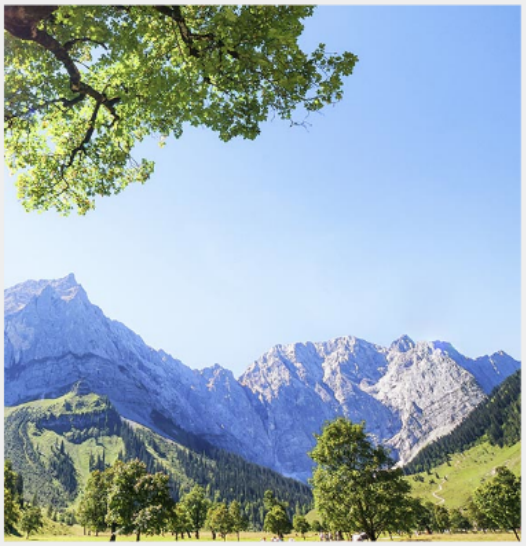
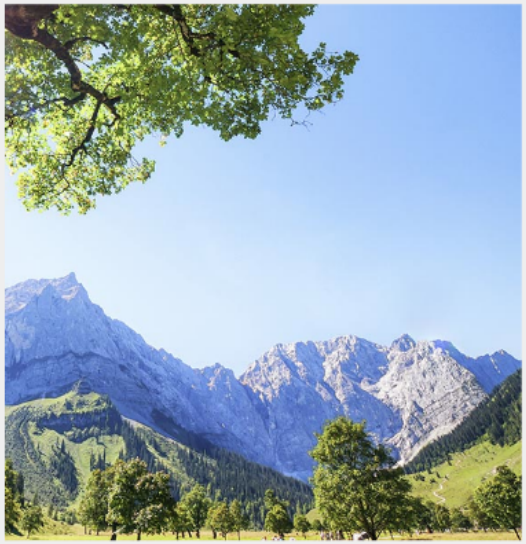
Apprendimento Bayesiano



[Capitolo 6, Mitchell]

- Teorema di Bayes
- Ipotesi MAP e ML
- algoritmi di apprendimento MAP
- Principio MDL (Minimum description length)
- Classificatore Ottimo di Bayes
- Apprendimento Ingenuo di Bayes (Naive Bayes)
- Richiamo di Reti Bayesiane
- Algoritmo EM (Expectation Maximization)

Metodi Bayesiani



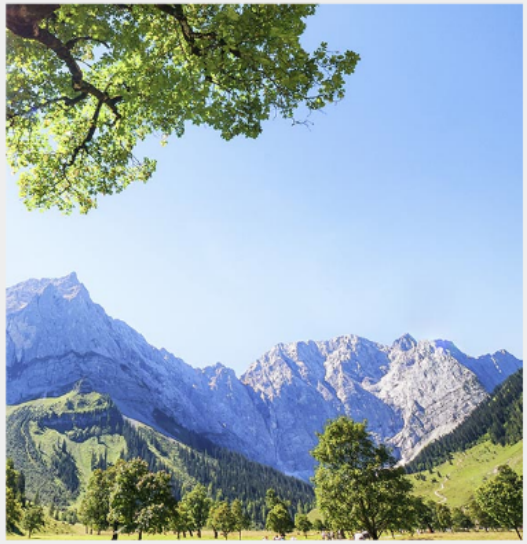
Forniscono metodi computazionali di apprendimento:

- Apprendimento Naive Bayes
- Apprendimento di Reti Bayesiane
- Combinazione di conoscenza a priori (probabilità a priori) con dati osservati
- Richiedono probabilità a priori

Forniscono un framework concettuale utile

- Forniscono il “gold standard” per la valutazione di altri algoritmi di apprendimento
- Interpretazione del “Rasoio di Occam”

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 D
- $P(h|D)$ = probabilità di h dati D
- $P(D|h)$ = probabilità di D data h

Scelta 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

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 $P(h_i) = P(h_j)$ allora si può ulteriormente semplificare, e scegliere la ipotesi *"Maximum likelihood"* (ML)

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

Apprendimento "Bruta Forza" dell'ipotesi MAP

1. Per ogni ipotesi h in H, calcola la probabilità a posteriori

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

2. 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 Find-S

Assumiamo di fissare le istanze $\langle x_1, \ldots, x_m \rangle$

Assumiamo D essere l'insieme dei valori desiderati $D = \langle c(x_1), \ldots, c(x_m) \rangle$

