Apprendimento Automatico

Metodi Bayesiani

Fabio Aiolli

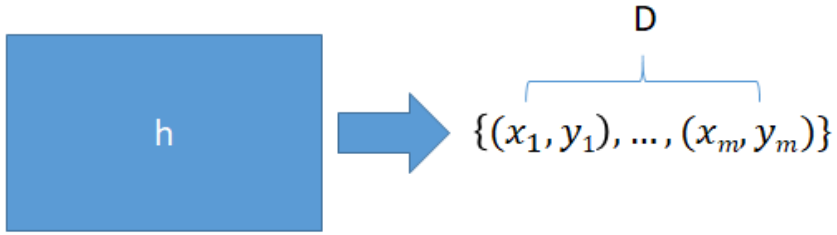
11 Dicembre 2017

Metodi Bayesiani

I metodi Bayesiani forniscono tecniche computazionali di apprendimento (Naive Bayes, Reti Bayesiane, ecc.):

- Gli esempi di training osservati vanno ad incrementare o decrementare la probabilità che una ipotesi sia corretta
- Combinazione di conoscenza a priori sulle ipotesi con dati osservati
- Predizioni di probabilità
- Classificazione mediante combinazione di ipotesi multiple, pesate con la loro probabilità
- Definiscono il caso ideale di predizione ottimale (anche se intrattabile computazionalmente)
- Utili per l'interpretazione di algoritmi non probabilistici
- Difficoltà pratica: necessitano della definizione di molte probabilità. Se esse non sono disponibili inizialmente, devono essere stimate facendo opportune ipotesi sulle distribuzioni.
- Difficoltà pratica: Computazionalmente costosi, richiedono molti esempi per la stima corretta dei parametri

Teorema di Bayes



$$P(h|D) = \frac{P(D|h)P(h)}{P(D)}$$

- $P(h)$: probabilità a priori della ipotesi h
- $P(D)$: probabilità a priori dei dati di apprendimento
- $P(h|D)$: probabilità di h dati D
- $P(D|h)$: probabilità di D data h

Scelta della ipotesi

$$P(h|D) = \frac{P(D|h)P(h)}{P(D)}$$

In generale si vuole selezionare l'ipotesi più probabile dati i dati di apprendimento, detta ipotesi maximum a posteriori h_{MAP}:

$$\begin{aligned} h_{MAP} &= \arg\max_{h \in H} P(h|D) \\ &= \arg\max_{h \in H} \frac{P(D|h)P(h)}{P(D)} \\ &= \arg\max_{h \in H} P(D|h)P(h) \end{aligned}$$

Se assumiamo probabilità uniforme sulle ipotesi, $P(h_i) = P(h_j)$, allora possiamo scegliere la ipotesi maximum likelihood h_{ML}:

$$h_{ML} = \arg\max_{h \in H} P(D|h)$$

Un esempio

Diagnosi medica: probabilità che un dato paziente abbia una particolare forma di tumore

$$P(cancer) = .008 \quad P(\neg cancer) = .992$$
$$P(\oplus|cancer) = .98 \quad P(\ominus|cancer) = .02$$
$$P(\oplus|\neg cancer) = .03 \quad P(\ominus|\neg cancer) = .97$$

Supponiamo di osservare un nuovo paziente per il quale i test di laboratorio hanno dato esito positivo $\oplus$. Quale è la probabilità che il paziente sia effettivamente affetto da tumore?

$$P(cancer|\oplus) \propto P(\oplus|cancer)P(cancer) = .0078$$
$$P(\neg cancer|\oplus) \propto P(\oplus|\neg cancer)P(\neg cancer) = .0298$$

Apprendimento "brute force" della ipotesi MAP

- Per ogni ipotesi $h \in H$, calcola la probabilità a posteriori

$$P(h|D) = \frac{P(D|h)P(h)}{P(D)}$$

- Restituisci l'ipotesi h_{MAP} con la probabilità a posteriori più alta

$$h_{MAP} = \arg\max_{h \in H} P(h|D)$$

Interpretazione Find-S

Si consideri l'apprendimento di concetti (funzioni booleane)

