

Uvod u obradu prirodnog jezika

8.1. Generativni protiv diskriminativnih modela

Branko Žitko

prevedeno od: Dan Jurafsky, Chris Manning

- Do sada smo razmatrali **generativne modele**
 - modeli jezika, Naivni Bayes
- Danas se često koriste **uvjetni odnosno diskriminativni probabilistički modeli** kod obrade prirodnog jezika, prepoznavanja govora, vraćanju informacija (strojnom učenju općenito) jer:
 - su veoma precizni
 - omogućavaju uključivanje mnogih lingvistički značajnih osobina
 - omogućavaju automatsku izgradnju jezično nezavisnih modula za obradu prirodnog jezika

Združeni protiv uvjetnog modela

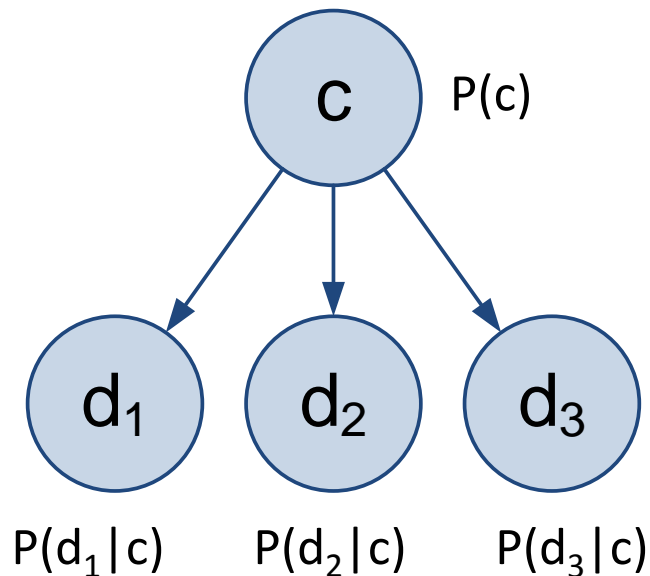
- Neka imamo skup $\{(d, c)\}$ uparenih promatranih podataka d i skrivenih klasa c
- **Združeni (generativni) model** postavljaju vjerojatnosti nad promatranim podacima d i skrivenim stvarima $P(c, d)$ (generiraju promatrane podatke temeljem skrivenih stvari)
 - Svi klasični statistički modeli obrade prirodnog jezika:
 - n-gram modeli, Naivni Bayesovi klasifikatori, skriveni Markovljevi modeli, probabilističke kontekstno neovisne gramatike, ...

Združeni protiv uvjetnog modela

- **Uvjetni (diskriminativni) model** ostavlja podatke kako jesu i postavljaju vjerojatnost nad skrivenim stvarima za dane podatke $P(c|d)$
 - logistička regresija, modeli maksimalne entropije, uvjetna slučajna polja
 - stroj s potpornim vektorima, perceptroni ... su diskriminativni modeli koji nisu probabilistički

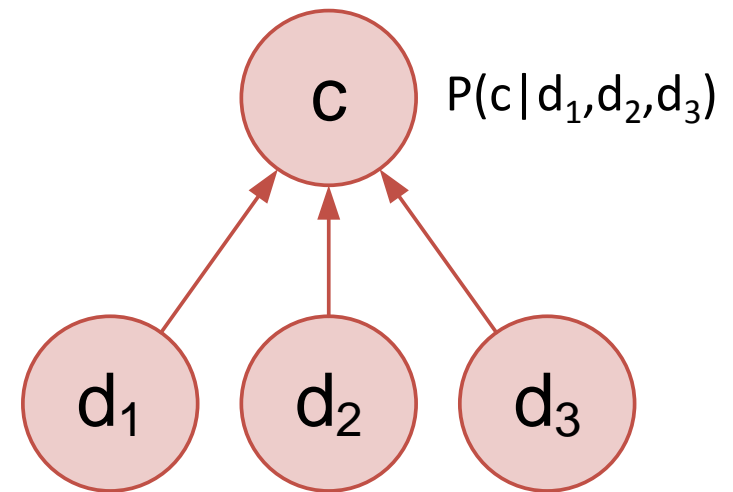
Bayesova mreža / grafički model

- Bayesova mreža za vrhove ima slučajne varijable i lukove za direktne zavisnosti
- Neke varijable se promatraju, dok su neke skrivene
- Svaki čvor je mali klasifikator (tablica uvjetne vjerojatnosti) temeljem dolaznih lukova



