A mesterséges intelligencia alapjai

statisztikai tanulási módszerek

tartalom

- statisztikai tanulás
- maximum-likelihood paramétertanulás: diszkrét modellek
- naiv Bayes-modellek
- Bayes-hálóstruktúrák tanulása
  - maximum-likelihood
  - lineáris regresszió

statisztikai tanulás

Bayes - tanulás

tények - adott területet leíró valószínűségi változók konkrét megvalósulása

- $\mathbf{d}=d_1,...,d_N$ (tanuló adatok: a *j*. kísérlet eredménye d_j, a D_j v.v. értéke)

hipotézis - elmélet: hogyan működik a világ?

- H - hipotézis változó, h_1, h_2, … értékek $\mathcal{P}(H)$ a priori eloszlással
- $P(h_i|\mathbf{d})=\alpha P(\mathbf{d}|h_i)P(h_i)$ - ahol $P(\mathbf{d}|h_i)$ likelihood

A következő X mennyiségre vonatkozó predikció

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

cukorkás példa

Tegyük fel, hogy a gyártó 5 fajta csomagolásban küldi a terméket:

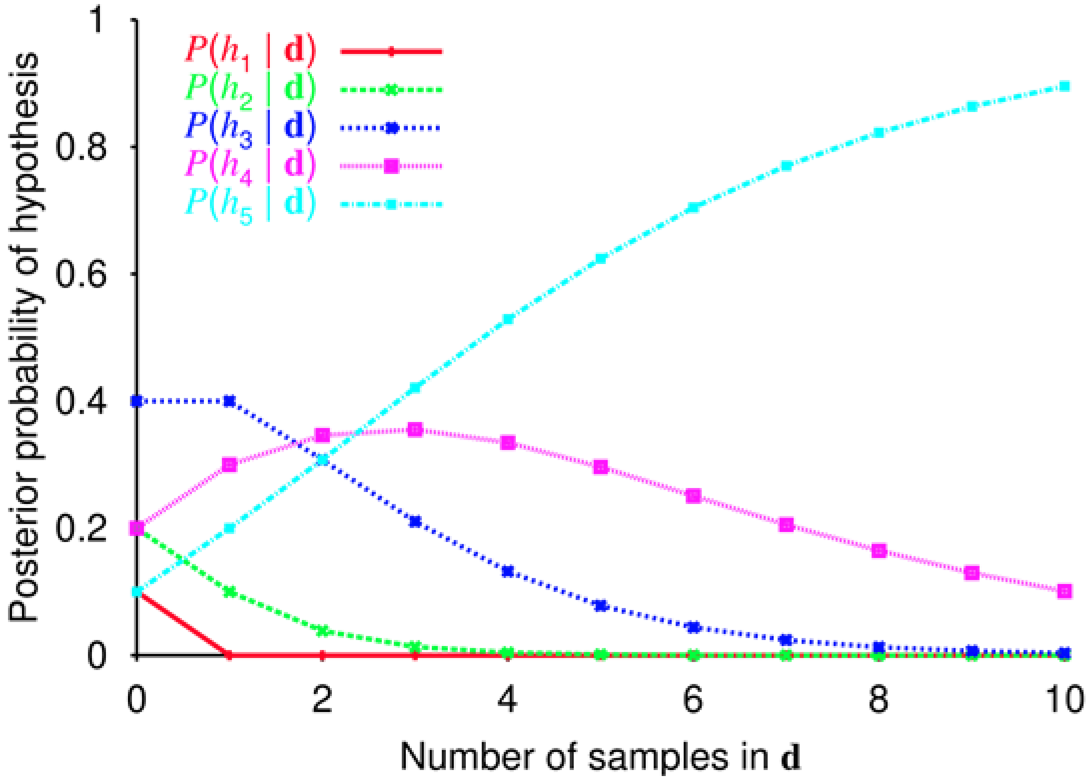
- h1 (10%): 100% meggyes cukorka
- h2 (20%): 75% meggyes cukorka + 25% citromos
- h3 (40%): 50% meggyes cukorka + 50% citromos
- h4 (20%): 25% meggyes cukorka + 75% citromos
- h5 (10%): 100% citromos cukorka

A csomagot megbontva 10 db citromos cukorkát veszünk ki találomra.

