Statističko učenje

Poglavlje 20.1–2

Prema slajdovima Stuarta Russella i Tomislava Šmuca (hvala)!

Sadržaj

- ♢ Bayesovo učenje — optimalno predviđanje
- ♢ Maksimalno *a posteriori* učenje (MAP)
- ♢ Učenje maksimalne izglednosti (ML)
  (engl. likelihood = izglednost, vjerodostojnost)
- ♢ Naivna Bayesova klasifikacija
- ♢ Mrežno Bayesovo učenje
  – maks. izgledno parametarsko učenje s potpunim podacima
  – linearna regresija

Učenje i vjerojatnosno učenje

Strojno učenje: naći najbolju ili najuspješniju tzv. hipotezu ili model
= ona koja dobro opisuje podatke koji su nam dostupni za učenje.

Cilj: koristiti tu hipotezu/model za predviđanje na ostalim/budućim podacima, izvan skupa za učenje.

Zato: dobro = dovoljno jednostavno, dobro za generalizaciju, ali ne "predobro" ili "presloženo" (pretreniranje), jer je loše za generalizaciju.

Vjerojatnosni pristup: hipoteza H = slučajna varijabla i
najbolja hipoteza = najvjerojatnija hipoteza (model).

Standardni model vjerojatnosnog ili Bayesovog učenja:
Hipoteze iz H imaju neku početnu distribuciju vjerojatnosti $\mathbf{P}(H)$.
Učenje na podacima = ažuriranje te distribucije uz dane "dokaze" (podatke), prema Bayesovom pravilu (uvjetna vjerojatnost).

Bayesovo učenje — diskretni model

Diskretni prostor hipoteza:

H je diskretna slučajna varijabla, tzv. "hipotezna" varijabla, s vrijednostima h_i, za $i = 1, \ldots, m$ (konačna), ili $i \in \mathbf{N}$ (niz).

Početna ili a priori distribucija vjerojatnosti $\mathbf{P}(H)$ je poznata.

Pretpostavka: varijabla H nije izravno opaziva = nije u podacima (tj. kao da je skrivena varijabla — usporediti poslije s klasifikacijom).

Podaci za učenje ili treniranje su: $\mathbf{d} = d_1, \ldots, d_N$ ("d" = data), svaki podatak ili opažanje d_j = uzorak (ishod) slučajne varijable D_j.

Interpretacija je kao ranije kod "dokaza" u Bayesovim mrežama:

$\mathbf{D}$ je skup svih pripadnih varijabli D_j, a $\mathbf{D} = \mathbf{d}$ je opažena vrijednost.

U danom trenutku, imamo N takvih uzoraka, s tim da N može rasti, kako dobivamo nova opažanja.

Bayesovo učenje — osnovni pojmovi

Prema Bayesovom pravilu, za dosad dane podatke $\mathbf{d}$, svaka hipoteza h_i ima a posteriori vjerojatnost:

$$P(h_i|\mathbf{d}) = \frac{P(\mathbf{d}|h_i)\,P(h_i)}{P(\mathbf{d})} = \alpha\,P(\mathbf{d}|h_i)\,P(h_i).$$

Nazivi pojedinih vrijednosti:

$P(h_i)$ = a priori (početna) vjerojatnost hipoteze h_i,

$P(h_i|\mathbf{d})$ = a posteriori (naknadna) vjerojatnost od h_i, za dane $\mathbf{d}$

= izglednost hipoteze za dane podatke,

$P(\mathbf{d}|h_i)$ = izglednost podataka $\mathbf{d}$, uz danu hipotezu h_i.

Općenito: a posteriori vjerojatnost hipoteze h_i ovisi o

– njezinoj prethodnoj vjerojatnosti,

– samim podacima za učenje $\mathbf{d}$,

– vjerojatnosti dobivanja baš tih podataka, uz danu hipotezu h_i.

Potpuno Bayesovo učenje — predviđanje

Predviđanje nepoznate vrijednosti X (slučajna varijabla), za dane $\mathbf{d}$,
uz pretpostavku da svaka hipoteza određuje (zadaje) neku distribuciju vjerojatnosti za X,
tj. traži se distribucija $\mathbf{P}(X|\mathbf{d})$, ili pojedinačna vjerojatnost $P(x_k|\mathbf{d})$.

