \frac{2}{9} \cdot \frac{2}{9} \approx 0.0001$$

Naivni Bayes nije baš toliko naivan!

- Veoma brz, malo prostora zauzima
- robustan na nevažne osobine
 - nevažne osobine se međusobno poništavaju ne utječući na rezultat
- Dobar kod domena s mnogo jednako važnih osobina
 - za razliku od stabla odluke koja pate od fragmentacije – pogotovo kod malo podataka
- Optimalan ako stoji pretpostavka o nezavisnosti: ako je pretpostavljena nezavisnost točna, onda se radi o optimalnom Bayesovom klasifikatoru
- dobra ovisna osnova za klasifikaciju teksta
- Postoje i drugi, precizniji klasifikatori

Uvod u obradu prirodnog jezika

6.7. Preciznost, odziv i F mjera (Precision, Recall and F measure)

Branko Žitko

prevedeno od: Dan Jurafsky, Chris Manning

2 za 2 tablica slučaja

- 2 skupa podataka
 - točni entiteti
 - odabrani entiteti
- 4 moguća slučaja
 - TP – stvarno pozitivni (True positives)
 - FP – lažno pozitivni (False Positives)
 - FN – lažno negativni (False Negatives)
 - TN – stvarno negativni (True Negatives)

točni entiteti

		točni entiteti	
		točno	nije točno
odabrani entiteti	odabrano	TP	FP
	nije odabrano	FN	TN

2 za 2 tablica slučaja: primjer

- Primjer
 - TP – sustav je točno rekao za spam da je spam
 - FP – sustav je pogrešno rekao za ne-spam da je spam
 - FN – sustav je pogrešno rekao za spam da je ne-spam
 - TN – sustav je točno rekap ne-spam da je ne-spam

	spam	ne-spam
spam	TP	FP
nije spam	FN	TN

Točnost

- Točnost ($Acc = Accuracy$) kao mjera

$$Acc = \frac{TP+TN}{TP+FP+TN+FN}$$

	točno	nije točno
odabrano	TP	FP
nije odabrano	FN	TN

Točnost

- Točnost kao mjera nije dobra za mali skup točnih podataka
- Recimo da promatramo 100000 Web stranica i samo 10 njih opisuje marku cipela.
- Ako napravimo najjednostavniji klasifikator koji za svaku stranicu kaže da ne opisuje marku cipela, onda ćemo dobiti veliku točnost

$$Acc = \frac{TP+TN}{TP+FP+FN+TN} =$$

$$\frac{0+99990}{0+0+10+99990} = \frac{99990}{100000} = 99.99\%$$

	marka cipela	ostalo
odabrano	0	0
nije odabrano	10	99990

Preciznost i odziv

- **Preciznost P** : % odabranih elemenata koji su točni
- **Odziv R** : % točnih elemenata koji su odabrani

	točno	nije točno
odabrano	TP	FP
nije odabrano	FN	TN

$$P = \frac{TP}{TP + FP}$$

$$R = \frac{TP}{TP + FN}$$

Preciznost i odziv

$$R = \frac{0}{10} = 0\%$$

	marka cipela	ostalo
odabrano	TP = 0	FP = 0
nije odabrano	FN = 10	TN = 99990

$$P = \frac{10}{40} = 25\%$$
$$R = \frac{10}{10} = 100\%$$

	marka cipela	ostalo
odabrano	TP = 10	FP = 30
nije odabrano	FN = 0	TN = 99960

Kombinirana mjera: F

- **F mjera:** Kombinirana mjera koja procjenjuje Preciznost/Odziv je (težinska harmonijska sredina)

$$F = \frac{1}{\alpha \frac{1}{P} + (1 - \alpha) \frac{1}{R}} = \frac{(\beta^2 + 1)PR}{\beta^2 P + R}$$

- Harmonijska sredina je konzervativni prosjek
- Obično se koristi balansirana F1 mjera

$$\text{za } \beta = 1 \text{ (odnosno, } \alpha = \frac{1}{2} \text{) } F1 = \frac{2PR}{P+R}$$

Uvod u obradu prirodnog jezika

6.8. Evaluacija

Branko Žitko

prevedene od: Dan Jurafsky, Chris Manning

Više od dvije klase: skupovi binarnih klasifikatora

- "Bilo koja" viševrijednosna klasifikaciju
 - dokument može pripadati 0, 1, ili više klasa
- Za svaku klasu $c \in C$
 - napravi klasifikator γ_c kako bi razlikovali c od drugih klasa $c' \in C$
- Za dani testni dokument d ,
 - Evaluiraj pripadnost za svaku klasu koristeći svaku γ_c
 - d pripada svakoj klasi za koju γ_c vraća istinu

Više od dvije klase: skupovi binarnih klasifikatora

- "Jedna od" - **više vrijednosna** klasifikacija
 - Klase su međusobne isključive: svaki dokument pripada točno jednoj klasi
- Za svaku klasu $c \in C$
 - napravi klasifikator γ_c kako bi razlikovali c od drugih klasa $c' \in C$
- Za dani testni dokument d ,
 - Evaluiraj pripadnost za svaku klasu koristeći svaku γ_c
 - d pripada jednoj klasi za koju γ_c **vraća najveću vjerojatnost**