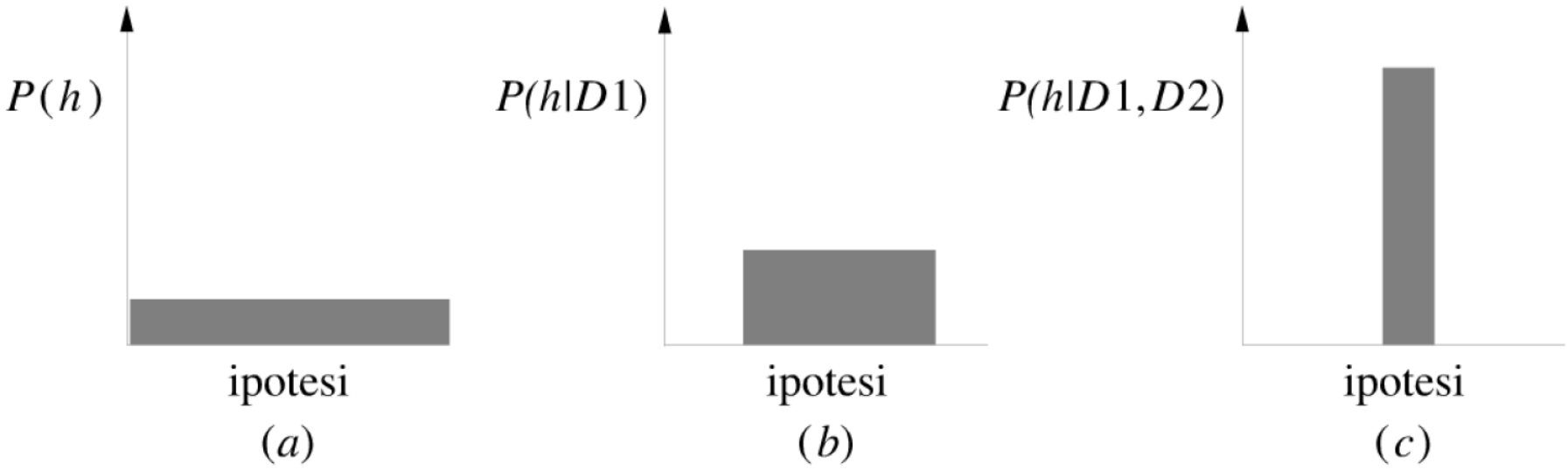
- spazio delle istanze X, spazio delle ipotesi H, esempi di apprendimento D
- si consideri l'algoritmo Find-S (restituisce l'ipotesi più specifica del version space $VS_{H,D}$)

- Quale sarebbe l'ipotesi MAP?
- Corrisponde a quella restituita da Find-S?

