

Lezione 6

Altri metodi di classificazione. Bayesian Learning Martedì, 16 Novembre 2004

Giuseppe Manco

Readings:

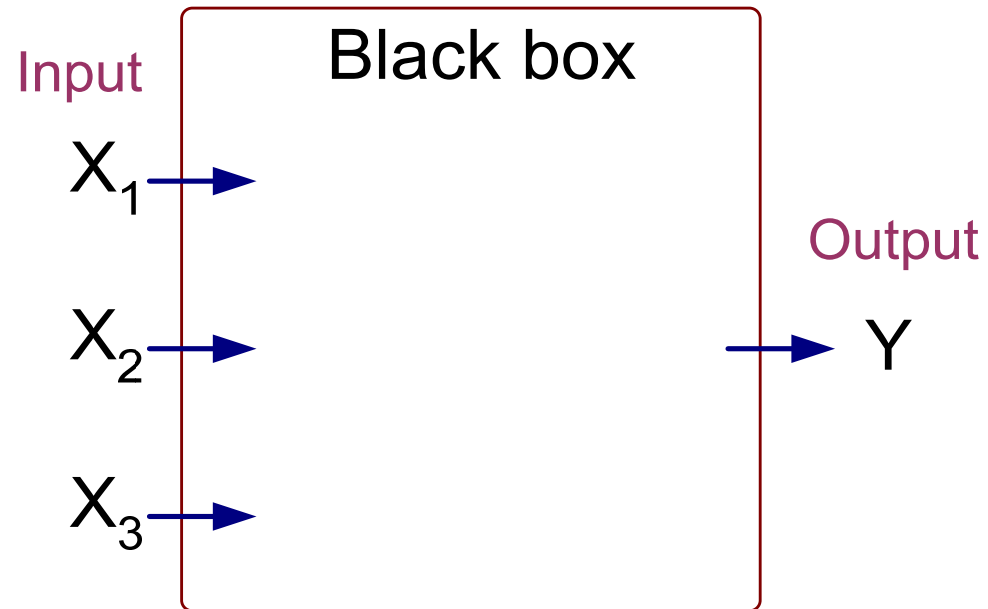
Sections 6.1-6.5, Mitchell

Chapter 2, Bishop

Chapter 4, Hand-Mannila-Smith

Artificial Neural Networks (ANN)

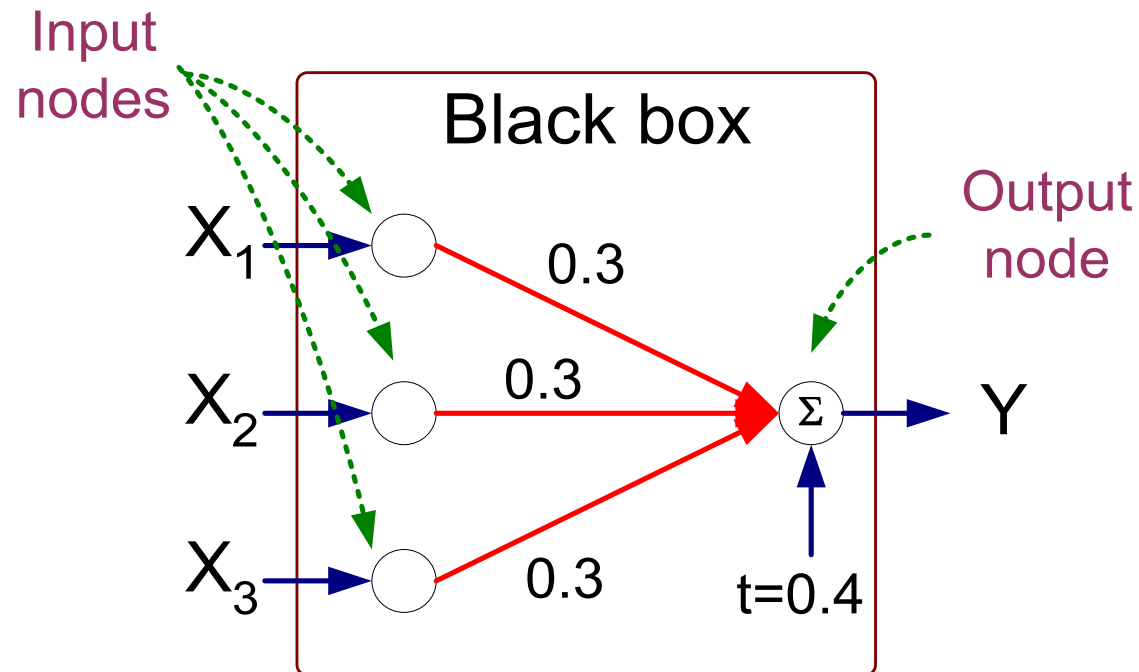
X_1	X_2	X_3	Y
1	0	0	0
1	0	1	1
1	1	0	1
1	1	1	1
0	0	1	0
0	1	0	0
0	1	1	1
0	0	0	0



L'output Y è 1 se almeno due dei tre input è uguale a 1.

Artificial Neural Networks (ANN)

X_1	X_2	X_3	Y
1	0	0	0
1	0	1	1
1	1	0	1
1	1	1	1
0	0	1	0
0	1	0	0
0	1	1	1
0	0	0	0

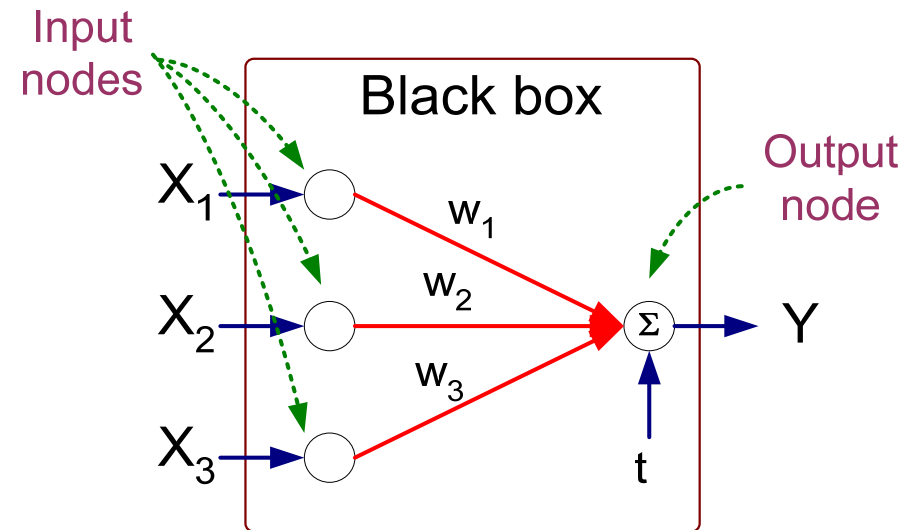


$$Y = I(0.3X_1 + 0.3X_2 + 0.3X_3 - 0.4 > 0)$$

$$\text{dove } I(z) = \begin{cases} 1 & \text{se } z \text{ è true} \\ 0 & \text{altrimenti} \end{cases}$$

