UNIVERSITÀ DEGLI STUDI DI SALERNO
DIPARTIMENTO DI INFORMATICA

Intelligenza Artificiale

Apprendimento statistico

a.a. 2021/2022

Outline

- Apprendimento bayesiano
- *Maximum a posteriori* (MAP) ed apprendimento con il metodo della massima verosimiglianza (ML)
- Reti di apprendimento bayesiane
  - Machine Learning con dati completi

Terminologia

- Dati: instanziazioni di alcune o tutte le variabili casuali che descrivono il dominio. Rappresentano delle *prove.*
- Ipotesi: teorie probabilistiche di come funziona un dominio

Esempio

- Supponiamo ci siano cinque tipi di sacchetti di caramelle:

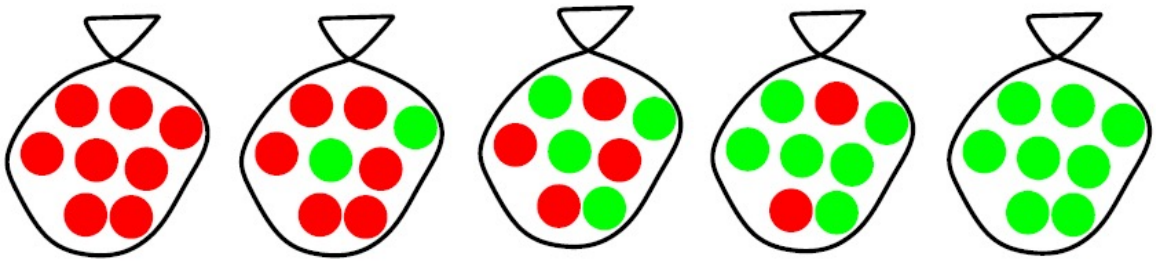
  h_1: 100% caramelle alla ciliegia

  h_2: 75% caramelle alla ciliegia + 25% caramelle al lime

  h_3: 50% caramelle alla ciliegia + 50% caramelle al lime

  h_4: 25% caramelle alla ciliegia + 75% caramelle al lime

  h_5: 100% caramelle al lime



Formulazione del problema

- Dato un nuovo sacchetto
  - una variabile d'ipotesi H con valori $h_1, h_2, \dots, h_5$ denota il tipo di sacchetto;
  - D_i è una variabile causale (ciliegia o lime);
  - dopo aver visto $D_1, D_2, \dots, D_N$, vogliamo predire il sapore (ossia il valore) di D_{N+1}

Apprendimento bayesiano completo

- Considera l'apprendimento bayesiano di una distribuzione di probabilità nello spazio delle ipotesi
- Calcola la probabilità di ogni ipotesi dai dati forniti e su questa base formula delle predizioni.
- H variabile d'ipotesi, con valori $h_1, h_2 \ \ldots, P(H)$ distribuzione a priori

- La j-esima osservazione d_j fornisce il risultato della variabile casuale D_j (ciliegia o lime) partendo dai dati di training $\mathbf{d} = d_1, \ldots, d_N$ precedentemente in possesso.
- Partendo dai dati disponibili fino a questo momento, la probabilità condizionata di ogni ipotesi si calcola:

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

dove $P(\mathbf{d}|h_i)$ viene chiamato verosimiglianza e **d** sono valori osservati da D

Apprendimento bayesiano completo

- Per effettuare una predizione su una quantità X sconosciuta:

$$\mathbf{P}(X|\mathbf{d}) = \Sigma_i \, \mathbf{P}(X|\mathbf{d}, h_i)P(h_i|\mathbf{d})$$
$$= \Sigma_i \, \mathbf{P}(X|h_i)\boldsymbol{P}(h_i|\mathbf{d})$$
$$= \Sigma_i \, \mathbf{P}(X|h_i)\mathbf{P}(\mathbf{d}|h_i)\mathbf{P}(h_i)/\mathbf{P}(\mathbf{d})$$

- Assumendo che h_i determini una distribuzione di probabilità su X.
- Le predizioni sfruttano una probabilità media ponderata sulla verosimiglianza rispetto alle ipotesi
- Una distribuzione per $P(h_i)$ è <0.1,0.2,0.4,0.2,0.1>
- Es. Supponiamo l'estrazione di caramelle da alcuni sacchetti:

- Di che tipo di sacchetto si tratta? Quale sarà il sapore della prossima caramella?