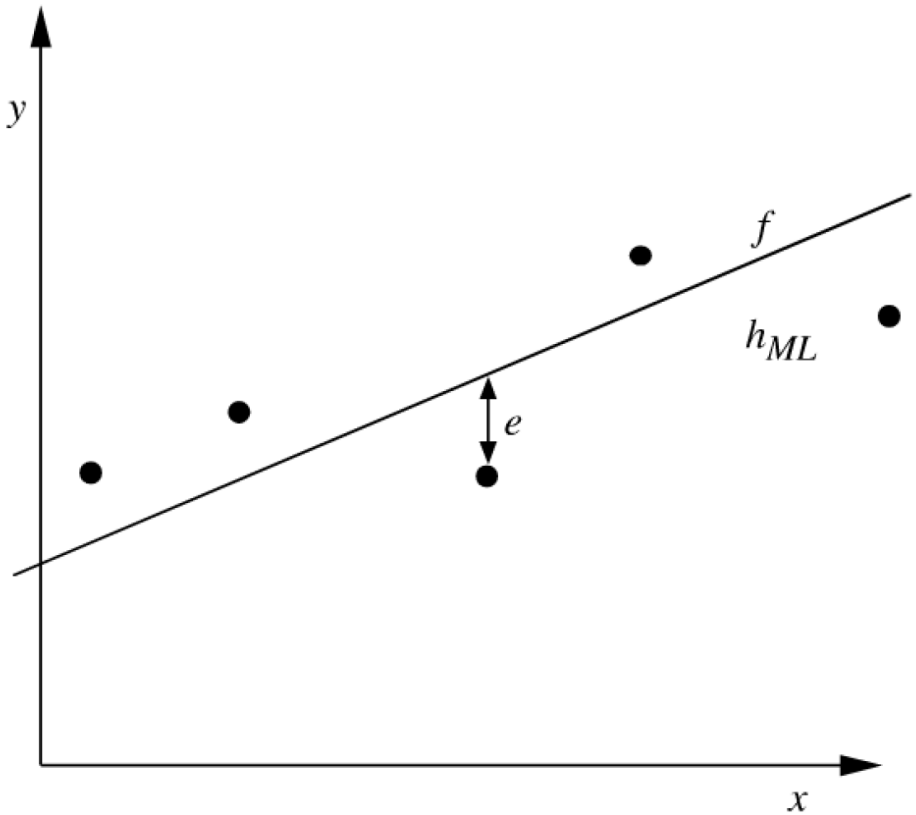
Interpretazione Apprendimento Concetti

- Assumiamo di fissare le istanze $\langle \mathbf{x}_1, \cdots, \mathbf{x}_m \rangle$
- Assumiamo D essere l'insieme dei valori desiderati (noise-free)
- Scegliamo $P(D|h)$:
  - $P(D|h) = 1$ se h è consistente con D, altrimenti $P(D|h) = 0$
- Scegliamo $P(h) = \frac{1}{|H|}$ per tutte le $h \in H$ (per Find-S definiamo $P(h_i) < P(h_j)$ se $h_i >_g h_j$)
- Allora, $P(h|D) = \begin{cases} \frac{1}{|VS_{H,D}|} & \text{se } h \text{ consistente con } D \\ 0 & \text{altrimenti} \end{cases}$

Evoluzione della probabilità a posteriori



Apprendimento di una funzione a valori reali

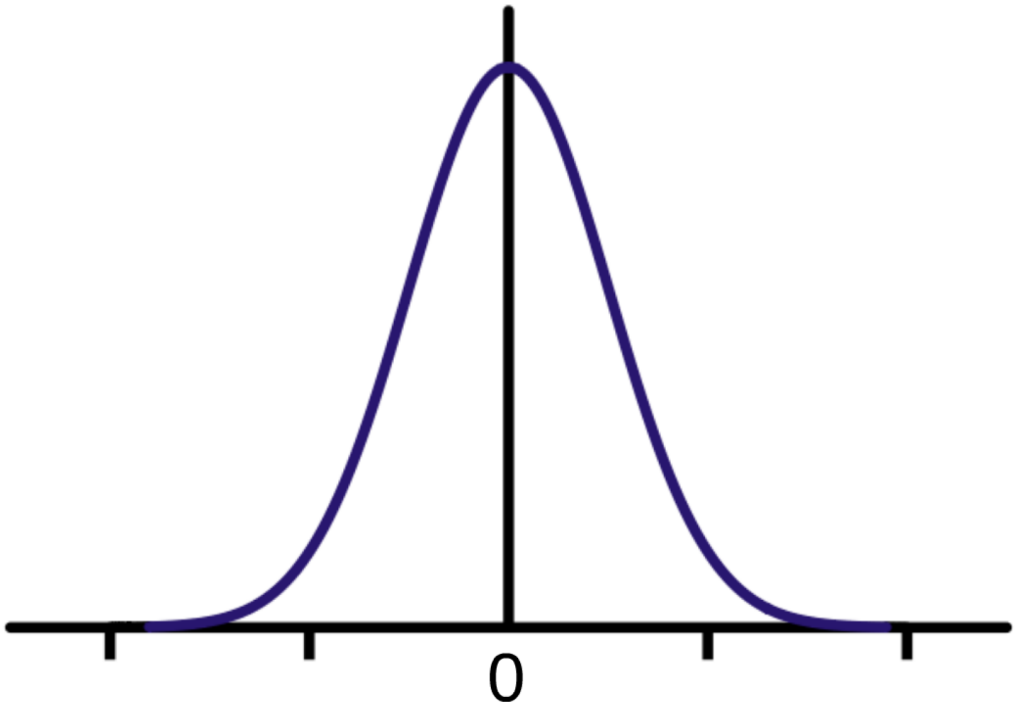


Si consideri una qualunque funzione target f a valori reali, esempi di apprendimento $\langle \mathbf{x}_i, d_i \rangle$, dove d_i presenta del rumore,

- $d_i = f(\mathbf{x}_i) + e_i$
- e_i è una variabile random (rumore) estratta indipendentemente per ogni $\mathbf{x}_i$ secondo una distribuzione Gaussiana con media 0.