Artificial Neural Networks (ANN)

- Il modello è un **assemblaggio di nodi interconnessi con associati dei pesi**
- L'output di ogni connessione somma ognuno dei valori in input, proporzionati dai pesi dei links

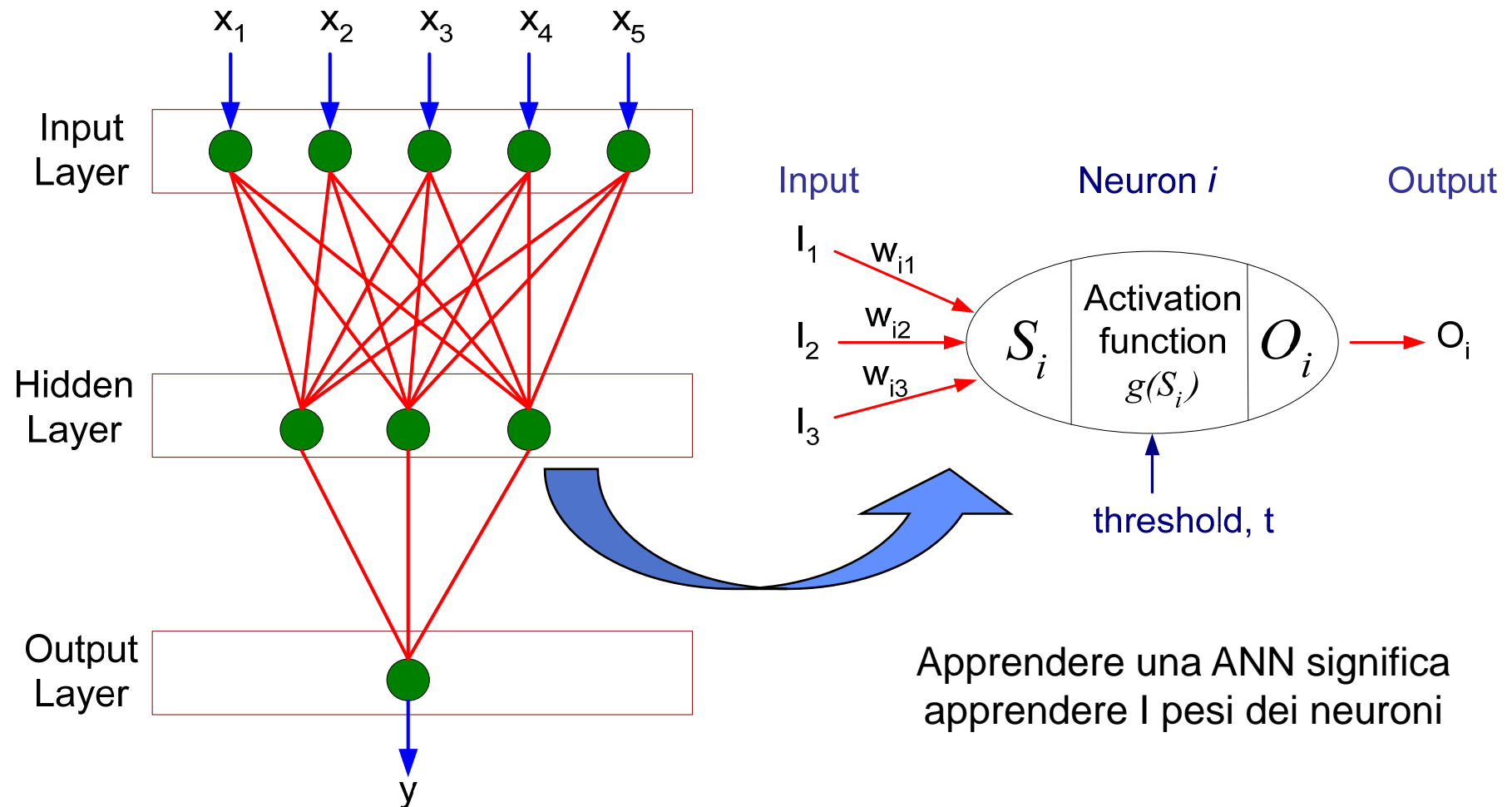


Perceptron Model

$$Y = I\left(\sum_i w_i X_i - t\right) \quad 0$$

$$Y = \text{sign}\left(\sum_i w_i X_i - t\right)$$

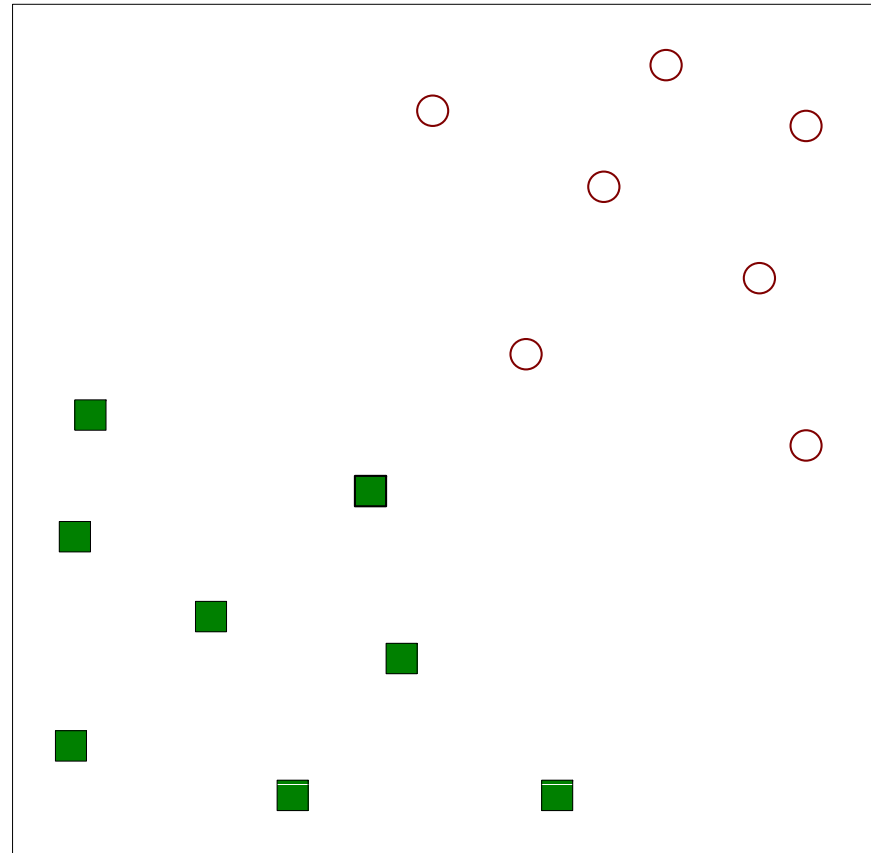
Sruttura generale di una ANN



Apprendimento di una ANN

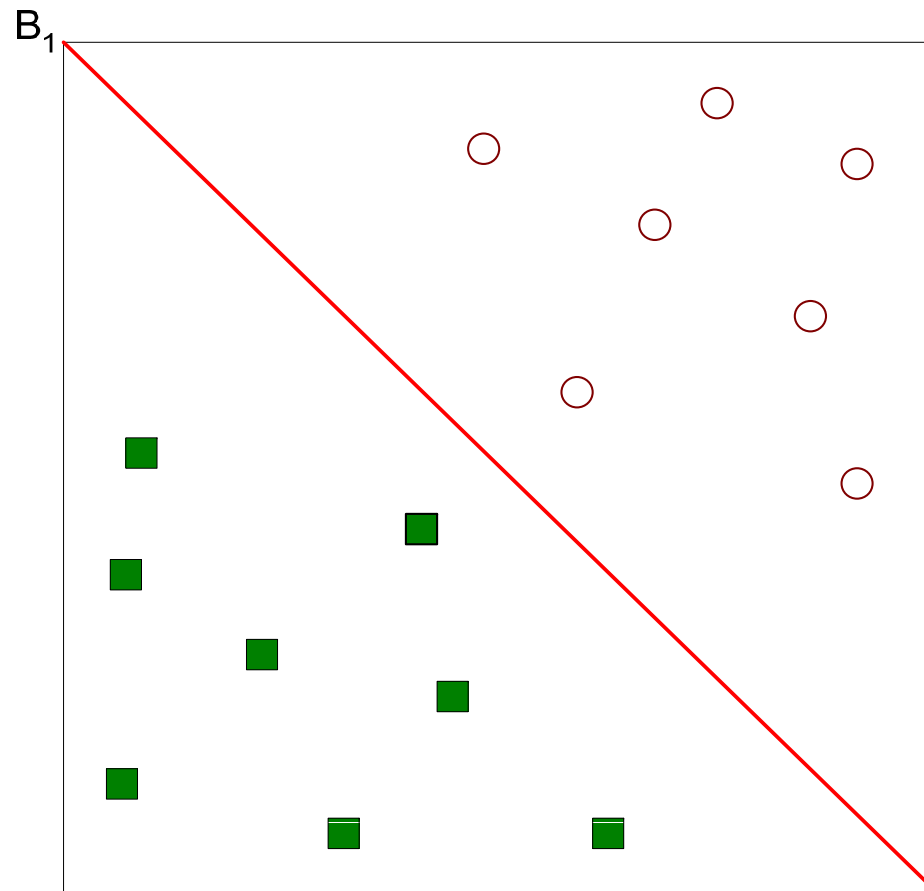