Apprendimento bayesiano completo

$$\mathbf{P}(X|\mathbf{d}) = \Sigma_i \, \mathbf{P}(X|\mathbf{d}, h_i) P(h_i|\mathbf{d})$$
$$= \Sigma_i \, \mathbf{P}(X|h_i) \boldsymbol{P}(h_i|\mathbf{d})$$
$$= \Sigma_i \, \mathbf{P}(X|h_i) \mathbf{P}(\mathbf{d}|h_i) \mathbf{P}(h_i) / \mathbf{P}(\mathbf{d})$$

- Distribuzione per $P(h_i)$ è <0.1,0.2,0.4,0.2,0.1>
- La verosimiglianza dei dati è calcolata partendo dal presupposto che le osservazioni siano ***indipendentemente e identicamente distribuite***, cosi che:

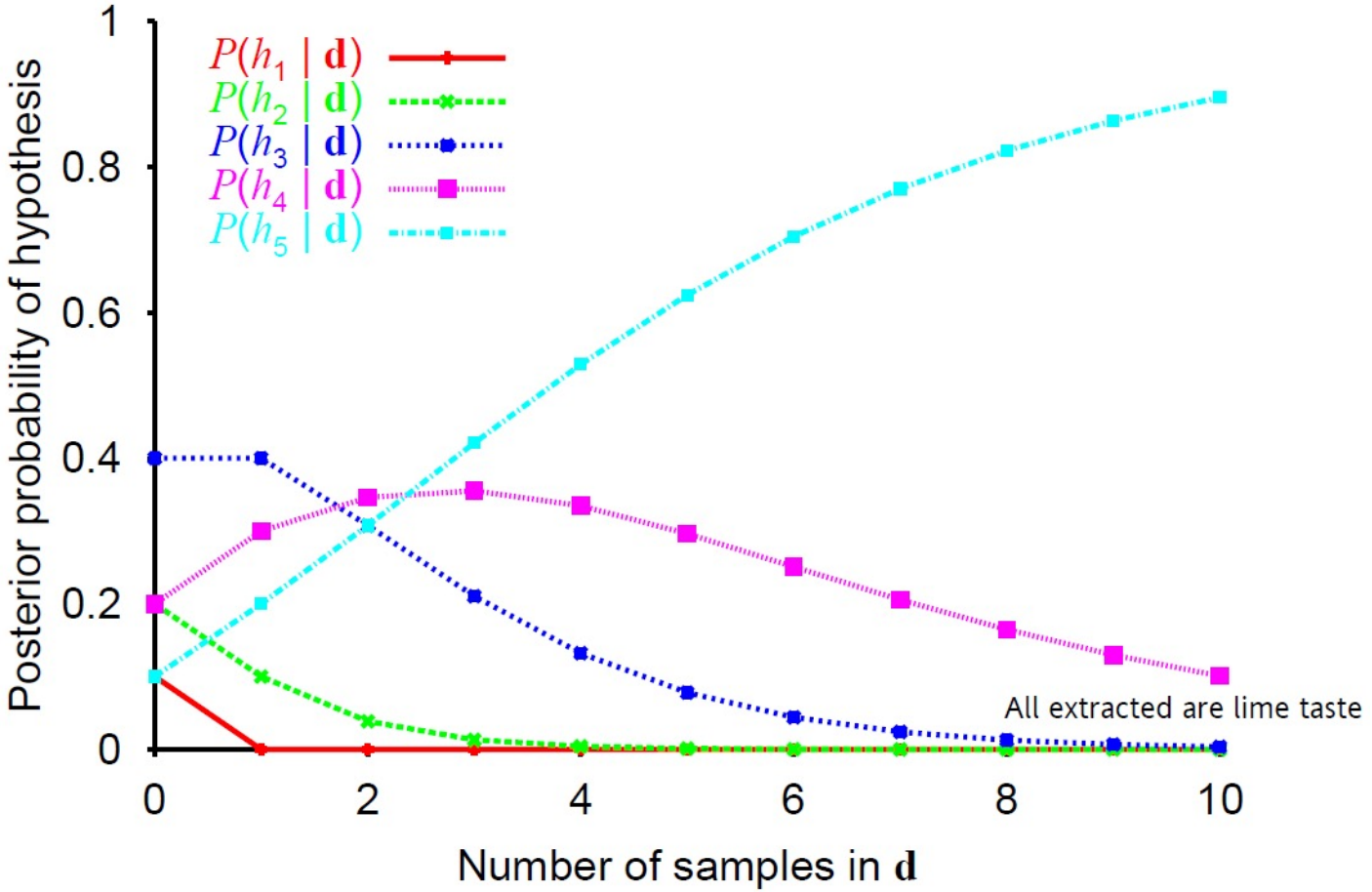
$$\mathbf{P}(\mathbf{d}|h_i) = \prod_j P(d_j, h_i)$$