Naivni Bayes

Generativni model



Logistička regresija

Diskriminativni model

Uvjetna protiv združene vjerodostojnosti

- Združeni model daje vjerojatnosti $P(d,c)$ i pokušava maksimizirati združenu vjerodostojnost
 - trivijalnim izborom težina: samo relativne frekvencije.
- Uvjetni model daje vjerojatnost $P(c|d)$. Uzima podatke kakve jesu i modelira samo uvjetnu vjerojatnost klase
 - teži se maksimizaciji uvjetne vjerodostojnosti
 - teže za napraviti

Uvjetni modeli rade dobro

- Primjer određivanja smisla riječi
- Čak i s istim osobinama, promjenom metode na uvjetnu se povećavaju performanse

Skup za treniranje	
Metoda	Točnost
Združena	86.8
Uvjetna	98.5

Skup za testiranje	
Metoda	Točnost
Združena	73.6
Uvjetna	76.1

Uvod u obradu prirodnog jezika

8.2. Osobine diskriminativnog modela

Branko Žitko

prevedeno od: Dan Jurafsky, Chris Manning

Osobine (features)

- **Osobina** f je elementarni dio dokaza koji povezuje aspekte onoga što promatramo d (podatak) i onoga što predviđamo c (klasa)
- Osobina je realna funkcija

$$f : C \times D \rightarrow R$$

Primjer osobina

$f_1(c,d) \equiv [c=\text{LOKACIJA} \wedge w_{-1}=\text{"u"} \wedge \text{imaPrvoSlovoVeliko}(w)]$

$f_2(c,d) \equiv [c=\text{LOKACIJA} \wedge \text{imaHrvatskeZnakove}(w)]$

$f_3(c,d) \equiv [c=\text{LIJEK} \wedge \text{završava}(w,\text{"c"})]$

LOKACIJA

u Split

LOKACIJA

u Čakovec

LIJEK

uzima Prozac

OSOBA

vidi Ivana

- Model će svakoj osobini pridružiti **težinu**:
 - Pozitivna težina govori da je vjerojatno točno
 - Negativna težina govori da vjerojatno nije točno

Očekivanja osobina

- Iskorištavaju se dva **očekivanja**:
 - točan ili predviđeni broj korištenja osobina

- Empirijski broj (očekivanje) osobine

$$\text{empirijski } E(f_i) = \sum_{(c,d) \in \text{promatran}(C,D)} f_i(c,d)$$

- Model očekivanja osobine

$$E(f_i) = \sum_{(c,d) \in (C,D)} P(c,d) f_i(c,d)$$

Osobine

- Kod obrade prirodnog jezika, osobina najčešće specificira
 1. binarnu funkciju svojstva ulaznih podataka
 2. određenu klasu

$$f_i(c, d) \equiv [\Phi(d) \wedge c = c_j] \quad [\text{vrijednost je 0 ili 1}]$$

- Svaka osobina bira podskup podataka i sugerira kako će ih označiti

Modeli zasnovani na osobinama

- Odlučivanje o mjestu u podacima se temelji na aktivnim **osobinama** na tom mjestu

Podaci: POSAO: Dionice su dosegle vrh ...	Podaci: ... za rekonstrukciju bankovnog: NOVAC duga	Podaci: N V A ... ulica je duga ...
Oznaka: POSAO Osobine: {..., dionice, su, dosegle, vrh, ...}	Oznaka: NOVAC Osobine: {..., w_{-1} = rekonstrukciju, w_1 = duga, $L=9$, ...}	Oznaka: A Osobine: { w = duga, t_{-1} = V, w_{-1} = je, ...}
Kategorizacija teksta	Smisao riječi	POS označavanje

Primjer: Kategorizacija teksta

- (Zhang and Oles 2001)
- Osobine su prisutnosti svake riječi u dokumentu i klasi dokumenta
- Testovi na klasičnim Reuters skupovima podataka
 - Naivni Bayes: 77.0% F1
 - Linearna regresija: 86.0%
 - Logistička regresija: 86.4%
 - SVM: 86.5%
- Rad ističe važnost regularizacije (izgladivanja) za uspješnu upotrebu diskriminativnih metoda.