Scegliamo $P(D|h)$:

- $P(D|h) = 1$ **se** h è consistente con D, **altrimenti** $P(D|h) = 0$

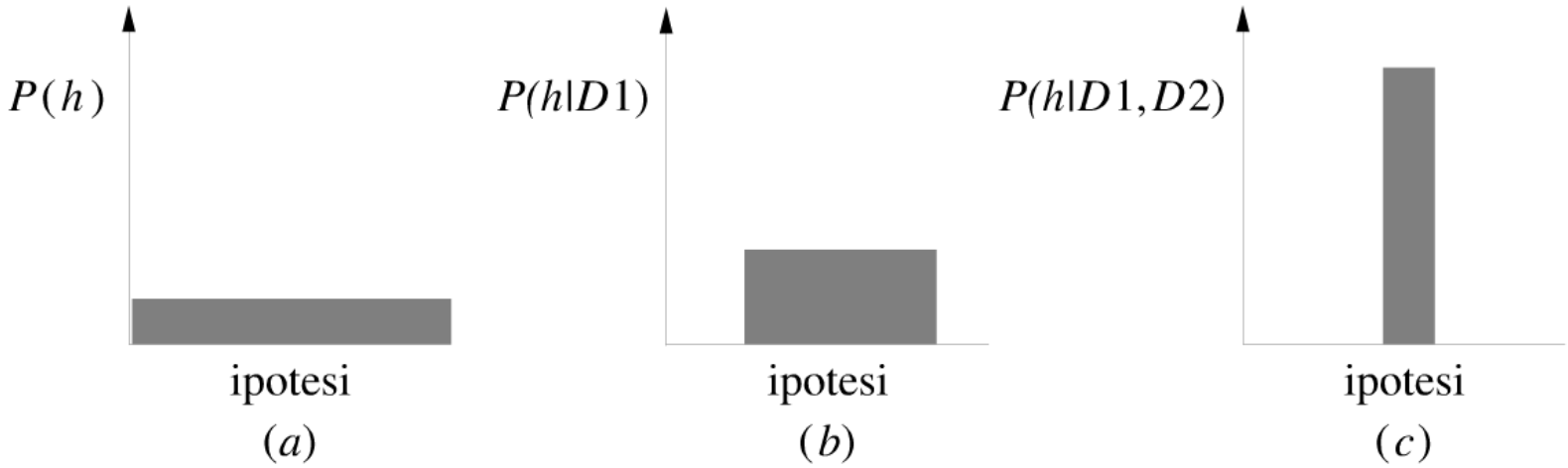
Scegliamo $P(h)$ essere la distribuzione *uniforme*

- $P(h) = \frac{1}{|H|}$ per tutte le h in H (per FIND-S, definire $P(h_i) < P(h_j)$ se $h_i >_g h_j$)