- Es. Supponiamo l'estrazione di caramelle da alcuni sacchetti:

- $\mathbf{P}(\mathbf{d}|h_3) = 0{,}5^{10}$
- Perché in h_3 la metà delle caramelle sono lime

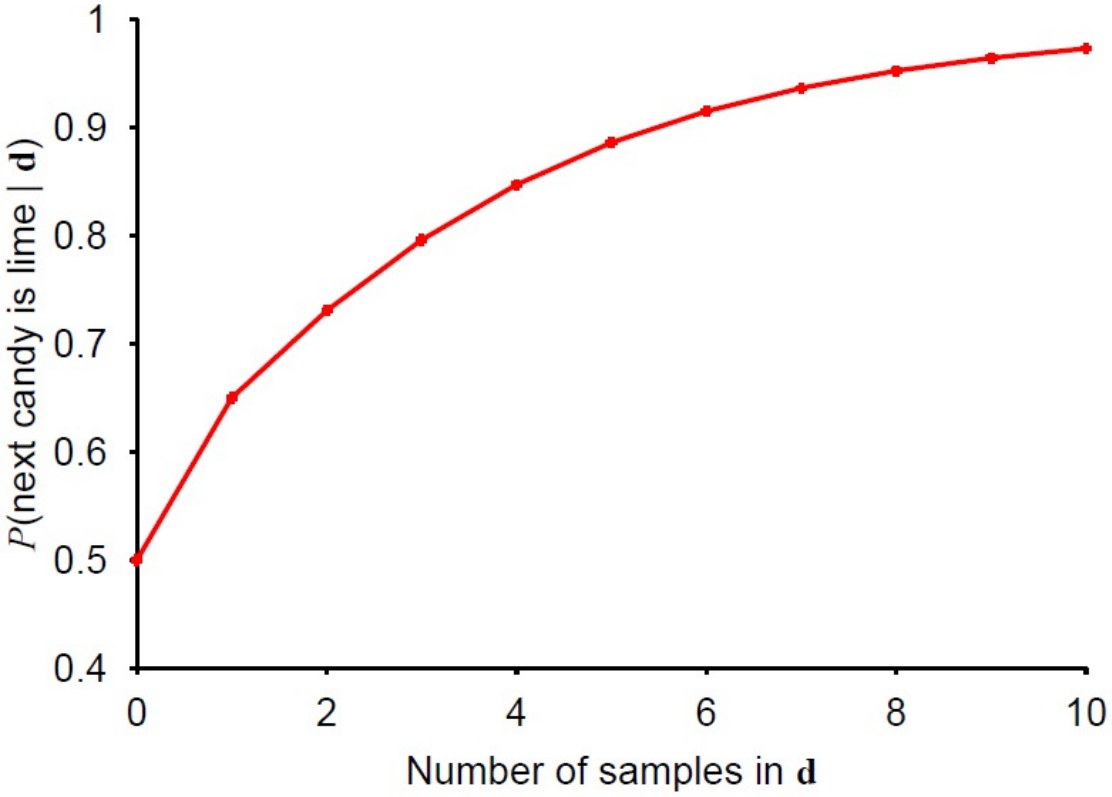
Probabilità condizionata delle ipotesi

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



Predizione di probabilità

$$\mathbf{P}(X = lime|\mathbf{d}) = \Sigma_i \mathbf{P}(X|\mathbf{d}, h_i)P(h_i|\mathbf{d})$$
$$= \Sigma_i \mathbf{P}(X|h_i)\boldsymbol{P}(h_i|\mathbf{d})$$
$$= \Sigma_i \mathbf{P}(X|h_i)\mathbf{P}(d|h_i)\mathbf{P}(h_i)/\mathbf{P}(d)$$



Apprendimento bayesiano completo

- Le ipotesi vere alla fine hanno dominano la predizione bayesiana
  - caratteristica dell'apprendimento bayesiano;

- La predizione bayesiana è ottimale; tuttavia il suo spazio delle ipotesi spesso molto grande o addirittura infinito

(es. 18,446,744,073,709,551,616 funzioni booleane su 6 attributi)

