Apprendimento Statistico

Capitolo 20, Sezioni 1–3

Outline

- ♢ Apprendimento Bayesiano
- ♢ Apprendimento
  - "maximum *a posteriori*" (MAP)
  - "maximum likelihood" (ML)
- ♢ Apprendimento in Reti Bayesiane
  - apprendimento di parametri con ML nel caso di dati completi
  - regressione lineare

Apprendimento Bayesiano

L'apprendimento è "visto" come un processo di aggiornamento Bayesiano di una distribuzione di probabilità definita su uno Spazio delle Ipotesi

H variabile "ipotesi": assume valori $h_1, h_2, \ldots$, e ha probabilità a priori $\mathbf{P}(H)$

j-esima osservazione d_j: realizzazione della variabile random D_j
Insieme di Apprendimento $\mathbf{d} = d_1, \ldots, d_N$

Dati gli esempi osservati fino ad un certo punto, una generica ipotesi ha probabilità a posteriori:

$P(h_i|\mathbf{d}) = \alpha P(\mathbf{d}|h_i)P(h_i)$ dove $P(\mathbf{d}|h_i)$ è chiamata likelihood

Le predizioni usano una media pesata della likelihood sulle ipotesi:

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

Non c'è bisogno di selezionare l'ipotesi migliore!

Esempio

Supponiamo che esistano 5 tipi diversi di confezioni di caramelle:

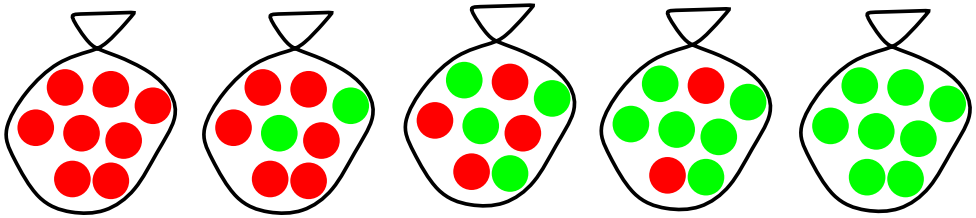
10% contengono h_1: 100% caramelle alla ciliegia

20% contengono h_2: 75% caramelle alla ciliegia + 25% caramelle al limone

40% contengono h_3: 50% caramelle alla ciliegia + 50% caramelle al limone

20% contengono h_4: 25% caramelle alla ciliegia + 75% caramelle al limone

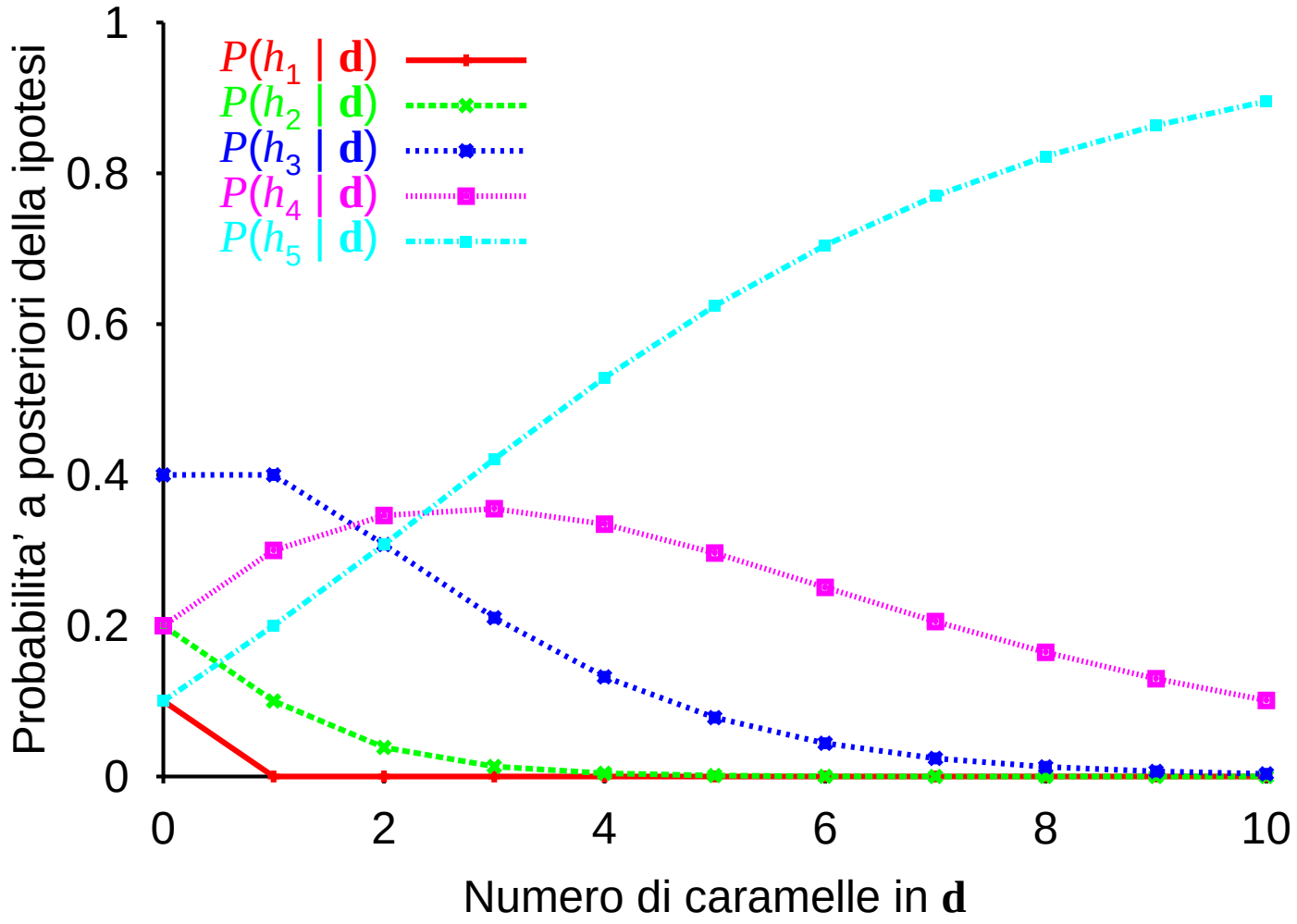
10% contengono h_5: 100% caramelle al limone