Allora l'ipotesi h_{ML} (maximum likelihood) è quella che minimizza:

$$h_{ML} = \arg\min_{h \in H} \sum_{i=0}^{m} (d_i - h(\mathbf{x}_i))^2$$



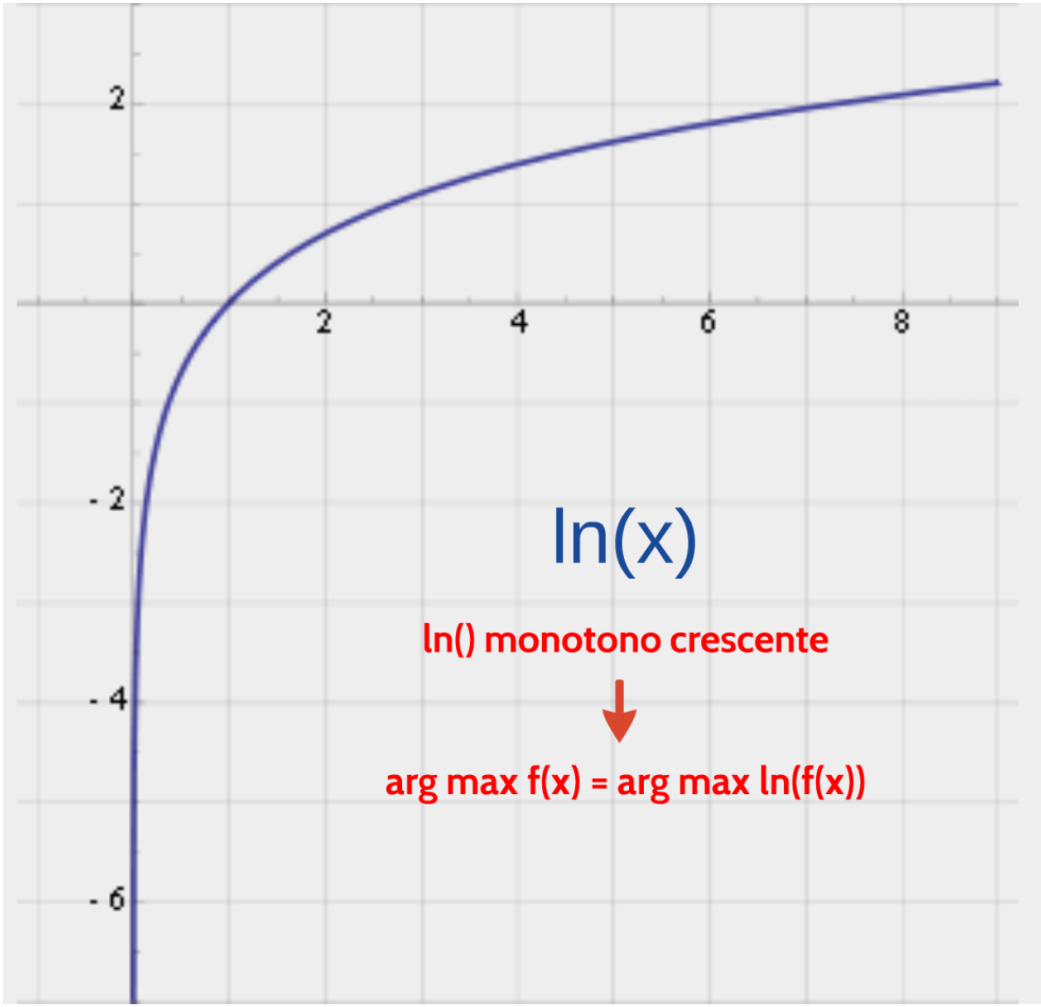
$$p(d_i|h) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(\overbrace{d_i - h(\mathbf{x}_i)}^{e_i})}$$

Apprendimento di una funzione a valori reali

$$
\begin{aligned}
h_{ML} &= \arg\max_{h \in H} p(D|h) \\
&= \arg\max_{h \in H} \prod_{i=1}^{m} p(d_i|h) \\
&= \arg\max_{h \in H} \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(d_i - h(\mathbf{x}_i))^2}
\end{aligned}
$$

che si tratta meglio massimizzando il logaritmo naturale..

Apprendimento di una funzione a valori reali



Apprendimento di una funzione a valori reali

$$
\begin{aligned}
h_{ML} &= \arg\max_{h \in H} \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(d_i - h(\mathbf{x}_i))^2} \\
&= \arg\max_{h \in H} \ln \left(\prod_{i=1}^{m} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(d_i - h(\mathbf{x}_i))^2} \right) \\
&= \arg\max_{h \in H} \sum_{i=1}^{m} \ln \left(\frac{1}{\sqrt{2\pi\sigma^2}} \right) - \frac{1}{2\sigma^2}(d_i - h(\mathbf{x}_i))^2 \\
&= \arg\max_{h \in H} \sum_{i=1}^{m} -\frac{1}{2\sigma^2}(d_i - h(\mathbf{x}_i))^2 \\
&= \arg\min_{h \in H} \sum_{i=1}^{m} \frac{1}{2\sigma^2}(d_i - h(\mathbf{x}_i))^2 = \arg\min_{h \in H} \sum_{i=1}^{m} (d_i - h(\mathbf{x}_i))^2
\end{aligned}
$$