Drugi primjeri klasifikacije s maksimalnom entropijom

- MAXENT klasifikator koristiti kadgod se želi pridružiti mjesto podataka jednoj od brojnih klasa:
 - Detekcija granice rečenice (Mikheev 2000)
 - Je li točka kraj rečenice ili kratica?
 - Sentimentalna analiza (Pang and Lee 2002)
 - unigrami, bigrami riječi, brojenje POS, ...
 - prijedlozi u POS (Ratnaparkhi 1998)
 - Pridružiti glagolu ili imenici? Osobine glavne imenice, prijedloga itd.
 - Opće odlučivanje prilikom parsiranja (Ratnaparkhi 1997; Johnson et al. 1999, etc.)

Uvod u obradu prirodnog jezika

8.3. Linearni klasifikator temeljen na osobinama (Feature-based linear classifier)

Branko Žitko

prevedeno od: Dan Jurafsky, Chris Manning

Linearni klasifikatori temeljeni na osobinama

- Linearni klasifikatori tijekom klasifikacije:
 - Linearna funkcija iz skupa osobina $\{f_i\}$ u skup klase $\{c\}$
 - Pridruživanje težine λ_i svakoj osobini f_i
 - Promatra se svaka klasa c za dane podatke d
 - Za par (c, d) , osobine biraju njihovim težinama
 - $\text{biraj}(c) = \sum \lambda_i f_i(c, d)$

	$\lambda_1=1.8$	$\lambda_2=-0.6$	$\lambda_3=0.3$
OSOBA	LOKACIJA	LIJEK	
<i>u Čakovec</i>	<i>u Čakovec</i>	<i>u Čakovec</i>	

- Izabrana klasa c koja maksimizira vjerojatnost
 $\sum \lambda_i f_i(c, d) = \text{LOKACIJA}$

Linearni klasifikatori temeljeni na osobinama

- Postoje razni načini odabira težina za osobine:
 - Perceptron: pronađi trenutni pogrešno klasificiran primjer i guraj težine u smjeru njegove korektne klasifikacije
 - Granične metode (SVM)

Linearni klasifikatori temeljeni na osobinama

- Eksponencijalni (log-linearni, maxent, logički, Gibbs) modeli:
 - Napravi probabilistički model iz linearne kombinacije $\sum \lambda_i f_i(c, d)$

$$P(c \mid d, \lambda) = \frac{\exp \sum_i \lambda_i f_i(c, d)}{\sum_{c'} \exp \sum_i \lambda_i f_i(c', d)}$$

Čini biranje pozitivnim

Normalizira biranje

- $P(\text{LOKACIJA} \mid \text{u Čakovec}) = e^{1.8}e^{-0.6} / (e^{1.8}e^{-0.6} + e^{0.3} + e^0) = 0.586$
- $P(\text{LIJEK} \mid \text{u Čakovec}) = e^{0.3} / (e^{1.8}e^{-0.6} + e^{0.3} + e^0) = 0.238$
- $P(\text{OSOBA} \mid \text{u Čakovec}) = e^0 / (e^{1.8}e^{-0.6} + e^{0.3} + e^0) = 0.176$

- Težine su parametri probabilističkog modela

Linearni klasifikatori temeljeni na osobinama

- Eksponencijalni (log-linearni, maxent, logistički, Gibbs) modeli:
 - Za ovaj oblik modela, izabrat ćemo parametre $\{\lambda_i\}$ koji **maksimiziraju uvjetnu vjerojatnost** podataka po ovom modelu.
 - Ne gradi se samo klasifikacija, već i distribucija vjerojatnosti nad klasifikacijom
 - postoje drugi (dobri) načini diskriminiranja klasa – SVM, perceptroni – ali nisu toliko trivijalne za interpretaciju distribucija nad klasama.

Veza sa logističkom regresijom

- Maxent modeli u OPJ su esencijalno isti kao i modeli višeklasne logističke regresije u statistici (ili strojnom učenju)

Uvod u obradu prirodnog jezika

8.4. Izgradnja MaxEnt modela

Branko Žitko

prevedeno od: Dan Jurafsky, Chris Manning

Izgradnja MaxEnt modela

- Definiranje osobina nad mjestima u podacima
 - osobine predstavljaju skupove mjesta u podacima koji su dovoljno karakteristični da zasluže parametre modela:
 - riječi, riječi s brojem, riječi koje završavaju na "iti", itd.
- Kodiranje Φ osobine kao jedinstveni string
 - Podatak će dovesti do niza stringova; aktivnih Φ osobina
 - Svaka osobina $f_i(c, d) \equiv [\Phi(d) \wedge c = c_j]$ dobiva težinu kao realni broj