Hipotezu H tretiramo kao "skrivenu" varijablu — marginalizacija ili zbrajanje po svim mogućim vrijednostima hipoteze H

$$\mathbf{P}(X|\mathbf{d}) = \sum_i \mathbf{P}(X \mid \mathbf{d}, h_i)\, P(h_i|\mathbf{d}).$$

Ključna stvar u modelu — vrlo razumna u praksi:
Za danu hipotezu H, nepoznata vrijednost X i podaci $\mathbf{D}$ su uvjetno nezavisni, tj. vrijedi $\mathbf{P}(X \mid \mathbf{d}, h_i) = \mathbf{P}(X|h_i)$, pa je

$$\mathbf{P}(X|\mathbf{d}) = \sum_i \mathbf{P}(X|h_i)\, P(h_i|\mathbf{d}).$$

Ovo je tzv. optimalno ili potpuno Bayesovo predviđanje učenjem.

Potpuno Bayesovo učenje — komentar

Prethodna formula za optimalno Bayesovo predviđanje

$$\mathbf{P}(X|\mathbf{d}) = \sum_i \mathbf{P}(X|h_i)\, P(h_i|\mathbf{d})$$

koristi sve hipoteze h_i, tj. nije bazirana na samo jednoj "najbolje" pogođenoj ili predviđenoj hipotezi!

Optimalno predviđanje je "težinska" srednja vrijednost
- predviđanja po svim pojedinačnim hipotezama, a
- "težine" su a posteriori vjerojatnosti pojedinih hipoteza.

Lijepo, ali skupo — u praksi, obično, koristimo neku aproksimaciju, tako da biramo jednu = "najbolju" hipotezu.

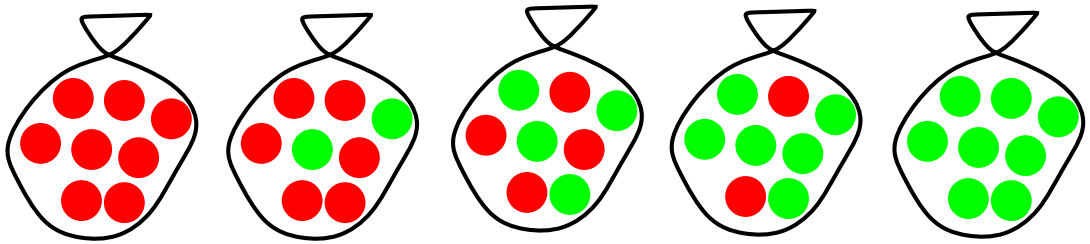
Opći pogled na gornju formulu: hipoteze h_i su "međuvrijednosti" između sirovih podataka i predviđanja.

Možemo i bez njih — uz dodatne vrijednosti u podacima (klasifikacija).

Primjer

Pretpostavimo da imamo pet različitih vrsta vrećica s bombonima:

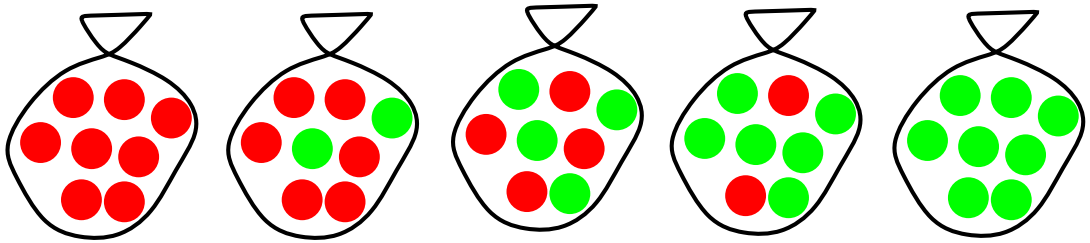
10% su h_1: 100% bomboni s višnjom
20% su h_2: 75% bomboni s višnjom + 25% bomboni s limunom
40% su h_3: 50% bomboni s višnjom + 50% bomboni s limunom
20% su h_4: 25% bomboni s višnjom + 75% bomboni s limunom
10% su h_5: 100% bomboni s limunom



Gledano izvana, vrećice se ne razlikuju — ne znamo koja je koja!