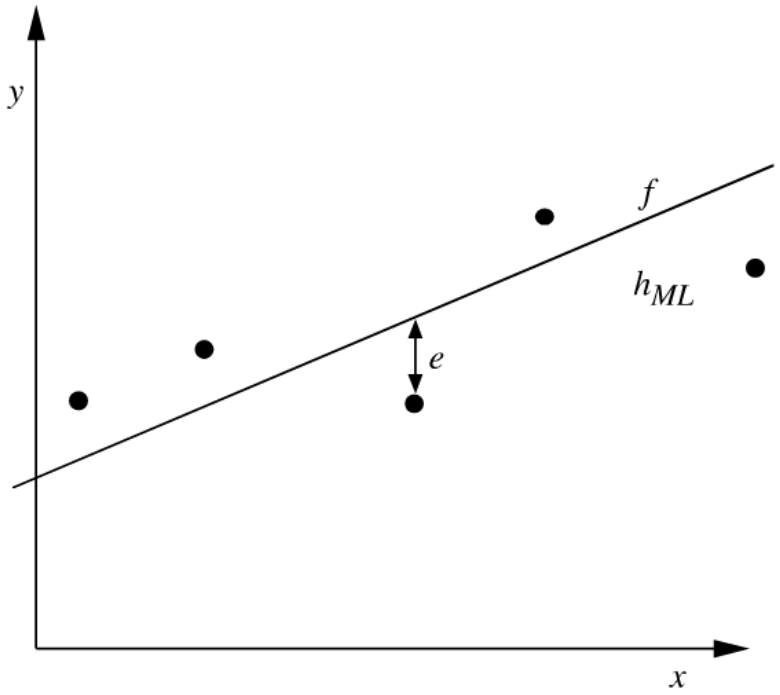
Allora,

$$P(h|D) = \begin{cases} \frac{1}{|VS_{H,D}|} & \text{se } h \text{ è consistente con } D \\ 0 & \text{altrimenti} \end{cases}$$

Evoluzione delle probabilità a posteriori



Apprendimento di una Funzione a Valori Reali

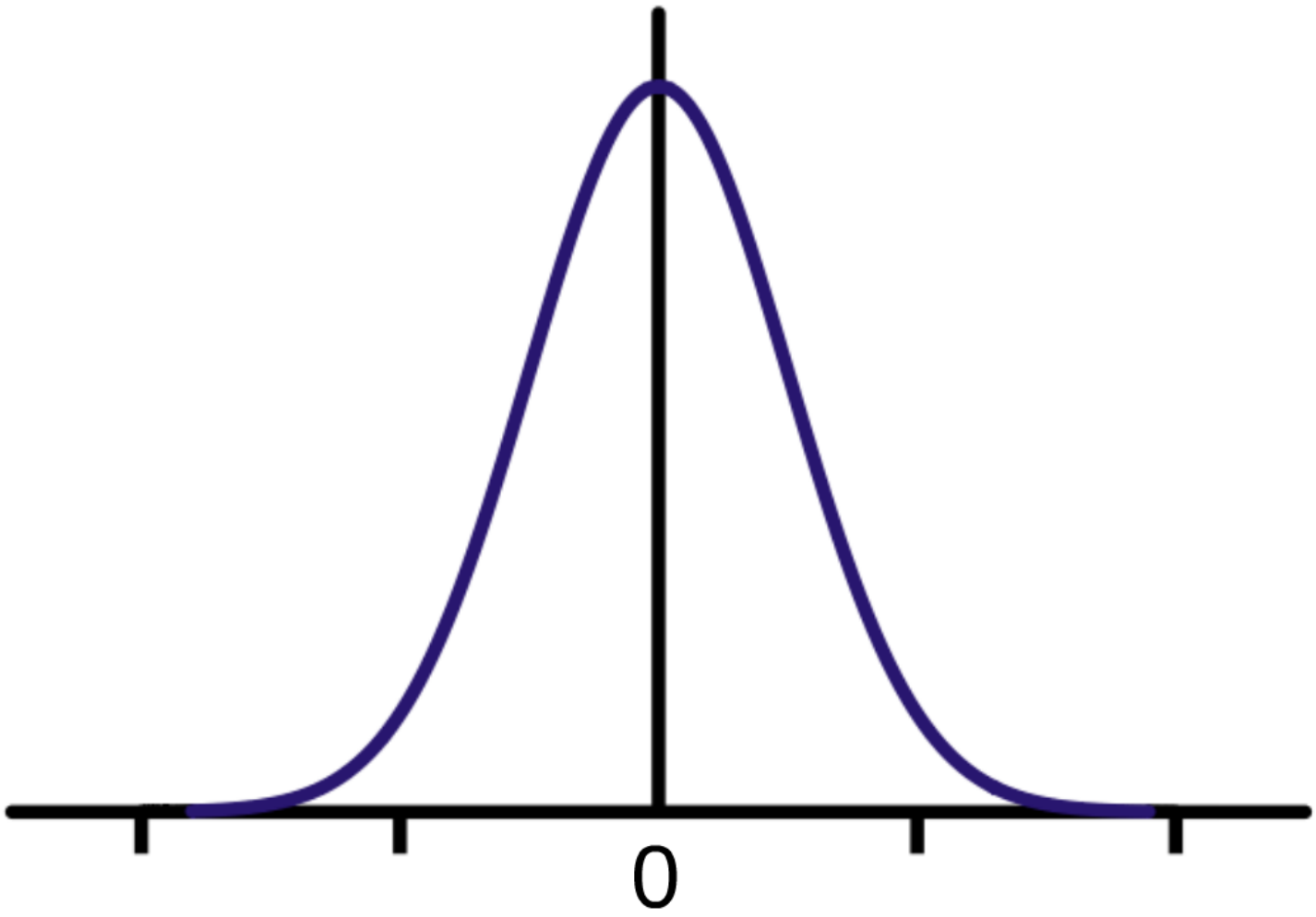


Si consideri una qualunque funzione target f a valori reali, esempi di apprendimento $\langle x_i, d_i \rangle$, dove d_i presenta del rumore