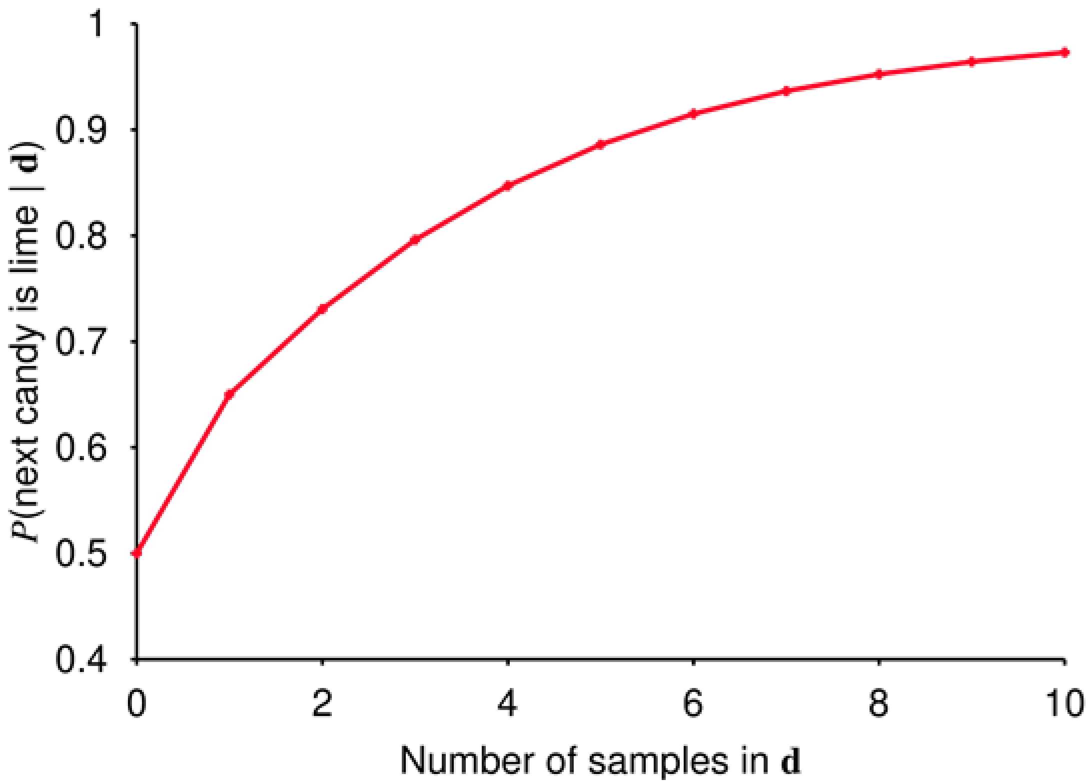
- Melyik fajta csomagolásból kaptunk?
- Milyen lesz a soron következő cukorka?

(egyforma és egyenletes eloszlást feltételezve)

$P(h_i|d_1,\ldots,d_N)$ a posteriori valószínűségek



$P(d_{N+1}=\text{citrom}|d_1,...,d_N)$ Bayes-predikció



maximum a posteriori hipotézis (MAP)

a hipotézisek tere gyakran kezelhetetlenül nagy (pl. korábban 6 változós logikai. fv.)

egyetlen, a legvalószínűbb hipotézis alapján végezzük a predikciót

- azon h_i alapján, mely maximalizálja $P(h_i|\mathbf{d})$-t
- maximalizálja $P(\mathbf{d}|h_i)P(h_i)$-t, és vele együtt $\log P(\mathbf{d}|h_i) + \log P(h_i)$ -t is
- utóbbi tekinthető azon bitek számának, ami kódolja a mintát az adott hipotézis esetén + kódolja a hipotézist. (1-nél kisebb számok miatt negatív mennyiség)
  - minimális hosszúságú leírás (MDL) - adatkódolás minimalizálása

3 citromos cukor után a MAP 100%-ra jósolja a negyedik cukor citromosságát, a Bayes csak 80%-ra.