Evaluacija: jedna od - viševrijednosna klasifikacija

- Kategorizacija maila u 3 klase: Hitno, Normalno, Spam

		<i>Zlatni standard</i>				
		Hitno	Normalno	Spam		
<i>Sustav</i>	Hitno	8	10	1	$P_H = \frac{8}{8+10+1}$	
	Normalno	5	60	50	$P_N = \frac{60}{5+60+50}$	
	Spam	3	30	200	$P_S = \frac{200}{3+30+200}$	
		$R_H = \frac{8}{8+5+3}$	$R_N = \frac{60}{10+60+30}$	$R_S = \frac{200}{1+50+200}$		

Evalvacija: jedna od - viševrijednosna klasifikacija

- Ako imamo više od jedne klase, kako se kombiniraju mjere u jednu mjeru?

	H	N	S
H	8	10	1
N	5	60	50
S	3	30	200

- Makro-prosjek:** izračunaj performanse za svaku klasu i onda prosjek

Hitno			Normalno			Spam		
	H	ne H		N	ne N		S	ne S
H	8	11	N	60	55	S	200	33
ne H	5	360	ne N	40	212	ne S	51	83
$P = \frac{8}{8+11} = 0.42$			$P = \frac{60}{60+55} = 0.52$			$P = \frac{200}{200+33} = 0.86$		

$$makroP = \frac{0.42+0.52+0.86}{3} = 0.60$$

Evaluacija: jedna od - viševrijednosna klasifikacija

- Ako imamo više od jedne klase, kako se kombiniraju mjere u jednu mjeru?

Hitno			Normalno			Spam		
	H	ne H		N	ne N		S	ne S
H	8	11	N	60	55	S	200	33
ne H	8	360	ne N	40	212	ne S	51	83

- Mikro-prosjek:** prikupi performanse svake klase, izračunaj tablicu slučaja, evaluiraj

	da	ne
da	268	99
ne	99	635

$$mikroP = \frac{268}{268+99} = 0.73$$

Razvojni testni skupovi i unakrsna validacija

Trening skup

Razvojni skup

Testni skup

- Mjera: $P/R/F1$ ili Acc
- Neviđeni testni skup
 - izbjeći prekoračenja (ugađanje prema razvojnom skupu)
 - ponekad je testni skup (ili razvojni skup) malen
- Unakrsna validacija (cross validation) nad višestrukim podjelama
 - rukovanje greškama uzorkovanja nad više skupova podataka
 - skupljanje rezultata za svaku podjelu
 - izračunati prosjek rezultata

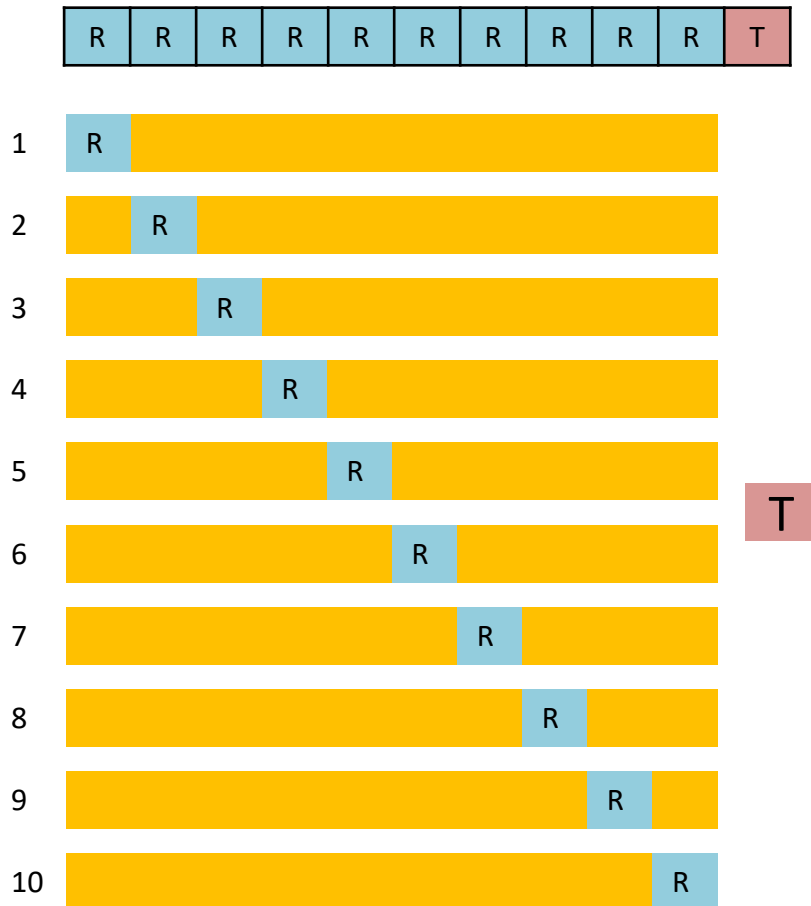
Razvojni testni skupovi i unakrsna validacija

Trening skup

Razvojni
skup

Testni skup

10-struka unakrsna validacija (10-fold cross validation)



Uvod u obradu prirodnog jezika

6.9. Testiranje statističke značajnosti

Branko Žitko

prevedene od: Dan Jurafsky, Chris Manning

Testiranje statističke značajnosti

Kako znamo koji klasifikator je bolji?