- Si inizializza l'insieme dei pesi (w_0, w_1, \dots, w_k)
- Si aggiustano i pesi in modo da rendere l'output di una rete consistente con le etichette di classe degli esempi del training set
 - **Funzione obiettivo:**
$$E = \sum_i [Y_i - f(w_i, X_i)]^2$$
 - **Trovare i pesi w_i 's che minimizzano la funzione obiettivo**
 - e.g., algoritmo backpropagation

Support Vector Machines



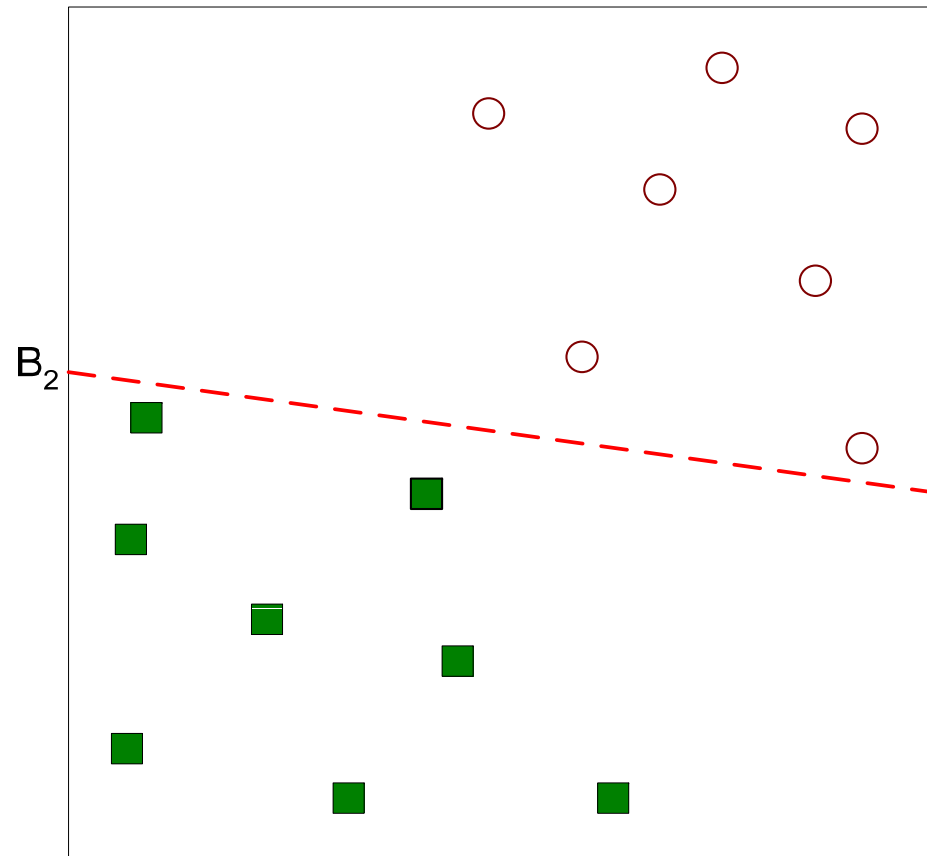
- **Obiettivo: trovare un iperpiano lineare (un decision boundary) che separa le classi**