Sok adat esetén a MAP és Bayes konvergál

maximum likelihood hipotézis

- feltétel: egyenletes $\mathcal{P}$(H) a priori eloszlás
  - MAP speciális esete
- statisztikában nagyon elterjedt, standard módszer
- hasznos, ha a hipotézisek komplexek, kezdeti eloszlást nehéz meghatározni
- nagy adathalmaz esetén megközelíti a Bayes- és MAP-tanulást

teljes adattal történő tanulás

paraméter-tanulás teljes adat alapján

- **teljes adat**: mindegyik adatpont értéket tartalmaz a megtanulandó valószínűségi modell valamelyik paraméterére
- paramétertanulás: egy rögzített struktúrájú modell paramétereinek megtalása

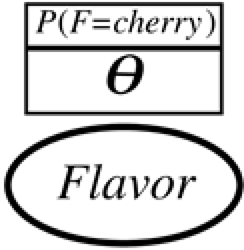
Probléma

- új gyártótól érkező cukorkák, ismeretlen a meggy aránya.
- paraméter: θ - a meggycukorkák aránya ($0 \leq \theta \leq 1$)
- hipotézis: h_θ - végtelen sok lehetőség

Modell

- kibontott cukor íze → csomagban lévő cukrok aránya

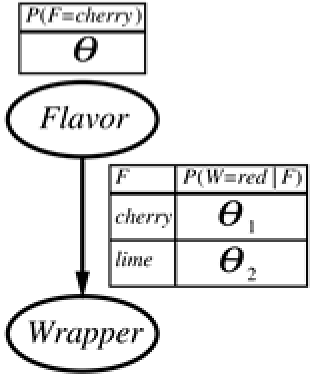
paraméter-tanulás Bayes-hálózatban



- Tegyük fel, hogy kiválasztunk *N* cukorkát, melyből *m* db. meggy, *c* db. citrom.
- ennek valószínűsége: $P(\mathbf{d}|h_\theta) = \prod_j P(d_j|h_\theta) = \theta^m(1-\theta)^c$,
  - mert független, egyenletes eloszlás
- Maximalizáljuk ezt az értéket! Ua. mint ha a log likelihood függvényt max.
- $L(\mathbf{d}|h_\theta) = \mathit{log}\, P(\mathbf{d}|h_\theta) = \sum_j \mathit{log} P(d_j|h_\theta) = m\, \mathit{log}\, \theta + c\, \mathit{log}\, (1-\theta)$
- deriváljuk a kifejezést θ szerint, majd nézzük meg, hol egyenlő 0-val!
- $dL(\mathbf{d}|h_\theta)/d\theta = m/\theta - c/(1-\theta) = 0$, így $\theta=m/(m+c)=m/N$
  - a hipotézis azt állítja, hogy a csomagban érvényes arány megegyezik a kibontott cukroknál megfigyelt aránnyal.
- ha az adathalmaz kicsi, az ML 0 valószínűséget rendel a még meg nem történt eseményekhez

több paraméter használata

A gyártó eltérő cukorkacsomagolást használ, de nem következetes: zöld csomagolásban is lehet meggyes cukorka.



paramétereink:

θ - meggyes cukorkák aránya,

θ_1- meggyes cukorkát pirosba csomagoltak

θ_2- citromos cukorkát pirosba csomagoltak

- $P(Íz=m,Cs=z|h_{\theta,\theta 1,\theta 2})=P(Íz=m|h_{\theta,\theta 1,\theta 2})P(Cs=z|Íz=m,h_{\theta,\theta 1,\theta 2})=\theta(1-\theta_1)$

több paraméter használata - 2

Kibontunk N cukorkát, ebből m meggyes, c citromos. p_m és p_c a pirosba csomagoltak száma, míg z_m és z_c a zöldbe.

