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.

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

Syntactic frame and verb bias in aphasia: Plausibility judgments of undergoer-subject sentences

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 $C = \{c_1, c_2, \ldots, c_j\}$
- Izlaz:
  - predviđena klasa $c \in C$

Metode klasifikacije: ručno pisana pravila

- Pravila temeljena na kombinacijama riječi i drugih osobina
  - spam: crna-lista-adresa ILI ("$" 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 $C = \{c_1, c_2, \ldots, c_j\}$
  - skup za trening m ručno označenih dokumenata $(d_1, c_1), \ldots, (d_m, c_m)$
- 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

$$Y(\text{Volim ---------- Sladak ----------}\\ \text{satiričnim ----------}\\ \text{super ----------}\\ \text{zabavne ---------- hirovit -}\\ \text{romantičan ----- ismijava --------}\\ \text{----------}\\ \text{preporučio ----------}\\ \text{nekoliko ----------}\\ \text{---------- ponovno ----------}\\ \text{----------})=c$$

Reprezentacija vreće riječi

$Y($

volim	2
sladak	2
preporučio	1
ismijava	1
super	1
…	…

$)=c$

Vreća riječi kod klasifikacije dokumenta

Testni dokument

parser
jezik
oznaka
prijevod
…

?

Obrada prirodnog jezika

parser
oznaka
treniranje
prijevod
jezik
…

Strojno učenje

učenje
treniranje
algoritam
skupljanje
mreža
…

Obrada prirodnog jezika

parser
oznaka
treniranje
prijevod
jezik
…

Objektno orijentirano programiranje

klasa
metoda
atribut
objekt
model
…

Planiranje

planiranje
zaključivanje
vrijeme
plan
jezik…
…

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

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

Naivni Bayesov klasifikator

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

MAP
=
Maximum a posteriori
=
najvjeroratnija klasa

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

Bayesovo pravilo

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

Izbacivanje nazivnika

Naivni Bayesov klasifikator

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

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

Dokument d prikazan kao skup osobina $x_1 \ldots x_n$

Naivni Bayesov klasifikator

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

vjerodostojnost

prior

- Koliko često se klasa c pojavljuje
  - možemo izračunati relativne frekvencije u korpusu
- Kako odrediti vjerodostojnost od d i osobina $x_1 \ldots x_n$
  - $O(|X|^n * |C|)$ parametara
  - Može se samo procijeniti ako imamo veliki broj primjera za testiranje

Multinominalna naivna Bayesova pretpostavka nezavisnosti

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

- **Pretpostavka kod vreće riječi**
  - pozicija nije važna
- **Uvjetna nezavisnost**
  - vjerojatnost osobina $P(x_i \mid c_j)$ su nezavisne za danu klasu c

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

Multinominalni naivni Bayesov klasifikator

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

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

Primjena multinominalnog naivnog Bayesovog klasifikatora

pozicije ← sve pozicije riječi u testnom dokumentu

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

Uvod u obradu prirodnog jezika

6.4. Učenje naivnog Bayesa

Branko Žitko

prevedeno od: Dan Jurafsky, Chris Manning

Učenje multinominalnog naivnog Bayesovog modela

- Prvi pokušaj: procjena maksimalne vjerodostojnosti
  - jednostavno koristi frekvencije podataka

$$\hat{P}(c_j) = \frac{broj\_dokumenata(C = c_j)}{N_{dokument}}$$

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

Procjena parametara

- Koliko puta se riječ w_i pojavljuje među svim riječima u dokumentu teme c_j

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

- Kreira se mega-dokument za temu *j* tako što se povežu svi dokumenti teme *j*
  - koristi se frekvencija od w iz mega-dokumenta
  - npr. $D_{pozitivno}$, $D_{negativno}$