Support Vector Machines



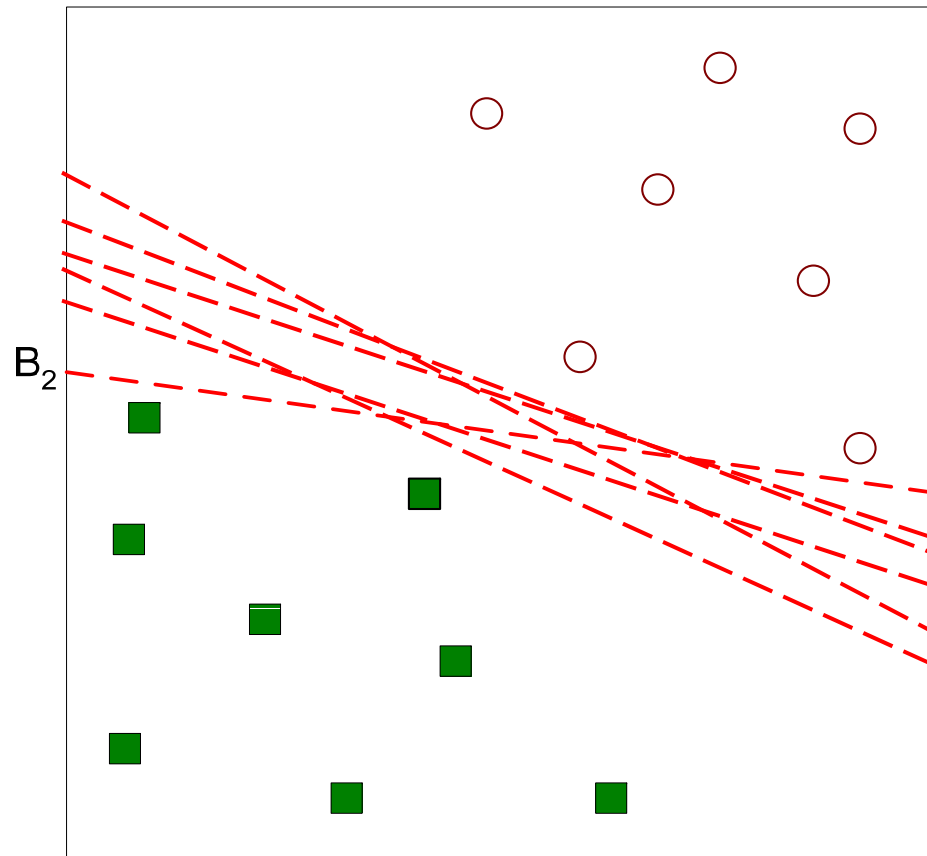
- Una possibile soluzione

Support Vector Machines



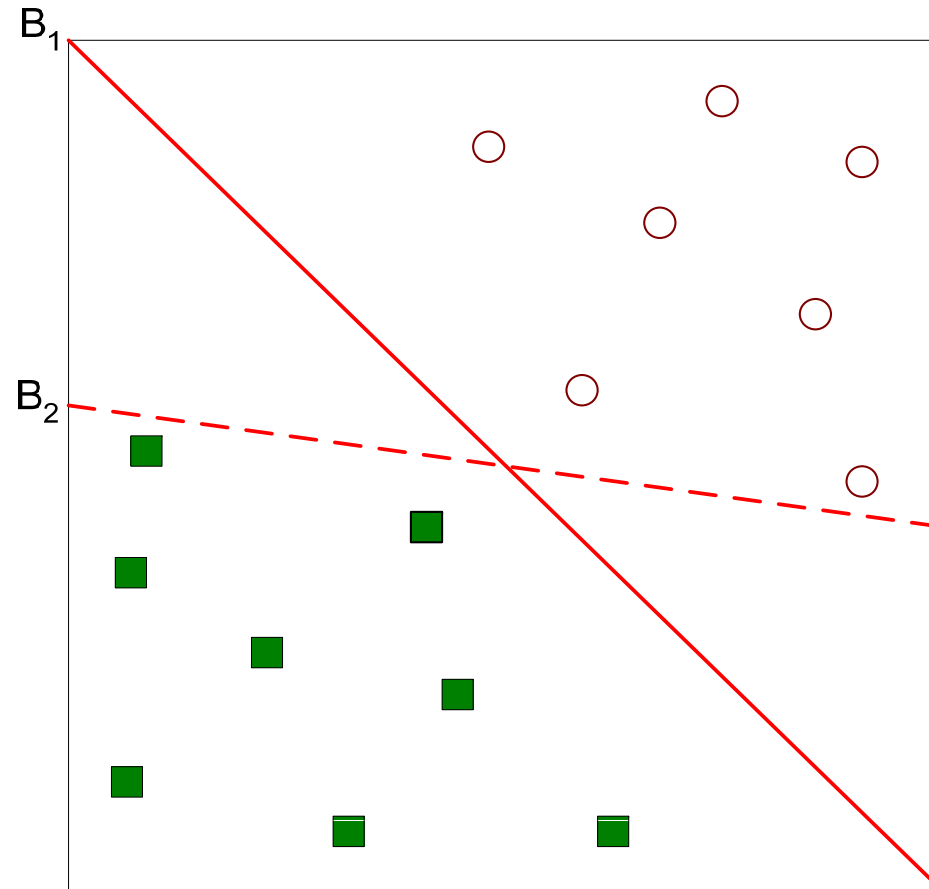
- Un'altra soluzione possibile