Approssimazione MAP

- Apprendimento Maximum a posteriori (MAP):

  h_{MAP} è h_i che massimizza $P(h_i|\mathbf{d}) \cong P(\mathbf{d}|h_i)P(h_i)$

- Le predizioni con h_{MAP} sono approssimativamente bayesiane $\mathbf{P}(X|\mathbf{d}) \cong \boldsymbol{P}(\boldsymbol{X}|h_{\text{MAP}})$
- Trovare le ipotesi MAP è molto più semplice dell'apprendimento bayesiano
- Nell'esempio $h_{\text{MAP}} = h_5$ dopo aver mangiato 3 caramelle a lime
  - Quindi un agente MAP predirà che la quarta caramella sia lime con probabilità 1 (0,8 è invece la predizione bayesiana), all'aumentare dei dati si avvicina a quella bayesiana

- Entrambe le tecniche fanno uso della distribuzione a priori $P(h_i)$ per ridurre la complessità

- Per le ipotesi deterministiche, $P(\mathbf{d}|h_i)$ vale 1 se consistente, 0 altrimenti

  ⟶ h_{MAP} = l'ipotesi più semplice consistente con i dati

Approssimazione MAP

- Apprendimento Maximum a posteriori (MAP):

$$h_{\text{MAP}} \text{ è } h_i \text{che massimizza } P(h_i|\mathbf{d}) \cong P(\mathbf{d}|h_i)P(h_i)$$

- Equivale a minimizzare $-\log_2 P(\mathbf{d}|h_i) - \log_2 P(h_i)$
  - $\log_2 P(h_i)$ equivale al numero di bit necessari a specificare l'ipotesi h_i
  - $\log_2 P(d|h_i)$ numero di bit aggiuntivi richiesti per la specifica dei dati fissata l'ipotesi h_i

- L'apprendimento MAP sceglie h_i che più comprime i dati, detta anche *minimum description length* (MDL)

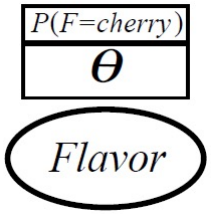
- Nell'esempio di prima $\log_2 P(h_5) = \log_2 1 = 0$ non serve alcun bit

Approssimazione ML

- Per dataset di grandi dimensioni, la distribuzione a priori $P(h_i)$ diventa irrilevante
- Apprendimento con massima verosimiglianza (Maximum Likelihood):
  - Scegliere h_{ML} massimizzando $P(\mathbf{d}|h_i)$, i.e. ottenere in maniera semplice il migliore adattamento ai dati; ipotesi di massima verosimiglianza
- È identico al MAP per la distribuzione a priori, laddove però essa risulti uniforme
  - (che è ragionevole se tutte le ipotesi sono della stessa complessità)
- ML è il metodo “standard" (non-Bayesiano) per l’apprendimento statistico

Apprendimento parametri ML nelle reti Bayesiane

Abbiamo un sacchetto da un nuovo produttore; frazione θ di caramelle alla ciliegia? Qualsiasi θ è possibile: continuum d'ipotesi h_θ



θ è un parametro per questa famiglia di modelli semplici

E' ragionevole adottare l'approccio ML (i due gusti ugualmente prob.)

Supponiamo di scartare N caramelle, c alla ciliegia e $l = N - c$ al lime.

Queste sono osservazioni indipendenti ed identicamente distribuite, pertanto

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

Massimizzandolo con riferimento a θ – che risulta più facile per la verosimiglianza logaritmica

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

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

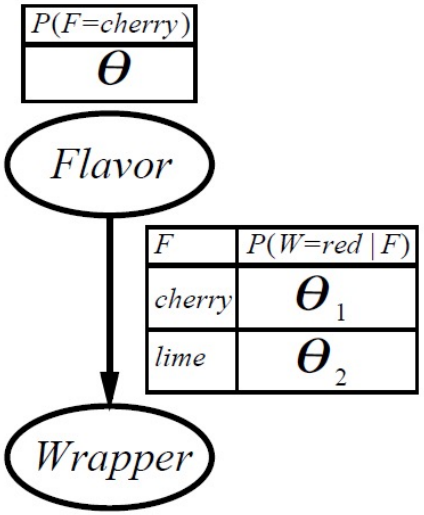
Problema: alcuni eventi potrebbero avere valore 0 qualora non fossero stati osservati

Parametri multipli

L'incartamento rosso/verde dipende probabilisticamente dal sapore.