Problem kod maksimalne vjerodostojnosti

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

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

- Nulte vjerojatnosti se ne mogu izbjeći

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

Laplace (dodaj 1) izglađivanje za naivnog Bayesa

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

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

Multinominalni naivni Bayes: Učenje

- Iz korpusa za treniranje izvuci *Rječnik = V*
- Izračunaj $P(c_j)$
  - za svaku klasu c_j u C
    - $dokument_j \leftarrow$ svi dokumenti klase cj

$$P(c_j) \leftarrow \frac{|dokument_j|}{|\text{ukupan broj dokumenata}|}$$

- Izračunaj $P(w_k|c_j)$
  - $Tekst_j \leftarrow$ mega-dokument koji sadrži sve $dokument_j$
  - za svaku riječ w_k iz *V*
    - $n_k \leftarrow$broj pojavljivanja od w_k u $Tekst_j$

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

Laplace (dodaj 1) izglađivanje: nepoznate riječi

- Ubaci jednu dodatnu riječ u rječnik, "nepoznata riječ" w_u

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

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

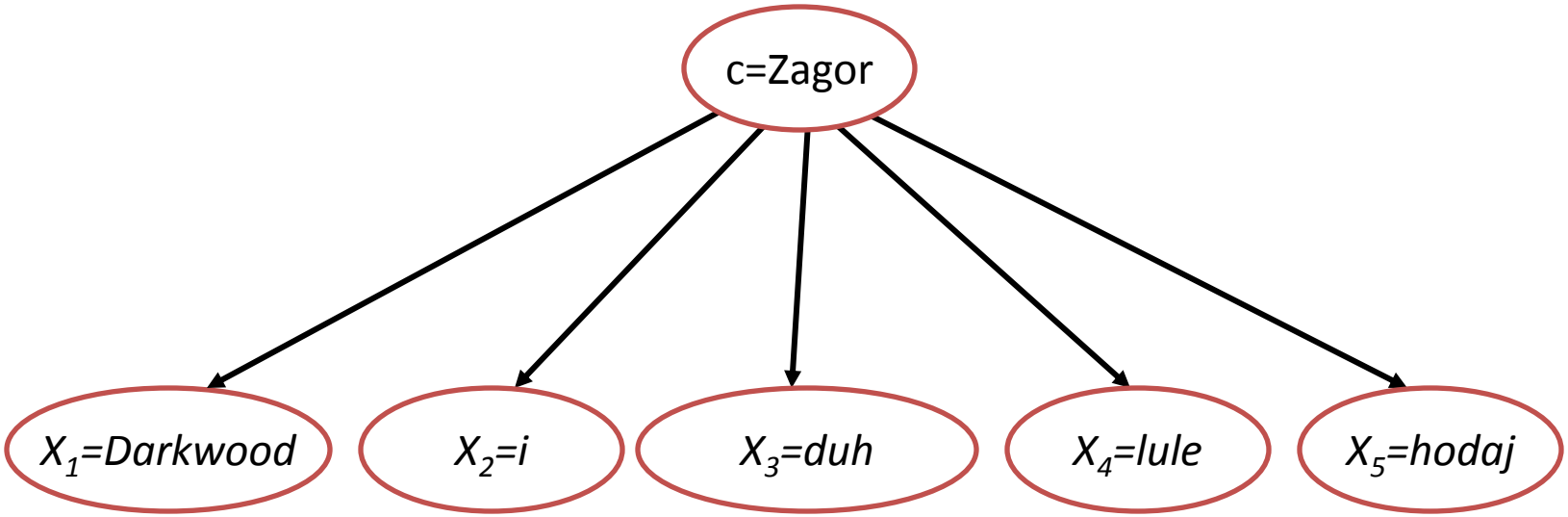
Uvod u obradu prirodnog jezika

6.5. Odnos s modelom jezika

Branko Žitko

prevedeno od: Dan Jurafsky, Chris Manning

Generativni model multinominalnog naivnog Bayesa



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 se pridružuje P(w|c)
- Svakoj rečenici se pridružuje P(s|c)=∏P(w|c)

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

P(s|pozitivno)=0.1*0.1*0.01*0.05*0.01=0.0000005

Svaka klasa je unigram

- Koja klasa pridružuje veću vjerojatnost rečenici s?