Stvarno, pretpostavljamo da su "vrećice" jako velike, tako da izvlačenje nekog bombona iz vreće ne mijenja vjerojatnost njezinog sadržaja, tj.
– $p(h_i)$ se time neće promijeniti.

Primjer — podaci za učenje, uvjetna nezavisnost



Zatim, vadimo 10 bombona iz neke (iste) vreće i dobivamo uzorak

$$d_1, \ldots, d_{10} = \bullet\bullet\bullet\bullet\bullet\bullet\bullet\bullet\bullet\bullet .$$

Ova opažanja gledamo kao "rastući" uzorak podataka za učenje, tj. u trenutku N je $\mathbf{d} = d_1, \ldots, d_N$, za $N = (0), 1, \ldots, 10$.

A priori distribucija hipoteza (prema proizvođaču bombona) je

$$\mathbf{P}(H) = \langle 0.1, 0.2, 0.4, 0, 2, 0.1 \rangle$$

Izglednost podataka računamo kao da su opažanja nezavisna, s istom distribucijom vjerojatnosti (velika vreća)

$$P(\mathbf{d}|h_i) = \prod_j P(d_j|h_i).$$

Primjer — pitanja, uvjetna nezavisnost

Pitanja:

Koja je to vrsta vrećice?

Kojeg će okusa biti sljedeći bombon?

Prvo pitanje: hipoteza h_i koja ima najveću a posteriori vjerojatnost.

$$P(h_i|\mathbf{d}) = \alpha\, P(\mathbf{d}|h_i)\, P(h_i).$$

Drugo pitanje: predviđanje za varijablu $X =$ "okus bombona"

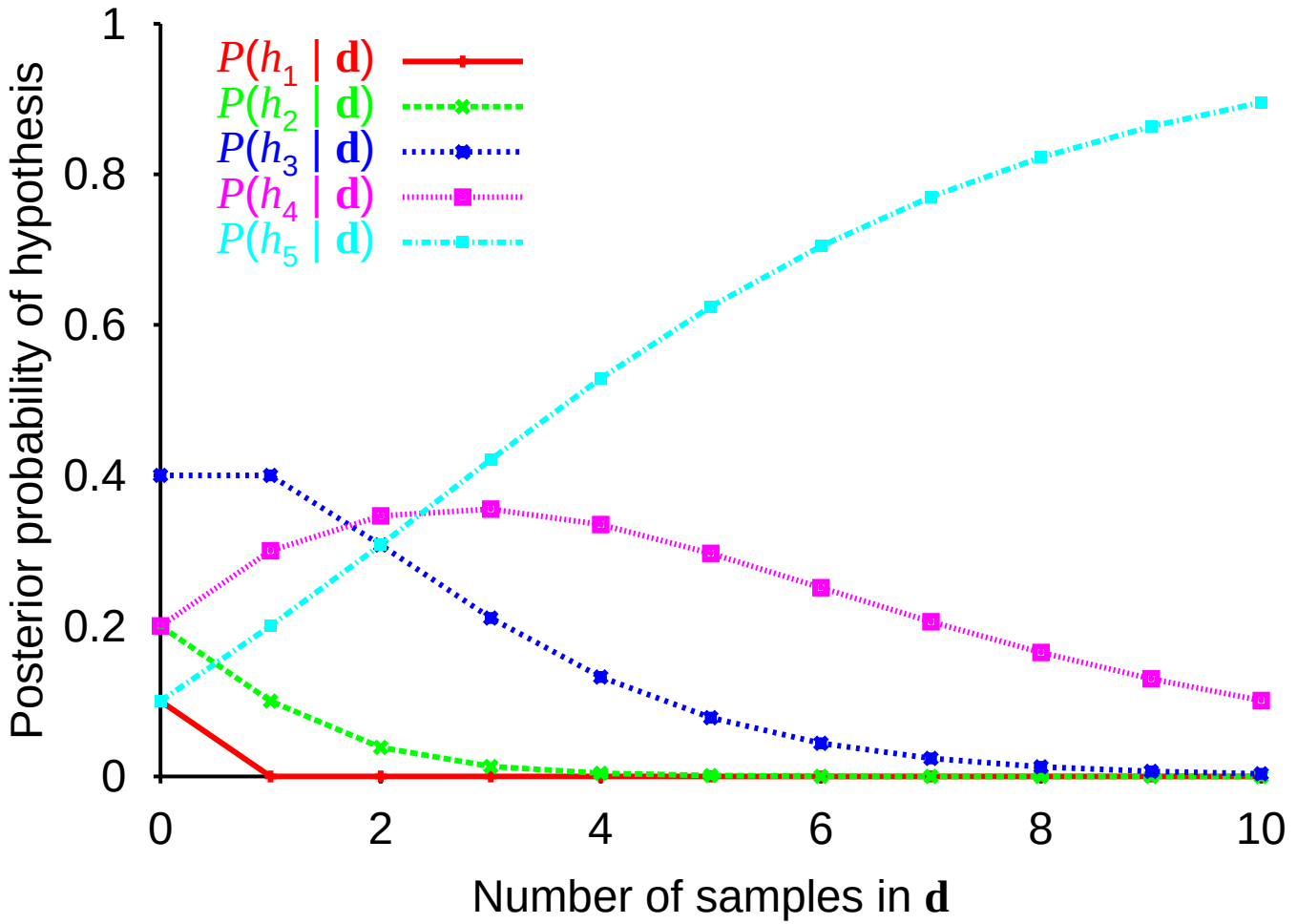
$$\mathbf{P}(X|\mathbf{d}) = \sum_i \mathbf{P}(X\,|\,\mathbf{d}, h_i)\, P(h_i|\mathbf{d}) = \sum_i \mathbf{P}(X|h_i)\, P(h_i|\mathbf{d}),$$