Izgradnja MaxEnt modela

- Osobine se često dodavaju tijekom razvoja modela kako bi označili pogreške
 - Često, najjednostavnije je dodavati osobine koje označavaju loše kombinacije
- Tada, za svaku težinu osobine, želimo biti u stanju izračunati:
 - uvjetnu vjerojatnost podataka
 - derivaciju vjerojatnosti sa težinom modela
 - koristi očekivanja za svaku osobinu po modelu
- Sada se mogu odrediti optimalne težine osobina

Uvod u obradu prirodnog jezika

8.5. Problem nadbrojavanja dokaza

Branko Žitko

prevedeno od: Dan Jurafsky, Chris Manning

Klasifikacija teksta: Azija ili Evropa?

Evropa

Monaco
Monaco

Monaco

Monaco
Monaco

Monaco
Hong
Kong

Podaci za treniranje

Hong
Kong
Monaco

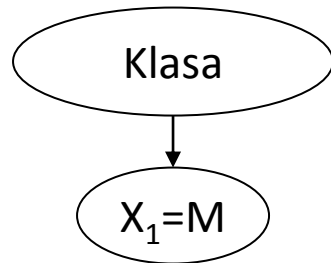
Monaco

Hong
Kong

Azija

Hong
Kong

Naivni Bayes (NB)
model



NB faktori

- $P(A) = P(E) = 4/8 = 1/2$
- $P(M|A) = 2/8 = 1/4$
- $P(M|E) = 6/8 = 3/4$

Predikcije

- $P(A, M) = 1/2 * 1/4 = 1/8$
- $P(E, M) = 1/2 * 3/4 = 3/8$
- $P(A|M) = (1/8) / (8/16) = 1/4$
- $P(E|M) = (3/8) / (8/16) = 3/4$

Klasifikacija teksta: Azija ili Evropa?

Evropa

Monaco
Monaco

Monaco

Monaco
Monaco

Monaco
Hong
Kong

Podaci za treniranje

Hong
Kong
Monaco

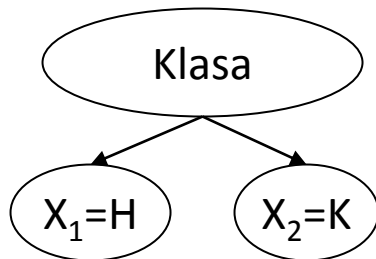
Monaco

Hong
Kong

Azija

Hong
Kong

Naivni Bayes (NB)
model



NB faktori

- $P(A) = 4/8 = 1/2$
- $P(E) = 4/8 = 1/2$
- $P(H|A) = P(K|A) = 3/8$
- $P(H|E) = P(K|E) = 1/8$

Predikcije

- $P(A, H, K) = 1/2 * 3/8 * 3/8 \propto 9$
- $P(E, H, K) = 1/2 * 1/8 = 1/8 \propto 1$
- $P(A|H, K) = 9/(9+1) = 9/10$
- $P(E|H, K) = 1/(9+1) = 1/10$

Klasifikacija teksta: Azija ili Evropa?

Evropa

Monaco
Monaco

Monaco

Monaco
Monaco

Monaco
Hong
Kong

Podaci za treniranje

Hong
Kong
Monaco

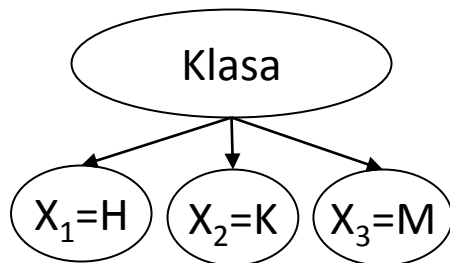
Monaco

Hong
Kong

Azija

Hong
Kong

Naivni Bayes (NB)
model



NB faktori

- $P(A) = 4/8 = 1/2$
- $P(E) = 4/8 = 1/2$
- $P(M|A) = 2/8 = 1/4$
- $P(M|E) = 6/8 = 3/4$
- $P(H|A) = P(K|A) = 3/8$
- $P(H|E) = P(K|E) = 1/8$