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

Klasa = negativno	
0.2	Ja
0.001	volim
0.01	ovaj
0.005	novi
0.1	film

P(s|pozitivno)=0.1*0.1* 0.01*0.05 *0.01=0.0000005

P(s|negativno)=0.2*0.001*0.01*0.005*0.1 =0.00000001

P(s|pozitivno)>P(s|negativno)

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	i
	d_2	Italija Italija Firenca	i
	d_3	Italija Ankona	i
	d_4	Pariz Francuska Italija	f
Test	d_5	Italija Italija Italija Pariz Francuska	?

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

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

Prior

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

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

Uvjetna vjerojatnost

$$P(Italija|i) = (5+1)/(8+6) = 6/14 = 3/7$$

$$P(Pariz|i) = (0+1)/(8+6) = 1/14$$

$$P(Francuska|i) = (0+1)/(8+6) = 1/14$$

$$P(Italija|f) = (1+1)/(3+6) = 2/9$$

$$P(Pariz|f) = (1+1)/(3+6) = 2/9$$

$$P(Francuska|f) = (1+1)/(3+6) = 2/9$$

Izbor klase

$$P(i|d_5) \propto 3/4 \times (3/7)^3 \times 1/14 \times 1/14 \approx 0.0003$$

$$P(f|d_5) \propto 1/4 \times (2/9)^3 \times 2/9 \times 2/9 \approx 0.0001$$

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 razliko 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, opoziv i F mjera (Precission, recall and F measure)

Branko Žitko

prevedeno od: Dan Jurafsky, Chris Manning

2 za 2 tablica slučaja

- 2 skupa
  - točni entiteti (i-istinito, l-lažno)
  - odabrani entiteti (p-pozitivno, n-negativno)
- 4 moguća slučaja
  - IP– istina pozitivna
  - LP – laž pozitivna
  - LN – laž negativna
  - IN – istina negativna

	točno	nije točno
odabrano	IP	LP
nije odabrano	LN	IN

2 za 2 tablica slučaja: primjer

- Primjer
  - IP– sustav je za spam rekao da je spam
  - LN – sustav je za ne-spam rekao da nije spam
  - LP – sustav je za ne-spam rekao da je spam
  - IN – sustav je za spam rekao da nije spam

	spam	ne-spam
spam	IP	LP
nije spam	IN	LN

Točnost

- Točnost kao mjera

$$Tocno = \frac{IP + LN}{IP + LP + IN + LN}$$

	točno	nije točno
odabrano	IP	LP
nije odabrano	IN	LN

Točnost

- Točnost kao mjera nije dobra za mali skup točnih podataka

$$\text{Tocno} = \frac{99990}{100000} = 99.99\%$$

	marka cipela	ostalo
odabrano	IP	LP
nije odabrano	nl = 10	ni = 99990

Preciznost i opoziv

- **Preciznost**: % odabranih elemenata koji su točni
- **Opoziv**: % točnih elemenata koji su odabrani

$$\text{Preciznost} = \frac{IP}{IP + LP}$$

$$\text{Opoziv} = \frac{IP}{IP + IN}$$

	marka cipela	ostalo
odabrano	IP	LP
nije odabrano	IN	LN

Preciznost i opoziv

$$\text{Preciznost} = \frac{0}{0+10000} = 0$$

$$\text{Opoziv} = \frac{0}{0+10} = 0$$

	marka cipela	ostalo
odabrano	IP = 0	LP =100000
nije odabrano	IN = 10	LN = 99990

$$\text{Preciznost} = \frac{8}{8+32} = 20\%$$

$$\text{Opoziv} = \frac{8}{8+40} = 16.66\%$$

	marka cipela	ostalo
odabrano	IP = 8	LP =32
nije odabrano	IN = 40	LN = 99960

Kombinirana mjera: F

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

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

- Harmonijska sredina je konzervativni prosjek
- Obično se koristi balansirana F1 mjera
  - za $\beta = 1$ (odnosno, $\alpha = ½$):
  - F1 = 2PO/(P+O)