Support Vector Machines



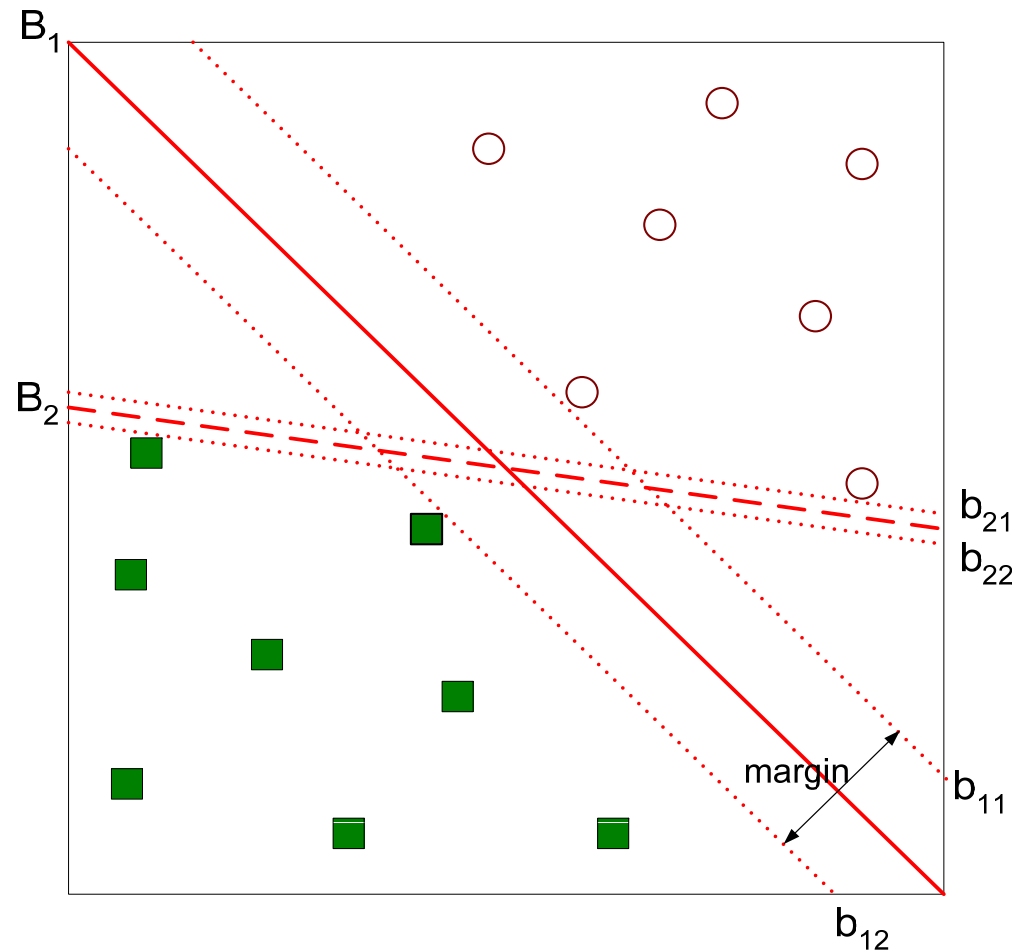
- Altre soluzioni

Support Vector Machines



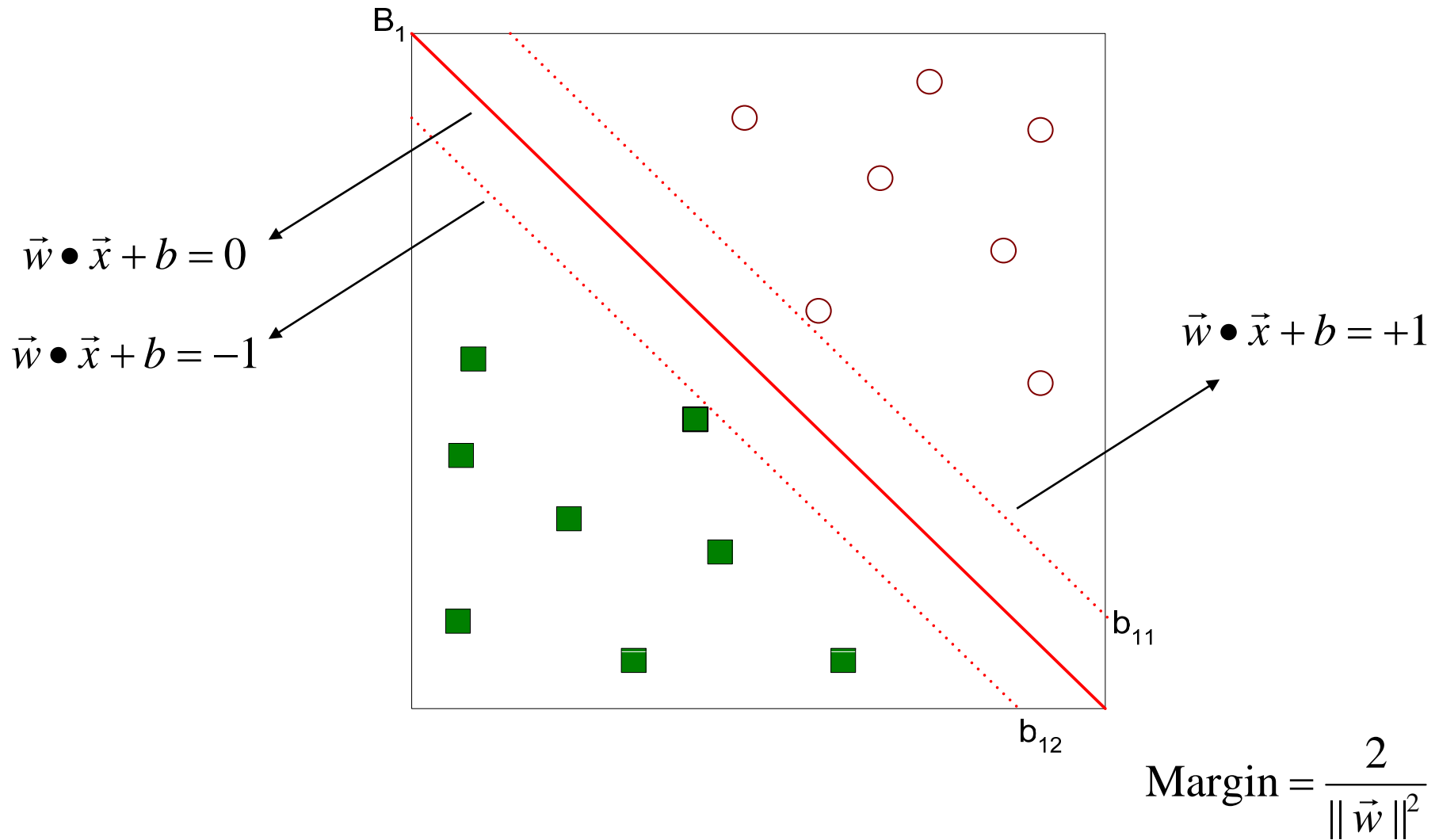
- Qual'è meglio? B_1 o B_2 ?
- Che significa "meglio"?