Predikcije

- $P(A, H, K, M) = 1/2 * 3/8 * 3/8 * 1/4 \propto 9$
- $P(E, H, K, M) = 1/2 * 1/8 * 1/8 * 3/4 \propto 3$
- $P(A|H, K, M) = 9/(9+3) = 9/12 = 3/4$
- $P(E|H, K, M) = 3/(9+3) = 3/12 = 1/4$

Naivni Bayes protiv MaxEnt modela

- Naivni Bayesovi modeli višestruko broje dokaze koji su povezani
 - Svaka osobina je pomnožena, iako postoji više osobina koje govore o istoj stvari
- MaxEnt modeli (u većini) rješavaju ovaj problem
 - To se postiže određivanjem težina osobinama rezultirajući da očekivanja modela odgovaraju promatranim očekivanjima

Uvod u obradu prirodnog jezika

8.6. Maksimizacija vjerodostojnosti

Branko Žitko

prevedeno od: Dan Jurafsky, Chris Manning

Vjerodostojnost eksponencijalnog modela

- Maksimalni (uvjetni) modeli vjerodostojnosti
 - za dani model, odaberi parametre tako da se maksimizira (uvjetne) vjerodostojnosti podataka

$$\log P(C | D, \lambda) = \sum_{(c,d) \in (C,D)} \log P(c | d, \lambda) = \sum_{(c,d) \in (C,D)} \log \frac{\exp \sum_i \lambda_i f_i(c, d)}{\sum_{c'} \exp \sum_i \lambda_i f_i(c', d)}$$

Vrijednost vjerodostojnosti

- (Logaritamska) uvjetna vjerodostojnost promatranih podataka (C, D) po maxent modelu je funkcija podataka i parametara λ :

$$\log P(C | D, \lambda) = \log \prod_{(c,d) \in (C,D)} P(c | d, \lambda) = \sum_{(c,d) \in (C,D)} \log P(c | d, \lambda)$$

- Ako nema mnogo vrijednosti od c , jednostavno se izračuna:

$$\log P(C | D, \lambda) = \sum_{(c,d) \in (C,D)} \log \frac{\exp \sum_i \lambda_i f_i(c, d)}{\sum_{c'} \exp \sum_i \lambda_i f_i(c', d)}$$

Vrijednost vjerodostojnosti

- Možemo izdvojiti dvije komponente:

$$\log P(C | D, \lambda) = \sum_{(c,d) \in (C,D)} \log \exp \sum_i \lambda_i f_i(c, d) - \sum_{(c,d) \in (C,D)} \log \sum_{c'} \exp \sum_i \lambda_i f_i(c', d)$$

$$\log P(C | D, \lambda) = N(\lambda) - M(\lambda)$$

- Derivacija je razlika derivacija svake komponente

Derivacija brojnika

$$\begin{aligned}\frac{\partial N(\lambda)}{\partial \lambda_i} &= \frac{\partial \sum_{(c,d) \in (C,D)} \log \exp \sum_i \lambda_i f_i(c,d)}{\partial \lambda_i} = \frac{\partial \sum_{(c,d) \in (C,D)} \sum_i \lambda_i f_i(c,d)}{\partial \lambda_i} \\ &= \sum_{(c,d) \in (C,D)} \frac{\partial \sum_i \lambda_i f_i(c,d)}{\partial \lambda_i} = \sum_{(c,d) \in (C,D)} f_i(c,d)\end{aligned}$$

- Derivacija brojnika je stvarni broj(f_i, c)

Derivacija nazivnika

$$\begin{aligned}\frac{\partial M(\lambda)}{\partial \lambda_i} &= \frac{\partial \sum_{(c,d) \in (C,D)} \log \sum_{c'} \exp \sum_i \lambda_i f_i(c', d)}{\partial \lambda_i} \\&= \sum_{(c,d) \in (C,D)} \frac{1}{\sum_{c''} \exp \sum_i \lambda_i f_i(c'', d)} \frac{\partial \sum_{c'} \exp \sum_i \lambda_i f_i(c', d)}{\partial \lambda_i} \\&= \sum_{(c,d) \in (C,D)} \frac{1}{\sum_{c''} \exp \sum_i \lambda_i f_i(c'', d)} \sum_{c'} \frac{\exp \sum_i \lambda_i f_i(c', d)}{1} \frac{\partial \sum_i \lambda_i f_i(c', d)}{\partial \lambda_i} \\&= \sum_{(c,d) \in (C,D)} \sum_{c'} \frac{\exp \sum_i \lambda_i f_i(c', d)}{\sum_{c''} \exp \sum_i \lambda_i f_i(c'', d)} \frac{\partial \sum_i \lambda_i f_i(c', d)}{\partial \lambda_i} \\&= \sum_{(c,d) \in (C,D)} \sum_{c'} P(c' | d, \lambda) f_i(c', d)\end{aligned}$$