s tim da se traži vrijednost x_k koja ima najveću vjerojatnost u $\mathbf{P}(X|\mathbf{d})$.

Uvjetna nezavisnost: ako znamo $h_i =$ koja je vreća, onda je vjerojatnost okusa bombona potpuno određena tipom vreće i nema veze s podacima, tj. vrijedi

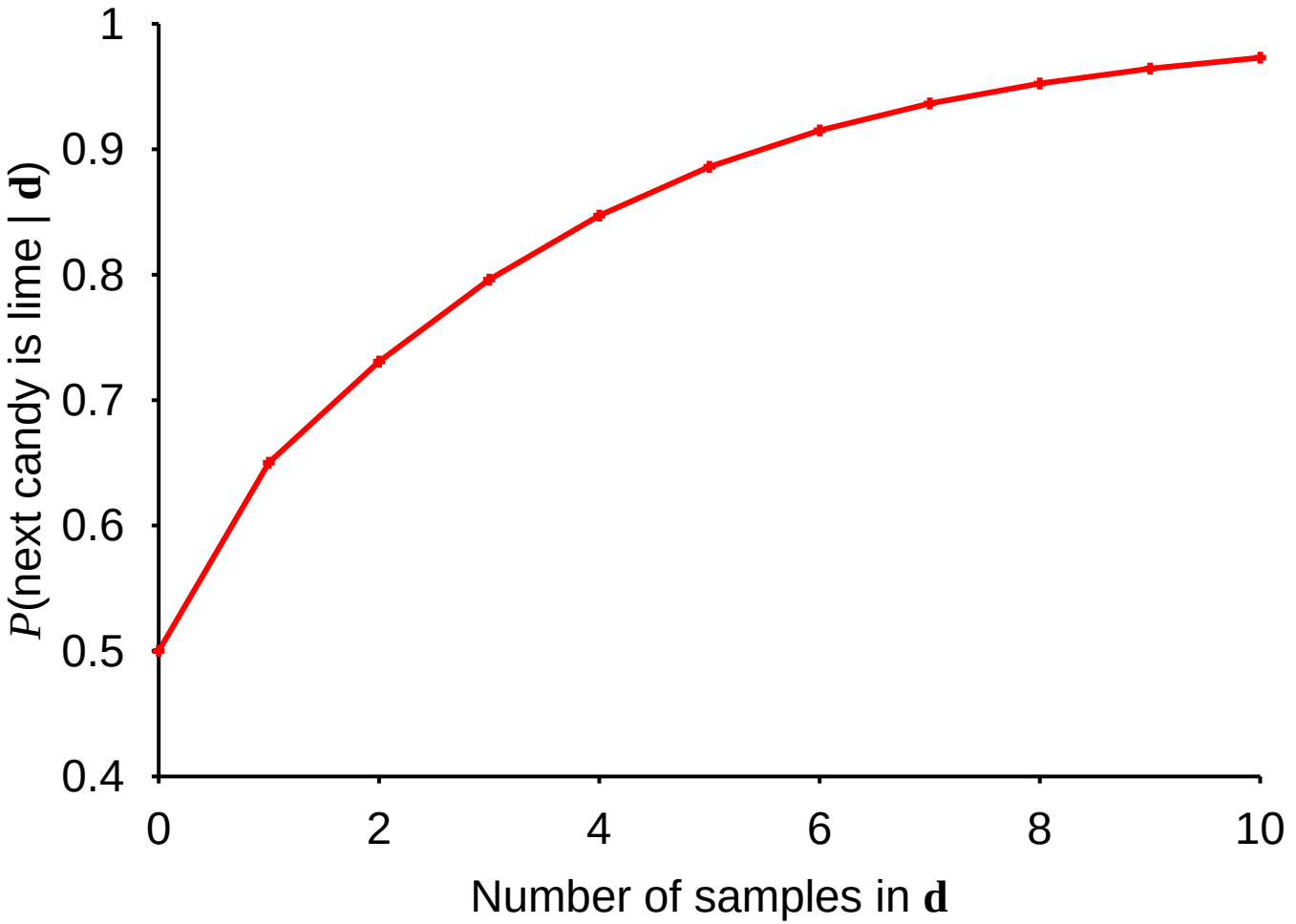
$$\mathbf{P}(X\,|\,\mathbf{d}, h_i) = \mathbf{P}(X|h_i).$$