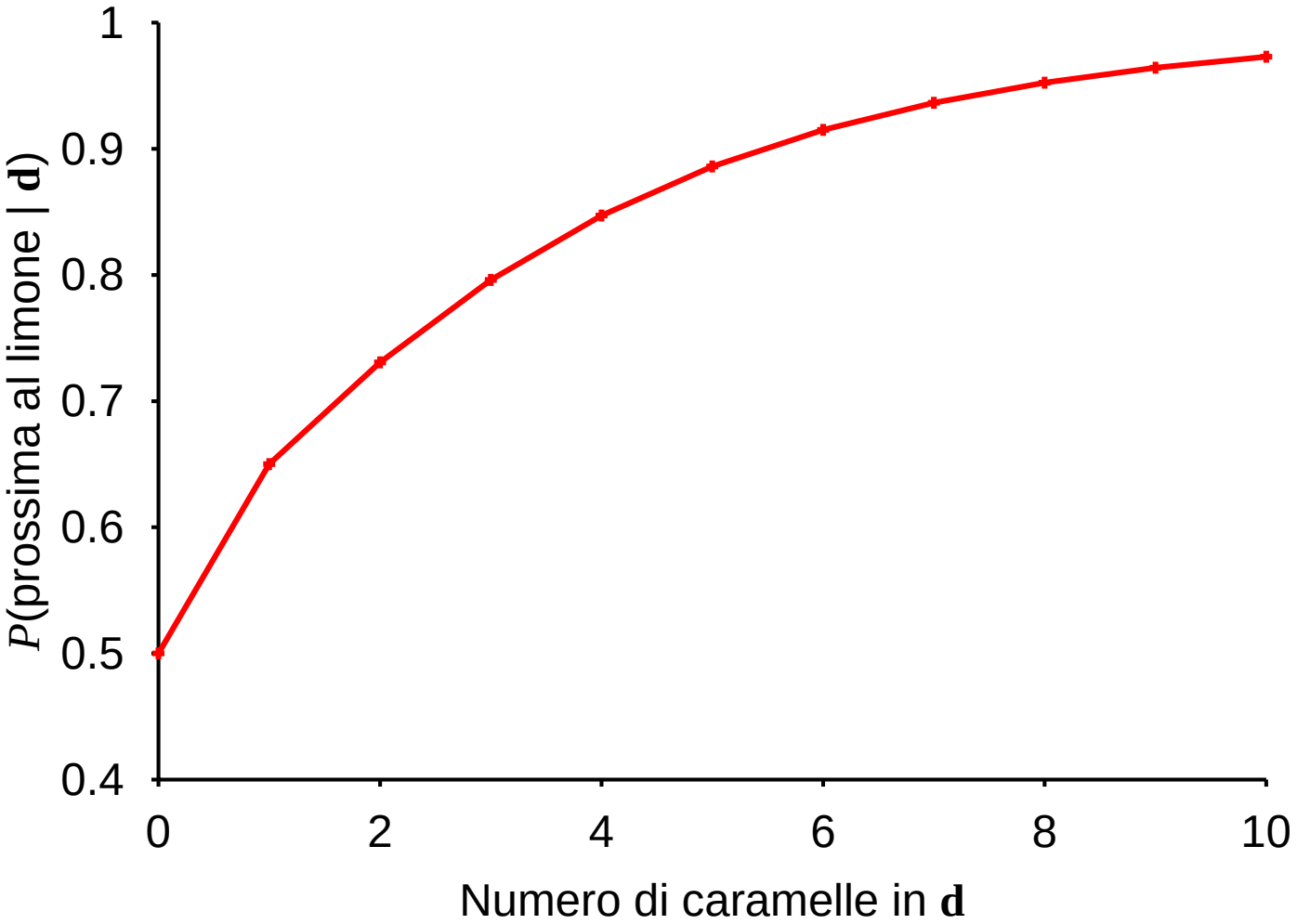
Osserviamo le caramelle estratte da una confezione: 