- Derivacija nazivnika je predviđen broj(f_i, c)

Derivacija

$$\frac{\partial \log P(C | D, \lambda)}{\partial \lambda_i} = \text{stvaran broj}(f_i, c) - \text{predviđen broj}(f_i, c)$$

- Optimalni parametri su oni kod kojih je predviđeno očekivanje jednako empirijskom očekivanju.
- Optimalna distribucija:
 - uvijek je jedinstvena (ali parametri ne moraju biti jedinstveni)
 - uvijek postoji (ako su odabrane osobine nad aktualnim podacima)
- Ovi modeli se zovu modeli maksimalne entropije

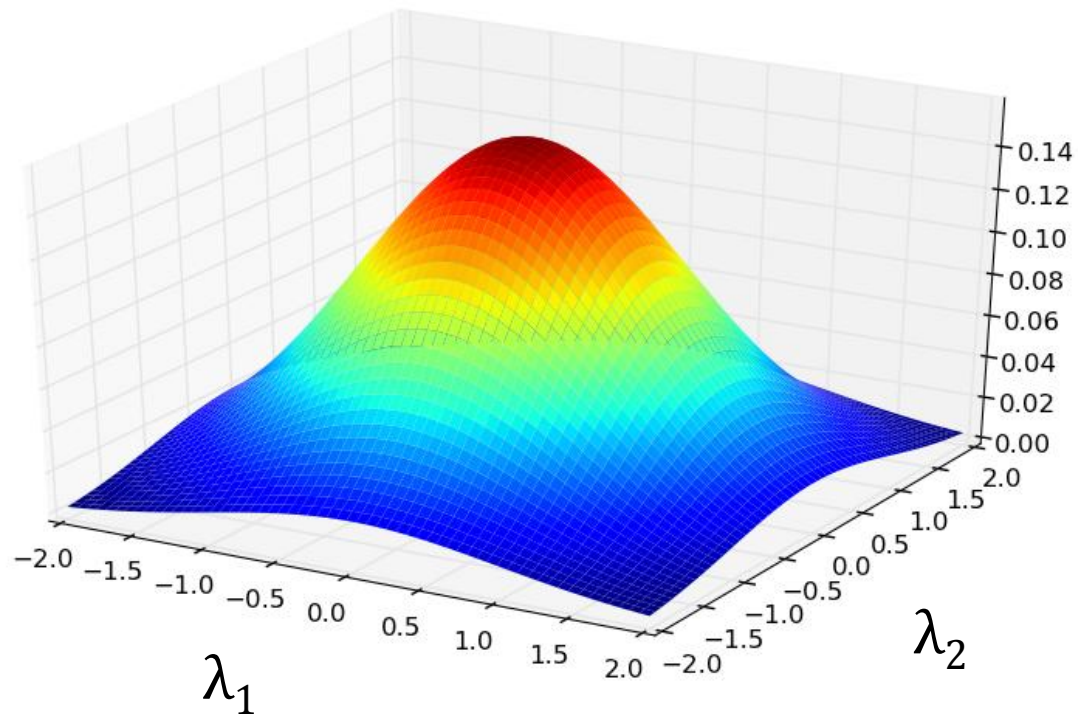
Pronalaženje optimalnih parametara

- Želimo odabrati parametre $\lambda_1, \lambda_2, \lambda_3 \dots$ koji maksimiziraju uvjetni logaritam vjerodostojnosti (conditional log likelihood) podataka za treniranje

$$CLogLik(D) = \sum_{i=1}^n \log P(c_i | d_i)$$

- Kako bi to napravili, pokazali smo kako se računaju parcijalne derivacije (gradijent)

Površina vjerodostojnosti



Pronalaženje optimalnih parametara

1. GD - Gradijentno spužtanje (Gradient Descent)
SGD - Stohastičko gradijentno spužtanje
2. GIS – Generalized Iterative Scaling
IIS – Improved iterative Scaling
3. CG – Conjugate Gradient
4. Kvazi-Newtonove metode