Support Vector Machines



- Trovare l'iperpiano che **massimizza** il margine => **B1** è meglio di **B2**

Support Vector Machines



Support Vector Machines

- **Si vuole massimizzare:** $\text{Margin} = \frac{2}{\|\vec{w}\|^2}$

- **Che è equivalente a minimizzare:** $L(w) = \frac{\|\vec{w}\|^2}{2}$

- **Sotto i vincoli:**

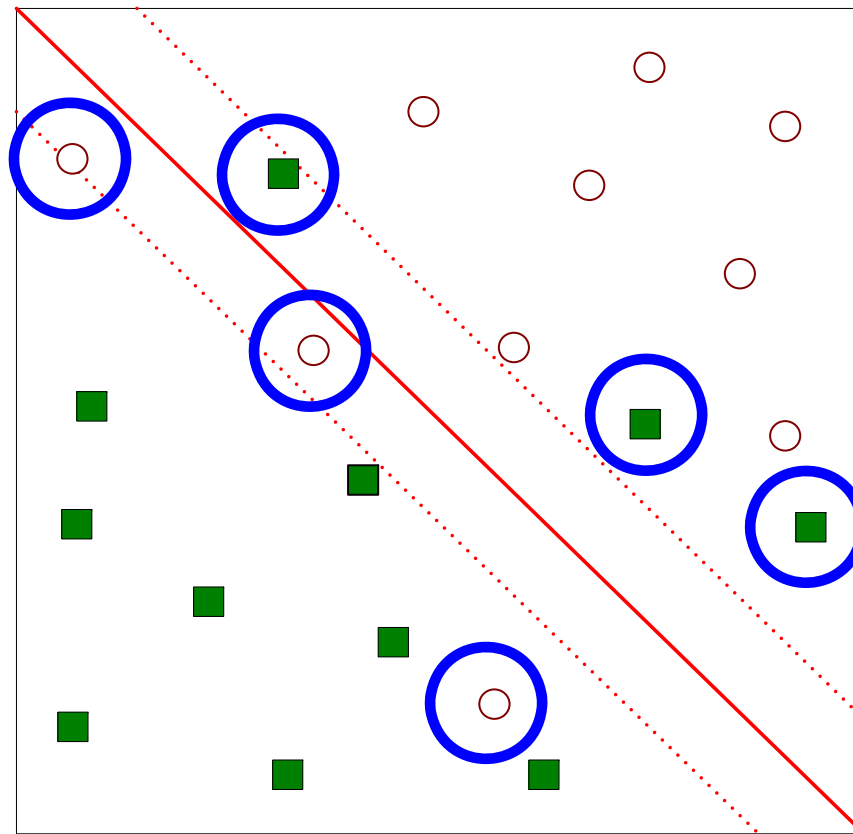
$$f(\vec{x}_i) = \begin{cases} 1 & \text{if } \vec{w} \bullet \vec{x}_i + b \geq 1 \\ -1 & \text{if } \vec{w} \bullet \vec{x}_i + b \leq -1 \end{cases}$$

- **Problema di ottimizzazione con vincoli**

- Approcci numerici alla soluzione (e.g., programmazione quadratica)

Support Vector Machines

- Che succede se il problema non è linearmente separabile?



Support Vector Machines

- Che succede se il problema non è linearmente separabile?

– Introduciamo variabili di comodo

– Si minimizza:

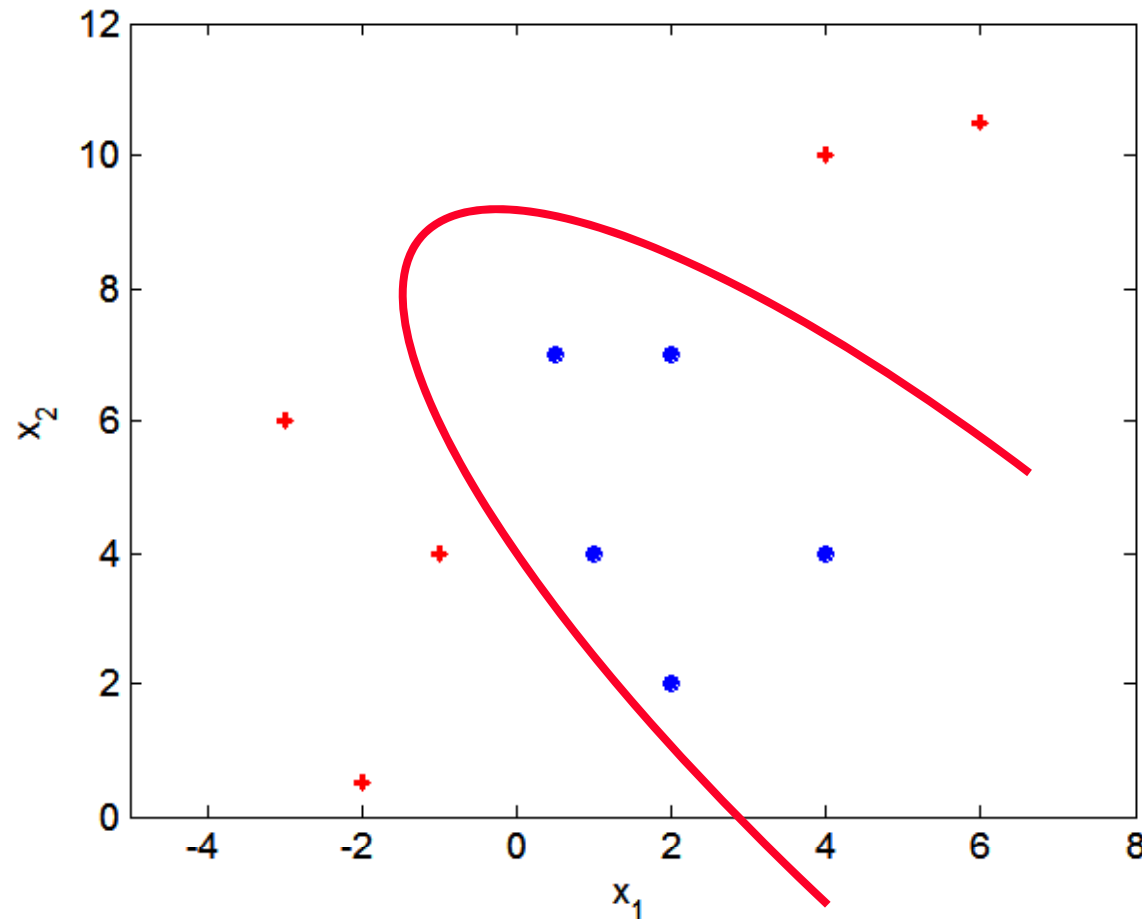
$$L(w) = \frac{\|\vec{w}\|^2}{2} + C \left(\sum_{i=1}^N \xi_i \right)$$