Di che tipo è la confezione? Quale sarà il gusto della prossima caramella?

Probabilità a posteriori delle ipotesi



Probabilità di predizione



Approssimazione MAP

Sommare su tutte le ipotesi è spesso impraticabile
(p.e., esistono 18,446,744,073,709,551,616 funzioni booleane distinte per 6 attributi)

Maximum a posteriori (MAP): sceglie l'ipotesi h_{MAP} che massimizza $P(h_i|\mathbf{d})$

Cioè, quella che massimizza $P(\mathbf{d}|h_i)P(h_i)$ o $\log P(\mathbf{d}|h_i) + \log P(h_i)$

I termini logaritmici possono essere visti come (il negativo del)
bit per codificare i dati data l'ipotesi + # bit per codificare l'ipotesi
Questa è l'idea base del principio di minimum description length (MDL)

Per ipotesi deterministiche, $P(\mathbf{d}|h_i)$ è 1 se h_i consistente con $\mathbf{d}$, 0 altrimenti
$\Rightarrow$ MAP = ipotesi consistente più semplice

Approssimazione ML

Per insiemi di apprendimento grandi, la probabilità a priori diventa irrilevante

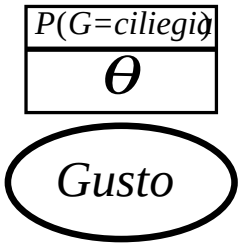
Maximum likelihood (ML): sceglie l'ipotesi h_{ML} che massimizza $P(\mathbf{d}|h_i)$

Cioè, quella che "descrive" meglio i dati (minimizza errori);
identica a MAP con probabilità a priori uniforme sulle ipotesi
(ragionevole se tutte le ipotesi hanno la stessa complessità)

ML è il metodo "standard" (non-Bayesiano) per l'apprendimento statistico

Apprendimento in Reti Bayesiane con ML

Confezione di marca diversa; frazione θ di caramella alla ciliegia?
Qualunque θ è possibile: continuum di ipotesi h_θ
θ è un parametro per la seguente semplice famiglia (binomiale) di modelli:



Supponiamo di scartare N caramelle, c alla ciliegia e $\ell = N - c$ al limone
Queste sono osservazioni i.i.d. (indipendenti, identicamente distribuite), quindi

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

Cerchiamo il massimo rispetto a θ— più facile per log-likelihood:

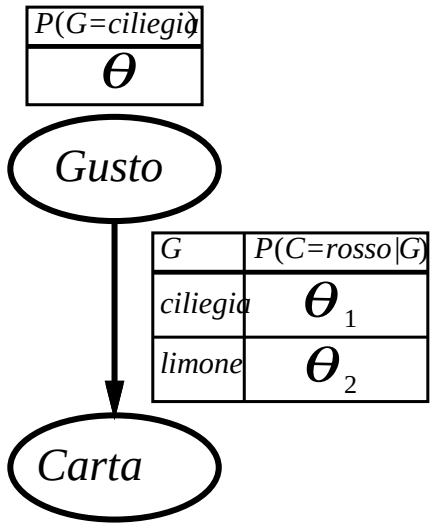
$$L(\mathbf{d}|h_\theta) = \log P(\mathbf{d}|h_\theta) = \sum_{j=1}^{N} \log P(d_j|h_\theta) = c\log\theta + \ell\log(1-\theta)$$

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

Sembra funzionare, ma ci sono problemi per contatori a 0 !