- $P(\mathbf{d}|h_{\theta,\theta1,\theta2}) = \theta^m(1-\theta)^c\ \theta_1^{pm}(1-\theta_1)^{zm}\ \theta_2^{pc}(1-\theta_2)^{zc}$
- $L = \big(m\ \mathit{log}\ \theta + c\ \mathit{log}\ (1-\theta)\big)+\big(p_m\ \mathit{log}\ \theta_1 + z_m\ \mathit{log}\ (1-\theta_1)\big)+\big(p_c\ \mathit{log}\ \theta_2 + z_c \mathit{log}\ (1-\theta_2)\big)$
- külön vesszük a deriváltakat θ,θ_1 és θ_2 szerint,
- $dL/d\theta = m/\theta - c/(1-\theta)$ $\theta=m/(m+c)$
- $dL/d\theta_1 = p_m/\theta_1 - z_m/(1-\theta_1)$ $\theta_1=p_m/(p_m+z_m)$ - a legjobb közelítés
- $dL/d\theta_2 = p_c/\theta_2 - z_c/(1-\theta_2)$ $\theta_2=p_c/(p_c+z_c)$ megfigyelt arány

teljes adatok esetén a Bayes-háló paramétertanulási problémája elkülönülő tanulási problémákra dekomponálható, egy-egy probléma egy-egy paraméterre

naiv Bayes-modellek

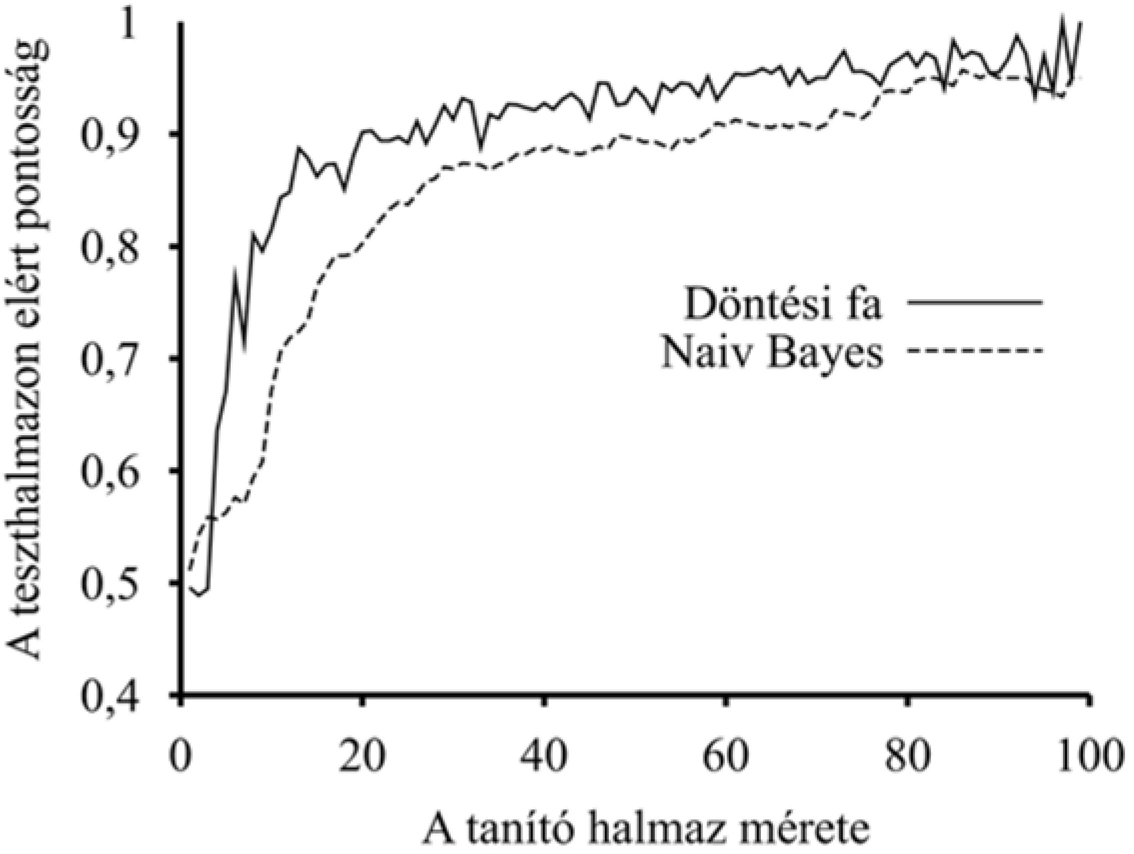
naiv Bayes-modell

A megjósolandó C osztály-változó a fa gyökere, az X_i attribútumváltozók a fa levelei.