– Sotto il vincolo:

$$f(\vec{x}_i) = \begin{cases} 1 & \text{se } \vec{w} \bullet \vec{x}_i + b \geq 1 - \xi_i \\ -1 & \text{se } \vec{w} \bullet \vec{x}_i + b \leq -1 + \xi_i \end{cases}$$

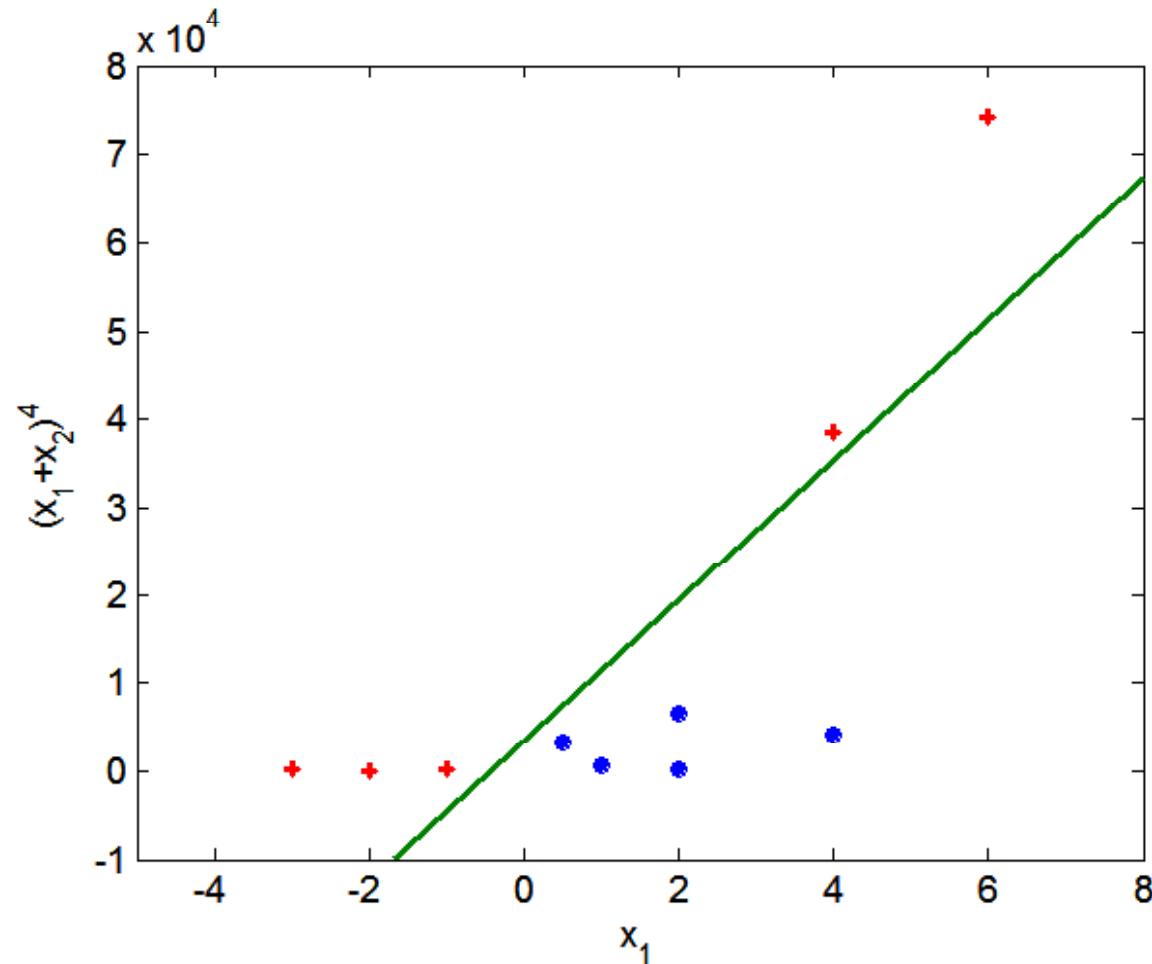
Nonlinear Support Vector Machines

- Che succede se il decision boundary non è lineare?



Nonlinear Support Vector Machines

- Si mappano i dati in uno spazio a più alte dimensioni



Concetti probabilistici, apprendimento probabilistico

- **Concetti probabilistici**

- Il concetto da apprendere è una funzione $c: X \rightarrow [0, 1]$
- $c(x)$, il valore target, denota la probabilità che l'etichetta 1 (*True*) sia assegnata a x
- Sostanzialmente, quello che abbiamo ottenuto finora

- **Probabilistic (i.e., Bayesian) Learning**

- Utilizzo di un criterio probabilistico nella selezione di un'ipotesi h
 - Esempio: l'ipotesi “più probabile” *considerando D*: MAP hypothesis
 - Esempio: h per cui D è “più probabile”: max likelihood (ML) hypothesis
- NB: h può essere una qualsiasi funzione

Alcune definizioni di base

- Spazio degli eventi (Ω): Dominio di una variabile casuale X
- Misura di probabilità $P(\bullet)$
 - P , è una misura su Ω
 - $P(X = x \in \Omega)$ è una misura della fiducia in $X = x$
- Assiomi di Kolmogorov
 - 1. $\forall x \in \Omega . 0 \leq P(X = x) \leq 1$
 - 2. $P(\Omega) \equiv \sum_{x \in \Omega} P(X = x) = 1$
 - 3. $\forall X_1, X_2, \dots \ni i \neq j \Rightarrow X_i \wedge X_j = \emptyset .$

$$P\left(\bigcup_{i=1}^{\infty} X_i\right) = \sum_{i=1}^{\infty} P(X_i)$$

- Probabilità congiunta: $P(X_1 \wedge X_2) \equiv$ dell'evento $X_1 \wedge X_2$
- indipendenza: $P(X_1 \wedge X_2) = P(X_1) \cdot P(X_2)$