Parametri multipli

Il colore rosso/verde della carta dipende probabilisticamente dal gusto:



Likelihood per caramelle alla ciliegia (c) con carta verde (v):

$$\begin{aligned} & P(G=c, C=v|h_{\theta,\theta_1,\theta_2}) \\ &= P(G=c|h_{\theta,\theta_1,\theta_2})P(C=v|G=c, h_{\theta,\theta_1,\theta_2}) \\ &= \theta \cdot (1-\theta_1) \end{aligned}$$

N caramelle, r_c ciliegia con carta rossa, etc.:

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

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

Parametri multipli

Le derivate di L contengono solo i parametri rilevanti:

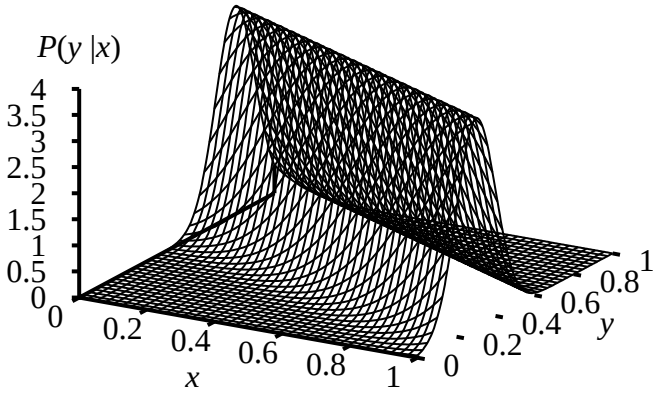
$$\frac{\partial L}{\partial \theta} = \frac{c}{\theta} - \frac{\ell}{1-\theta} = 0 \qquad \Rightarrow \quad \theta = \frac{c}{c+\ell}$$

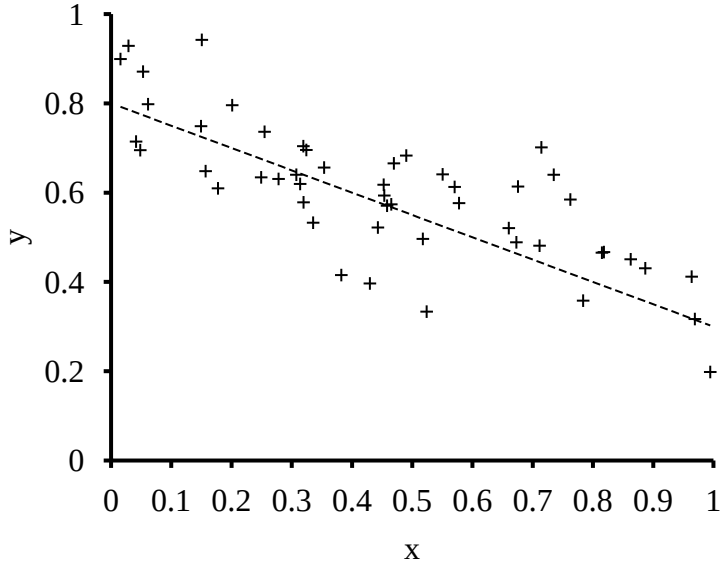
$$\frac{\partial L}{\partial \theta_1} = \frac{r_c}{\theta_1} - \frac{v_c}{1-\theta_1} = 0 \qquad \Rightarrow \quad \theta_1 = \frac{r_c}{r_c+v_c}$$

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

Con dati completi, **i parametri si possono apprendere separatamente**

Esempio: modello Gaussiano lineare





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

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

Cioè, minimizzare la somma del quadrato degli errori restituisce la soluzione ML per una interpolazione lineare **assumendo rumore Gaussiano di varianza prefissata**

Riassunto

Apprendimento Bayesiano "puro" produce la migliore predizione possibile ma è intrattabile

Apprendimento MAP bilancia la complessità con l'accuratezza sui dati d'apprendimento

Maximum likelihood assume probabilità a priori uniforme, OK per grosse moli di dati

1. Scegliere una famiglia parametrizzata di modelli per descrivere i dati
richiede esperienza e a volte lo sviluppo di nuovi modelli

2. Scrivere la likelihood dei dati come funzione dei parametri
può richiedere la somma sulle variabili nascoste, cioè inferenza

3. Scrivere la derivata della log likelihood rispetto ad ogni parametro

4. Trovare i valori dei parametri per cui le derivate sono nulle
può essere difficile/impossibile; in molti casi occorrono tecniche avanzate di ottimizzazione