Uvod u obradu prirodnog jezika

6.8. Evaluacija

Branko Žitko

prevedene od: Dan Jurafsky, Chris Manning

Više od dvije klase: skupovi binarnih klasifikatora

- Kako provesti više-vrijednosnu 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 članstvo za svaku klasu koristeći svaku γ_c
  - d pripada svakoj klasi za koju γ_c vraća istinu

Više od dvije klase: skupovi binarnih klasifikatora

- Multinominalna klasifikacija
  - Klase su međusobne isključive: svaki dokument pripada točne 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 članstvo za svaku klasu koristeći svaku γ_c
  - d pripada jednej klasi za koju γ_c vraća istinu

Evaluacija: klasični Reuters-21578 skup podataka

- Većina korištenih skupova podataka, 21578 dokumenata (svaki 90 tipova, 200 pojavnica)
- 9603 skupova za trening, 3299 testnih članaka
- 118 kategorija
  - članak može pripadati u više kategorija
  - nauči 118 binarnih kategorija
- Prosječni dokument (od više kategorija) ima 1.24 klase
- Samo 10 od 118 kategorija su velike

Učestale kategorije

(#trening, #test)

- 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)

Reuters skup podataka za kategorizaciju teksta

<REUTERS TOPICS="da" 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 technelogy in all areas of the industry, the NPPC added. Reuter

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

Matrica konfuzije c

- Za svaki par klasa <c1, c2> koliko dokumenata klase c1 su netočne pridruženih klasi c2?
- <c3,c2>: 90 dokumenata o pšenici netočne pridruženih peradi

Dokumenti testneg skupa	Pridružen UK	Pridružen perad	Pridružen pšenica	Pridružen kava	Pridružen interes	Pridružen trgovina
Točne UK	95	1	13	0	1	0
Točne perad	0	1	0	0	0	0
Točne pšenica	10	90	0	1	0	0
Točne kava	0	0	0	34	3	7
Točne interes	-	1	2	13	26	5
Točne trgovina	0	0	2	14	5	10

Mjere po klasi

Opoziv

Dio dokumenata klase c_i točne klasificiranih:

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

Preciznost

Dio dokumenata pridruženih klasi c_i koji su baš iz klase c_i:

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

Točnost

Dio dokumenata točne klasificiranih:

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

Mikro i makro prosjek

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

- **Makro-prosjek**: Izračunaj performanse za svaku klasu, i onda prosjek
- **Mikro-prosjek**: Prikupi odlike za svaku klasu, izračunaj tablicu slučaja, evaluiraj

Mikro i makro prosjek: primjer

Klasa 1

	Istina: da	Istina: ne
Odabran: da	10	10
Odabran: ne	10	970

Klasa 2

	Istina: da	Istina: ne
Odabran: da	90	10
Odabran: ne	10	890

Tablica mikro-prosjeka

	Istina: da	Istina: ne
Odabran: da	100	20
Odabran: ne	20	1860

- Makro-prosjek preciznost: (0.5+0.9)/2 = 0.7
- Mikro-prosjek preciznost: 100/120 = 0.83
- Mikro-prosjek dominira nad učestalim klasama

Razvojni testni skupovi i unakrsna validacija

Skup za trening | Razvojni testni skup | Testni skup

- Mjera: P/O/F1 ili točnost
- Neviđeni testni skup
  - izbjeći prekoračenja (ugađanje prema testnom skupu)
  - više konzervativna procjena performansi
- Unakrsna validacija (cross validation) nad višestrukim podjelama
  - rukovanje greškama uzorkovanja nad više skupova podataka
  - skupljanje skupova rezultata za svaku podjelu
  - izračunati performansu razvojnog skupa