Apprendimento di una ipotesi che predice probabilità

$$P(D|h) = \prod_{i=1}^{m} P(\mathbf{x}_i, d_i|h) = \prod_{i=1}^{m} P(d_i|h, \mathbf{x}_i) P(\mathbf{x}_i)$$

$$P(d_i|h, \mathbf{x}_i) = \begin{cases} h(\mathbf{x}_i) & \text{if } d_i = 1 \\ 1 - h(\mathbf{x}_i) & \text{if } d_i = 0 \end{cases} = h(\mathbf{x}_i)^{d_i}(1 - h(\mathbf{x}_i))^{1-d_i}$$

$$P(D|h) = \prod_{i=1}^{m} h(\mathbf{x}_i)^{d_i}(1 - h(\mathbf{x}_i))^{1-d_i} P(\mathbf{x}_i)$$

$$\begin{aligned} h_{ML} &= \arg\max_{h \in H} \prod_{i=1}^{m} h(\mathbf{x}_i)^{d_i}(1 - h(\mathbf{x}_i))^{1-d_i} \\ &= \arg\max_{h \in H} \underbrace{\sum_{i=1}^{m} d_i \ln(h(\mathbf{x}_i)) + (1 - d_i) \ln(1 - h(\mathbf{x}_i))}_{-\text{cross entropy}} \end{aligned}$$

Classificazione più probabile per nuove istanze

Finora abbiamo cercato l'ipotesi più probabile dati i dati D (cioè h_{MAP})

Data una nuova istanza $\mathbf{x}$, quale è la classificazione più probabile?

Non necessariamente $h_{MAP}(\mathbf{x})$ è la classificazione più probabile..

Consideriamo per esempio la situazione seguente:

- tre possibili ipotesi:

$$P(h_1|D) = 0.4, \quad P(h_2|D) = 0.3, \quad P(h_3|D) = 0.3$$

- data una nuova istanza $\mathbf{x}$,

$$h_1(\mathbf{x}) = \oplus, \quad h_2(\mathbf{x}) = \ominus, \quad h_3(\mathbf{x}) = \ominus$$

- quale è la classificazione più probabile per $\mathbf{x}$?

Classificazione ottima di Bayes

Sia data una classe $v_j \in V$, otteniamo:

$$P(v_j|D) = \sum_{h_i \in H} P(v_j|h_i)P(h_i|D)$$

da cui deriva che la classificazione ottima (di Bayes) di una certa istanza è la classe $v_j \in V$ che massimizza tale probabilità, ovvero:

$$\arg\max_{v_j \in V} \sum_{h_i \in H} P(v_j|h_i)P(h_i|D)$$

Esempio di classificazione ottima di Bayes

$$v_{Bayes} = \arg\max_{v_j \in V} \sum_{h_i \in H} P(v_j|h_i)P(h_i|D)$$

Esempio:

$$P(h_1|D) = 0.4, \quad P(\ominus|h_1) = 0, \quad P(\oplus|h_1) = 1$$
$$P(h_2|D) = 0.3, \quad P(\ominus|h_2) = 1, \quad P(\oplus|h_1) = 0$$
$$P(h_3|D) = 0.3, \quad P(\ominus|h_3) = 1, \quad P(\oplus|h_3) = 0$$

pertanto:

$$\sum_{h_i \in H} P(\oplus|h_i)P(h_i|D) = 0.4 \qquad \sum_{h_i \in H} P(\ominus|h_i)P(h_i|D) = 0.6$$

e quindi:

$$v_{Bayes} = \arg\max_{v_j \in \{\ominus,\oplus\}} \sum_{h_i \in H} P(v_j|h_i)P(h_i|D) = \ominus$$

Classificatore di Gibbs

Il classificatore ottimo di Bayes può essere molto costoso da calcolare se ci sono molte ipotesi
Algoritmo di Gibbs:

- Scegliere una ipotesi a caso, con probabilità $P(h|D)$
- Usarla per classificare la nuova istanza

Fatto piuttosto sorprendente: assumendo che i concetti target siano estratti casualmente da H secondo una probabilità a priori su H, allora:

$$E[\epsilon_{Gibbs}] \leq 2E[\epsilon_{Bayes}]$$

Supponendo una distribuzione a priori uniforme sulle ipotesi corrette in H,

- Seleziona una qualunque ipotesi da VS, con probabilità uniforme
- Il suo errore atteso non è peggiore del doppio dell'errore ottimo di Bayes!