- $d_i = f(x_i) + e_i$
- e_i è una variabile random (rumore) estratta indipendente per ogni x_i secondo una distribuzione Gaussiana con media 0

Allora l'ipotesi h_{ML} è quella che minimizza:

$$h_{ML} = \arg\min_{h \in H} \sum_{i=1}^{m} (d_i - h(x_i))^2$$

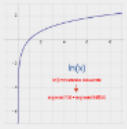


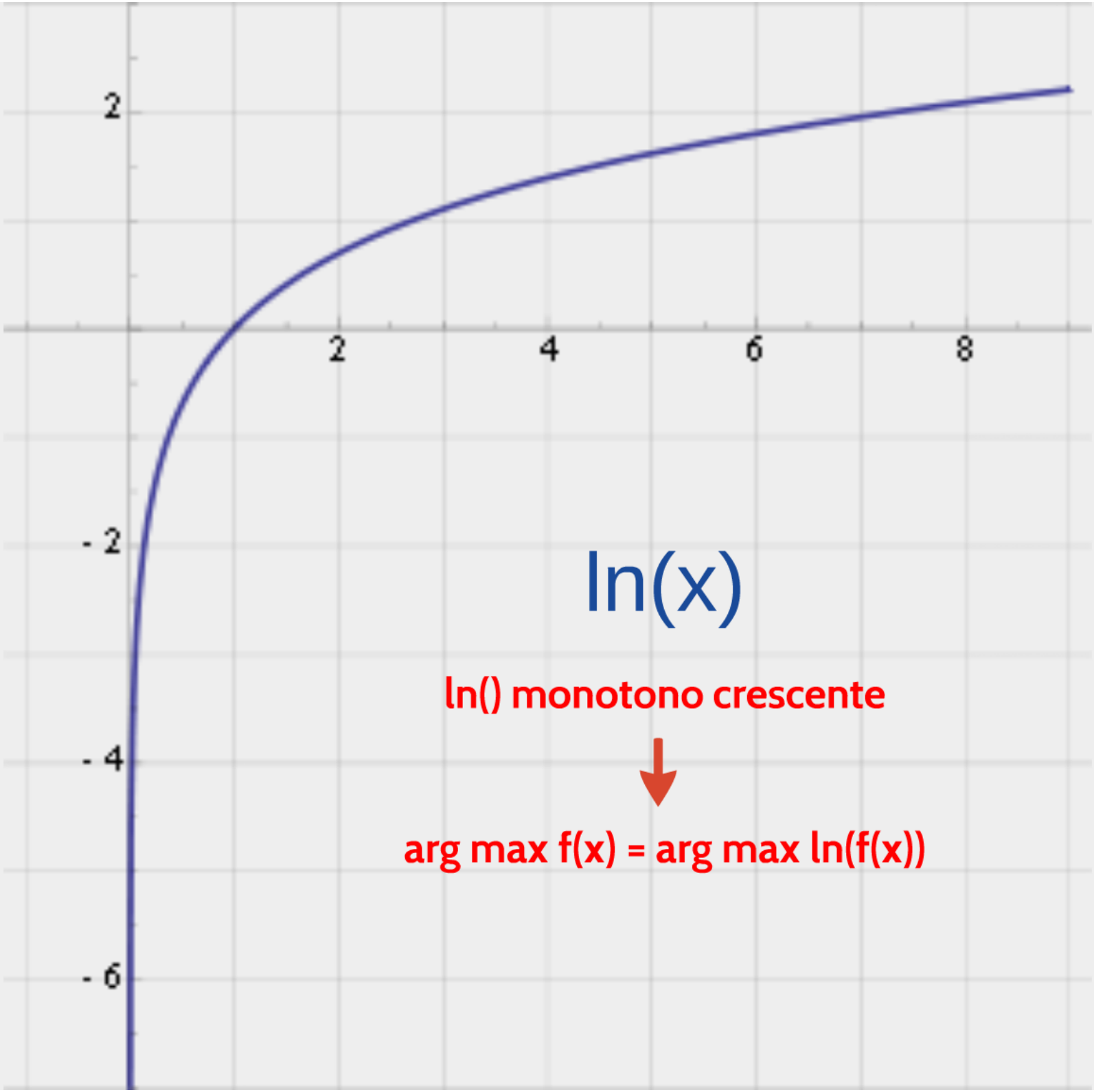
$$\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(\overbrace{d_i - h(x_i)}^{e_i})^2}$$