Aposteriorna vjerojatnost hipoteza



Za $N = 0$ (bez učenja) dobivamo a priori vjerojatnosti hipoteza.
Nakon $N = 3$, hipoteza h_5 postaje najizglednija!

Predviđanje vjerojatnosti



Prema optimalnom predviđanju, vjerojatnost limuna brzo ide prema 1.
To se slaže s "pravom" hipotezom h_5.

Maksimalna a posteriori aproksimacija — MAP

Optimalno predviđanje je skupo — sve hipoteze su "u igri"!
Zbrajanje preko cijelog prostora hipoteza je često neostvarivo,
na pr., 18 446 744 073 709 551 616 Booleovih funkcija sa 6 atributa.

Maksimalno a posteriori (MAP) učenje:
izabrati hipotezu h_{MAP} koja maksimizira $P(h_i|\mathbf{d})$,
za predviđanje je onda $\mathbf{P}(X|\mathbf{d}) \approx \mathbf{P}(X|h_{MAP})$.

Tj., treba maksimizirati $P(\mathbf{d}|h_i)\,P(h_i)$, ili $\log P(\mathbf{d}|h_i) + \log P(h_i)$.

Log članovi mogu se shvatiti kao (negativni) broj bitova za
kodiranje podataka uz danu hipotezu + kodiranje hipoteze.
To je osnovna ideja učenja minimalno dugog opisa (MDL)
(engl. "minimum description length").

Za determinističke hipoteze, $P(\mathbf{d}|h_i)$ je 1 ako je konzistentno, 0 inače
$\Rightarrow$ MAP = najjednostavnija konzistentna hipoteza (vidi znanost).

Maksimalno izgledna aproksimacija — ML

Za velike skupove podatake, a priori vjerojatnost postaje beznačajna
(pogledati prethodni primjer).

Učenje maksimalne izglednosti (ML):
izabrati h_{ML} koja maksimizira $P(\mathbf{d}|h_i)$.

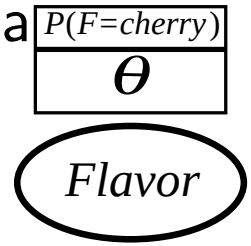
Tj., jednostavno se uzme najbolja aproksimacija podataka;
identično MAP-u za uniformne apriori vjerojatnosti
(što je razumno ako su sve hipoteze iste složenosti).

ML je "standardna" (ne-Bayesova) metoda statističkog učenja.

Problemi s ML za male skupove podataka = loša aproksimacija.