Za dane:

- Klasifikatore A i B
- Metriku M : $M(A, x)$ je performansa od A na testnom skupu x
- $\delta(x) = M(A, x) - M(B, x)$
- Želimo znati je li $\delta(x) > 0$ (A je bolji od B)
- $\delta(x)$ – **veličina učinka** (effect size)
- Ako se pokaže $\delta(x)$ pozitivnim, to može biti slučajnost samo za ovaj testni skup x .

Testiranje hipoteze

Dvije hipoteze

- nul-hipoteza: A nije bolji od B $H_0: \delta(x) \leq 0$
- hipoteza: A je bolji od B $H_1: \delta(x) > 0$
- Želimo isključiti H_0
- Stvori se slučajna varijabla X za sve testne skupove
- Pitamo, koliko je vjerojatno, ako je H_0 istina, da ćemo među testnim skupovima naići na vrijednost $\delta(x)$ koju promatramo.
- Formaliziramo kao p-vrijednost:
$$P(\delta(X) \geq \delta(x) \mid H_0 \text{ je istina})$$

Testiranje hipoteze

$P(\delta(X) \geq \delta(x) \mid H_0 \text{ je istina})$

p-vrijednost je vjerojatnost da ćemo naići na $\delta(x)$ uz pretpostavku da A nije bolji od B .

Ako je $\delta(x)$ velik (A ima $F1 = 0.9$, B ima $F1 = 0.2$)

- to bi bilo iznenađenje
- uz činjenicu da je H_0 istinita.
(p-vrijednost niska)

Ako je $\delta(x)$ malen (A ima $F1 = 0.2$, B ima $F1 = 0.9$)

- to nebi bilo iznenađenje
- uz pretpostavku da je H_0 istinita i
- da A zaista nije bolji od B .
(p-vrijednost visoka)

Testiranje hipoteze

$$P(\delta(X) \geq \delta(x) \mid H_0 \text{ je istina})$$

Vrlo mala p-vrijednost znači da je razlika koju smo uočili malo vjerojatna pod nul-hipotezom.

Odbacujemo nul-hipotezu.

Veoma mala: 0.05 ili 0.01

Rezultat (A je bolji od B) je **statistički značajan** ako promatrani δ ima vjerojatnost koja je manja od praga.

Testiranje hipoteze

Koristi se neparametarsko testiranje temeljeno na uzorkovanju:

- umjetno stvaramo mnogo verzija postavki eksperimenta

Na primjer:

- kreiramo veliki broj testnih skupova x'
- za svaki izračunamo $\delta(x')$
- dobivamo distribuciju
- odaberemo prag (npr. 0.01)
- ako za 99% testnih skupova vrijedi $\delta(x) > \delta(x')$
- onda zaključujemo da je δ našeg testnog skupa prava δ , a ne umjetna

Upareno Bootstrap testiranje

Može se primijeniti na bilo koju metriku (Acc, P, R, F1)

Bootstrap znači iterativno uzimati veliki broj malih uzoraka sa zamjenom iz originalnog velikog uzorka.

Upareno Bootstrap testiranje

Jednostavan primjer:

Klasifikacija teksta s testnim skupom x od 10 dokumenata

Rezultati sustava A i B nad x s 4 moguća ishoda:

AB - oboje točno

AB - oboje pogrešno

AB - A točan, B pogrešan

AB - A pogrešan, B točan

	1	2	3	4	5	6	7	8	9	10	A%	B%	δ
x	AB	AB	AB	AB	AB	AB	AB	AB	AB	AB	0.70	0.50	0.20

Upareno Bootstrap testiranje

Imamo distribuciju!

Možemo provjeriti koliko često A ima **slučajnu** prednost

Uz pretpostavku H_0 očekujemo $\delta(x^{(i)}) = 0$

Prebrojimo koliko puta $\delta(x^{(i)})$ prelazi 0 u odnosu na $\delta(x)$

$$\text{p-value} = \frac{1}{b} \sum_{i=1}^b 1(\delta(x^{(i)}) - \delta(x) \geq 0)$$

Upareno Bootstrap testiranje

Međutim, uzorke nismo izvlačili iz distribucije čija je srednja vrijednost 0.

Koristili smo originalni testni skup x koji je pristran (0.20) u korist sustava A.

p-vrijednost stoga računamo koliko često $\delta(x^{(i)})$ premašuje očekivanu vrijednost $\delta(x)$ s $\delta(x)$ ili više:

$$\text{p-value} = \frac{1}{b} \sum_{i=1}^b 1 \left(\delta(x^{(i)}) - \delta(x) \geq \delta(x) \right)$$

$$\text{p-value} = \frac{1}{b} \sum_{i=1}^b 1 \left(\delta(x^{(i)}) \geq 2\delta(x) \right)$$