Apprendimento di una Funzione a Valori Reali

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

che si tratta meglio massimizzando il logaritmo naturale...




2
2
4
6
8
- 2
- 4
- 6
ln(x)
ln() monotono crescente
arg max f(x) = arg max ln(f(x))

Apprendimento di una Funzione a Valori Reali

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

costante indipendente da h()

max -f(x) = min f(x)

fattore di scala: non modifica l'ottimo

Principio MDL (Minimum Description Length)

Rasoio di Occam: preferire l'ipotesi più semplice

MDL: preferire l'ipotesi h che minimizza

$$h_{MDL} = \arg\min_{h \in H} L_{C_1}(h) + L_{C_2}(D|h)$$

dove $L_C(x)$ è la lunghezza di descrizione di x sotto la codifica C

Esempio: H = alberi di decisione, D = etichette di allenamento

- $L_{C_1}(h)$ è # bit per descrivere l'albero h
- $L_{C_2}(D|h)$ è # bit per descrivere D dato h
  - Notare che $L_{C_2}(D|h) = 0$ se gli esempi sono classifi cati perfettamente da h. Basta descrivere solo le eccezioni
- Quindi h_{MDL} cerca un compromesso fra dimensione albero e numero errori

Principio MDL (Minimum Description Length)

$$
\begin{aligned}
h_{MAP} &= \arg\max_{h \in H} P(D|h)P(h) \\
&= \arg\max_{h \in H} \log_2(P(D|h)) + \log_2(P(h)) \\
&= \arg\min_{h \in H} -\log_2(P(D|h)) - \log_2(P(h))
\end{aligned}
$$

Dalla Teoria dell'Informazione:

> Il codice ottimo (il codice con lunghezza aspettata pi`u corta) per un evento con probabilità p è $-\log_2 p$ bit.

Pertanto:

- $-\log_2(P(h))$ è la lunghezza di h usando un codice ottimo
- $-\log_2(P(D|h))$ è la lunghezza di D dato h usando un codice ottimo

$\Rightarrow$ preferire l'ipotesi che minimizza

$$lunghezza(h) + lunghezza(errori)$$

Classificazione più probabile per nuove istanze

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

Data una nuova istanza x, qual' è la *classificazione* più probabile ?

- $h_{MAP}(x)$ non è la classificazione più probabile!

Consideriamo:

- 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 x,

$$h_1(x) = \oplus, \qquad h_2(x) = \ominus, \qquad h_3(x) = \ominus$$

- qual' è la classificazione più probabile per x?

Classificazione Ottima di Bayes

Classificazione Ottima di Bayes:

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

Esempio:

$$\begin{array}{lll} P(h_1|D) = 0.4, & P(\ominus|h_1) = 0, & P(\oplus|h_1) = 1 \\ P(h_2|D) = 0.3, & P(\ominus|h_2) = 1, & P(\oplus|h_2) = 0 \\ P(h_3|D) = 0.3, & P(\ominus|h_3) = 1, & P(\oplus|h_3) = 0 \end{array}$$

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

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

Classificatore di Gibbs

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

Algoritmo di Gibbs:

1. Scegliere una ipotesi a caso, secondo $P(h|D)$
2. Usarla per classificare la nuova istanza

Fatto sorprendente: assumiamo che i concetti target siano estratti casualmente da H secondo una probabilità a priori su H. Allora:

$$E[errore_{Gibbs}] \leq 2E[errore_{BayesOttimo}]$$

Supponendo distribuzione a priori uniforme su ipotesi corrette in H,

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