ML parametarsko učenje u Bayesovim mrežama

Vrećica s bombonima novog proizvođača; udio θ bombona s višnjom?
Moguć je bilo koji $\theta \in [0,1]$ — imamo kontinuum hipoteza h_θ
θ je parametar za ovu jednostavnu (binomnu) familiju modela



Pretpostavimo da smo razmotali N bombona, od toga
c s višnjama i $\ell = N - c$ s limunom,
To su nezavisna, jednako distribuirana opažanja, pa je izglednost za $\mathbf{d}$

$$P(\mathbf{d}|h_\theta) = \prod_{j=1}^{N} P(d_j|h_\theta) = \theta^c \cdot (1-\theta)^\ell$$

Maksimizacija toga po θ — što je jednostavnije za log-izglednost:

$$L(\mathbf{d}|h_\theta) = \log P(\mathbf{d}|h_\theta) = \sum_{j=1}^{N} \log P(d_j|h_\theta) = c\log\theta + \ell\log(1-\theta)$$

$$\frac{dL(\mathbf{d}|h_\theta)}{d\theta} = \frac{c}{\theta} - \frac{\ell}{1-\theta} = 0 \quad \Rightarrow \quad \theta = \frac{c}{c+\ell} = \frac{c}{N}$$

Izgleda razumno, ali postoji problem kad se dogodi 0 u brojanju!

ML parametarsko učenje — nastavak

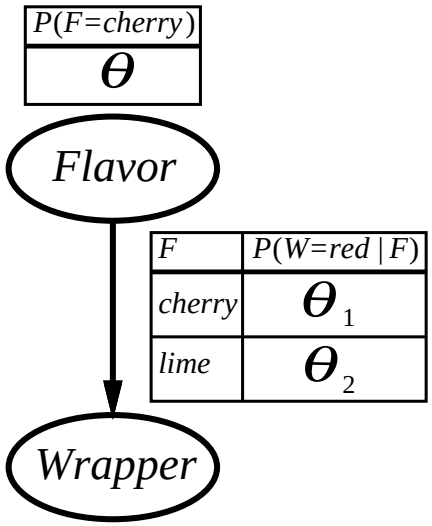
Riječima: h_{ML} hipoteza kaže da je stvarni udio višnji u vreći jednak opaženom udjelu do tada, tj. puno posla za očitu stvar!

Primjer je prejednostavan, pa izgleda očito . . . , svejedno, to je standardna metoda za ML parametarsko učenje, s puno primjena:

1. Izabrati parametriziranu familiju modela koji opisuju sve podatke
   zahtijeva znatan uvid i ponekad nove modele
2. Napisati izglednost podataka kao funkciju parametara
   može zahtijevati zbrajanje preko nevidljivih varijabli, tj. zaključivanje
3. Napisati derivaciju log izglednosti obzirom na svaki parametar
4. Naći vrijednosti parametara tako da su derivacije jednake nuli
   može biti teško / nemoguće; pomažu moderne tehnike optimizacije.

Više parametara

Crveni / zeleni omoti bombona vjerojatnosno ovise o okusu — uvjetna distribucija za dani okus, s nepoznatim parametrima:



Izglednost za, na primjer, zeleno–omotani bombon s višnjom:

$$\begin{aligned} &P(F = cherry,\, W = green \,|\, h_{\theta,\theta_1,\theta_2}) \\ &= P(F = cherry | h_{\theta,\theta_1,\theta_2})\, P(W = green | F = cherry, h_{\theta,\theta_1,\theta_2}) \\ &= \theta \cdot (1 - \theta_1) \end{aligned}$$

Uzorak: N bombona, višnje: r_c s crvenim, g_c sa zelenim omotima, itd.

$$P(\mathbf{d}|h_{\theta,\theta_1,\theta_2}) = \theta^c(1-\theta)^\ell \cdot \theta_1^{r_c}(1-\theta_1)^{g_c} \cdot \theta_2^{r_\ell}(1-\theta_2)^{g_\ell}$$

$$\begin{aligned} L = &[c \log\theta + \ell \log(1-\theta)] \\ &+ [r_c \log\theta_1 + g_c \log(1-\theta_1)] + [r_\ell \log\theta_2 + g_\ell \log(1-\theta_2)] \end{aligned}$$

Više parametara — nastavak

Derivacije od L sadrže samo bitan parametar:

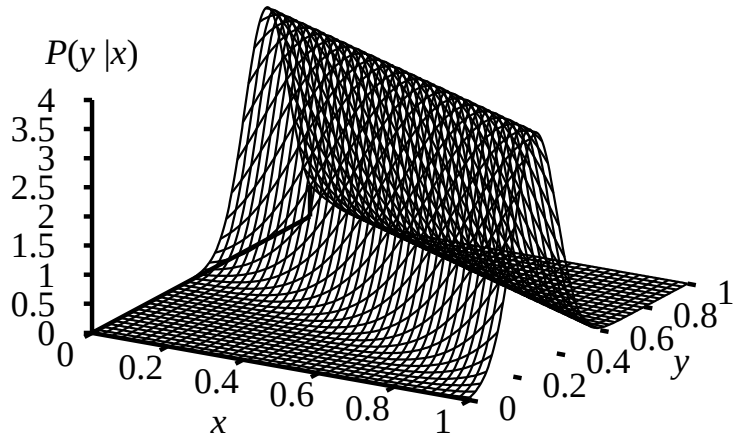
$$\frac{\partial L}{\partial \theta} = \frac{c}{\theta} - \frac{\ell}{1-\theta} = 0 \qquad \Rightarrow \quad \theta = \frac{c}{c+\ell}$$

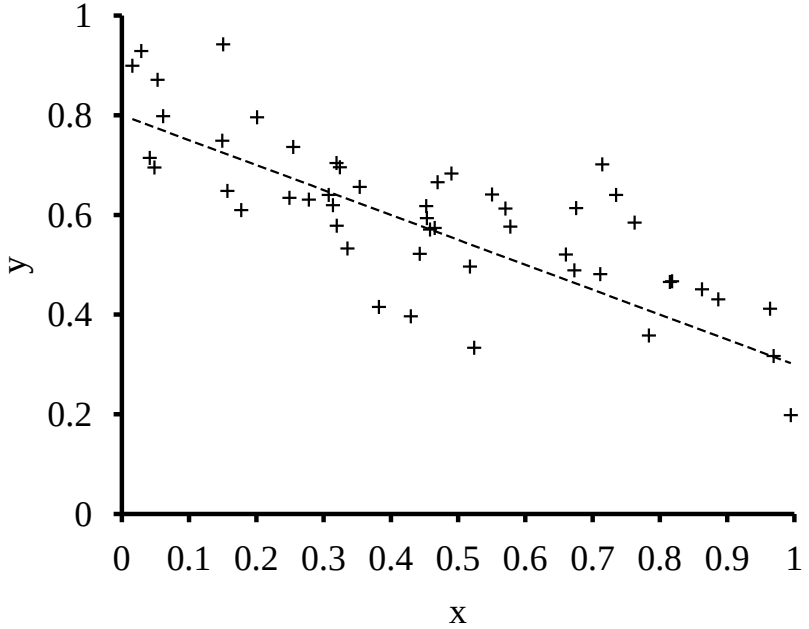
$$\frac{\partial L}{\partial \theta_1} = \frac{r_c}{\theta_1} - \frac{g_c}{1-\theta_1} = 0 \qquad \Rightarrow \quad \theta_1 = \frac{r_c}{r_c+g_c}$$

$$\frac{\partial L}{\partial \theta_2} = \frac{r_\ell}{\theta_2} - \frac{g_\ell}{1-\theta_2} = 0 \qquad \Rightarrow \quad \theta_2 = \frac{r_\ell}{r_\ell+g_\ell}$$

U slučaju potpunog skupa podataka, **parametri se mogu učiti zasebno**.

ML za neprekidni model — linearni Gaussov model





Maksimizacija $P(y|x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y-(\theta_1 x+\theta_2))^2}{2\sigma^2}}$ obzirom na θ_1, θ_2

$=$ minimizacija $E = \sum_{j=1}^{N} (y_j - (\theta_1 x_j + \theta_2))^2$

Minimizacija sume kvadrata grešaka daje ML rješenje za linearnu aproksimaciju uz pretpostavku da **Gaussov šum ima fiksnu varijancu**

Sažetak

Potpuno Bayesovo učenje daje najbolja moguća predviđanja ali nije ostvarivo

MAP učenje uravnotežuje složenost s točnošću na podacima za treniranje

Maksimalna izglednost pretpostavlja uniformnu apriori vjerojatnost, OK za velike skupove podataka