Probabilità (ad es.) di avere caramelle alla ciliegia nella carta verde:

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



N caramelle, r_c caramelle alla ciliegia in carta rossa, ecc…

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

$$L = [c \log\theta + \ell \log(1-\theta)]$$
$$+ [r_c \log\theta_1 + g_c \log(1-\theta_1)]$$
$$+ [r_\ell \log\theta_2 + g_\ell \log(1-\theta_2)]$$

Parametri multipli

Le derivate di L contengono solo i parametri rilevanti:

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

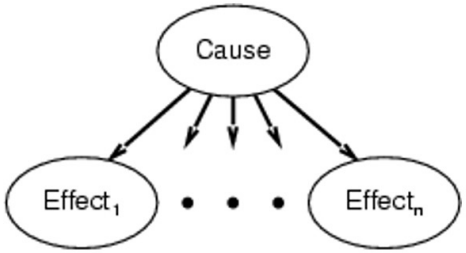
$$\frac{\partial L}{\partial \theta_1} = \frac{r_c}{\theta_1} - \frac{g_c}{1-\theta_1} = 0 \qquad \Rightarrow \quad \theta_1 = \frac{r_c}{r_c+g_c}$$

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

Con dati completi, i parametri possono essere appresi separatamente

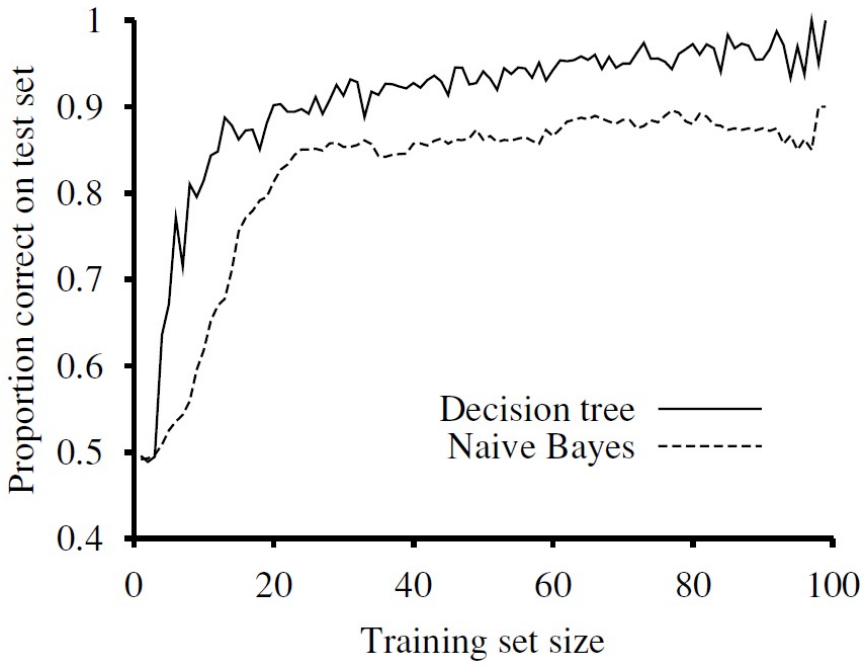
Modelli bayesiani Naive

- Rappresentano il modello di rete bayesiana comunemente usato nel machine learning



- La variabile C da predire è la radice, le variabili attributo x_1 sono le foglie
- Ingenuo perché assume che gli attributi sono condizionalmente indipendenti
- Una predizione deterministica può essere ottenuta scegliendo le classi più probabili
  - $P(C|x_1, x_2, \dots, x_n) = \alpha P(C)\Pi_i P(x_i|C)$

Modelli bayesiani Naive



- Non ha nessuna difficoltà con i dati rumorosi
  - Per *n* attributi booleani, ci sono solo $2n+1$ parametri, non è richiesta nessuna ricerca per trovare h_{ML}

Riassunto

- L'apprendimento bayesiano formula un apprendimento come una forma d'inferenza probabilistica, usando le osservazioni per aggiornare una distribuzione a priori attraverso le ipotesi
- L'apprendimento MAP seleziona una singola ipotesi più probabile, sfruttando i dati di training $P(\mathbf{d}|h_i)P(h_i)$
- Il metodo della massima verosimiglianza (ML) seleziona le ipotesi che massimizzano la verosimiglianza dei dati (uguale a MAP ma con una distribuzione a priori uniforme) $P(\mathbf{d}|h_i)$