$P(C=igaz)=\theta$,
$P(X_1=igaz|C=igaz)=\theta_1$, …,
$P(X_n=igaz|C=igaz)=\theta_n$

$\mathcal{P}(C|x_1,...,x_n) = \alpha\mathcal{P}(C)\prod_i P(x_i|C)$

naiv, mert feltételezi, hogy az attribútumok egymástól feltételesen függetlenek



ML paramétertanulás - folytonos eset

folytonos eset

$$P(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

$$L = \sum_{j=1}^{N} \log \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_j-\mu)^2}{2\sigma^2}} = N(-\log\sqrt{2\pi} - \log\sigma) - \sum_{j=1}^{N} \frac{(x_j-\mu)^2}{2\sigma^2}$$

$$\frac{\partial L}{\partial \mu} = -\frac{1}{\sigma^2} \sum_{j=1}^{N} (x_j - \mu) = 0 \implies \mu = \frac{\sum_j x_j}{N}$$

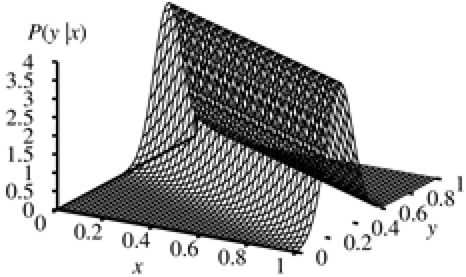
$$\frac{\partial L}{\partial \sigma} = -\frac{N}{\sigma} + \frac{1}{\sigma^3} \sum_{j=1}^{N} (x_j - \mu)^2 = 0 \implies \sigma = \sqrt{\frac{\sum_j (x_j-\mu)^2}{N}}$$

az átlag ML-becslése a mintaátlag, a szórás ML-becslése a minta átlagos szórásnégyzetének négyzetgyöke (*józan ész*)

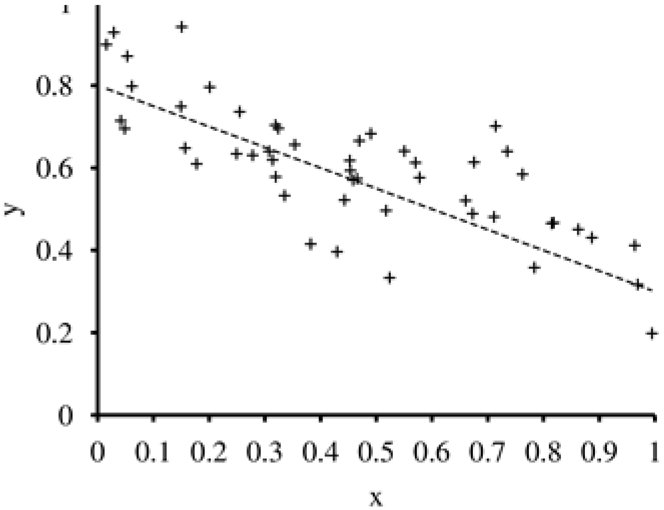
lineáris Gauss-modell, X f. szülő, Y f. gyerek

$$y = \theta_1 x + \theta_2 + \varepsilon$$

$$P(y|x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y-(\theta_1 x+\theta_2))^2}{2\sigma^2}}$$ maximalizálása



$$E = \sum_{j=1}^{N} (y - (\theta_1 x + \theta_2))^2$$ minimalizálása



E a jól ismert hibanégyzetek összege, ezt a lineáris regresszió minimalizálja; feltéve, hogy ε rögzített varienciájú zaj.

összefoglalás

- a teljes Bayes-tanulás adja a legjobb eredményeket, de gyakran kezelhetetlen bonyolultságú
- a MAP-tanulás egészséges kompromisszum a bonyolultság tekintetében
- az ML-tanulás egyenlő valószínűségű hipotéziseket feltételez, nagy méretű adatokra helyes
- tekintsünk egy paraméterezett családját a modellek egy halmazának
- írjuk le az adatok valószínűségét mint a paraméterek egy függvényét
    - ez a rejtett változók szerinti összegzést igényelhet
- keressük meg a paraméterek azon értékét, ahol a derivált 0
    - időnként nehéz feladat, optimalizációs módszerek segíthetnek