Il teorema di Bayes

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

- $P(h)$ \equiv Probabilità a priori dell'ipotesi h
 - Misura la credenza iniziale indipendentemente da qualsiasi informazione (e quindi a priori)
- $P(D)$ \equiv Prior dell'insieme D
 - Misura la probabilità dell'insieme D (i.e., expresses D)
- $P(h | D)$ \equiv Probabilità di h dato D
 - | denota condizionamento - $P(h | D)$ is probabilità condizionale (a posteriori)
- $P(D | h)$ \equiv Probabilità di D dato h
 - Probabilità di osservare D sapendo che vale h (modello "generativo")
- $P(h \wedge D)$ \equiv Probabilità congiunta di h e D

Da Bayes "Essay towards solving a problem in the doctrine of chances" (1763)

Thomas Bayes

Born: 1702 in London, England

Died: 1761 in Tunbridge Wells, Kent, England



La scelta delle ipotesi

- **Teorema di Bayes**

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

- **Ipotesi MAP**

- Si vuole l'ipotesi più probabile sullo specifico training set
- $\arg \max_{x \in \Omega} [f(x)]$ \equiv il valore di x nello spazio Ω che esibisce il più alto $f(x)$
- Maximum a posteriori hypothesis, h_{MAP}

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

- **Ipotesi ML**

- Assumiamo $p(h_i) = p(h_j)$ (tutte le ipotesi sono egualmente probabili)
- Si sceglie l'ipotesi che spiega meglio i dati, h_{ML}

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

Esempio: Moneta bilanciata [1]

- **Lancio della moneta**

- Spazio: $\Omega = \{Head, Tail\}$
- Scenario: la moneta è bilanciata o sbilanciata al 60% in favore di *Head*
 - $h_1 \equiv$ bilanciata: $P(Head) = 0.5$
 - $h_2 \equiv$ 60% bias : $P(Head) = 0.6$
- Obiettivo: decidere tra l'ipotesi di default (null) e l'alternativa

- **Distribuzione a Priori**

- $P(h_1) = 0.75, P(h_2) = 0.25$
- Riflette le credenze iniziali su H
- L'apprendimento è revisione delle credenze

- **Evidenze**

- $d \equiv$ singolo lancio, viene *Head*
- D: Cosa crediamo adesso?
- R: Calcoliamo $P(d) = P(d | h_1) P(h_1) + P(d | h_2) P(h_2)$

Esempio: Moneta bilanciata [2]

- **Inferenza Bayesiana: Calcoliamo $P(d) = P(d | h_1) P(h_1) + P(d | h_2) P(h_2)$**
 - $P(\text{Head}) = 0.5 \cdot 0.75 + 0.6 \cdot 0.25 = 0.375 + 0.15 = 0.525$
 - Questa è la probabilità dell'osservazione $d = \text{Head}$
- **Apprendimento bayesiano**
 - In base al teorema di Bayes
 - $P(h_1 | d) = P(d | h_1) P(h_1) / P(d) = 0.375 / 0.525 = 0.714$
 - $P(h_2 | d) = P(d | h_2) P(h_2) / P(d) = 0.15 / 0.525 = 0.286$
 - *Le credenze sono state spostate verso h_1*
 - MAP: crediamo ancora che la moneta sia bilanciata
 - Approccio ML (assumiamo priors identici)
 - Le credenze sono revisionate a partire da 0.5
 - C'è più sbilanciamento a favore di h_1
- **Ulteriore evidenza: Sequenza D di 100 lanci con 70 heads e 30 tails**
 - $P(D) = (0.5)^{70} \cdot (0.5)^{30} \cdot 0.75 + (0.6)^{70} \cdot (0.4)^{30} \cdot 0.25$
 - Ora $P(h_1 | D) \ll P(h_2 | D)$

Stima di densità

- **Obiettivo principale: stimare $P(D | h)$**
 - Grazie al teorema di Bayes, possiamo ottenere la probabilità “a posteriori”
- **Tre approcci**
 - **Metodi parametrici**
 - Si assume una forma funzionale per le densità
 - Si stimano i parametri di tali forme funzionali
 - **Metodi nonparametrici**
 - La forma funzionale è determinata dai dati
 - **Mixture models**
 - Combinazioni di molte possibili forme funzionali
 - Neural Networks

Esempio: Maximum Likelihood Estimation

- **Dati M parametri**
- **Si trovino i valori più probabili**
 - **Massimizzando la probabilità congiunta**
 - **Assumendo N istanze indipendenti**

$$\boldsymbol{\theta} = [\theta_1, \dots, \theta_M]$$

$$L(\mathbf{x}_1, \dots, \mathbf{x}_N) = \prod_{i=1}^N p(\mathbf{x}_i | \boldsymbol{\theta})$$

- **Minimizzando l'errore corrispondente**

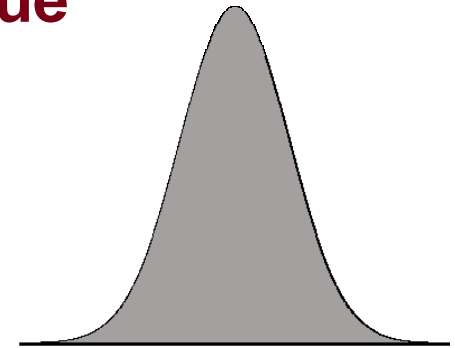
$$E = -\ln L(\mathbf{x}_1, \dots, \mathbf{x}_N) = -\sum_{i=1}^N \ln p(\mathbf{x}_i | \boldsymbol{\theta})$$

Il caso di dati gaussiani

- Assunzione tipica: gli attributi numerici hanno una distribuzione normale (*Gaussiana*)
- La densità di probabilità è definita da due parametri

– densità $f(x)$:

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



Dati gaussiani e MLE

- **Stimiamo la verosimiglianza**

$$L(\mu, \sigma^2) = \prod_{i=1}^n \frac{1}{\sigma \sqrt{2\pi}} e^{-(x_i - \mu)^2 / (2\sigma^2)} = \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-(1/2\sigma^2) \sum_{i=1}^n (x_i - \mu)^2}$$

- Al logaritmo:

$$\ln L(\mu, \sigma^2) = -\frac{n}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2$$

- Derivando:

$$\frac{\partial \ln L(\mu, \sigma^2)}{\partial \mu} = \frac{1}{\sigma^2} \sum_{i=1}^n (x_i - \mu) = 0$$

$$\frac{\partial \ln L(\mu, \sigma^2)}{\partial (\sigma^2)} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^n (x_i - \mu)^2 = 0$$

- Otteniamo

$$\hat{\mu} = \bar{X} \quad \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2$$

La più probabile classificazione delle nuove istanza

- **MAP and MLE: Limitazioni**
 - Il problema: “trovare l’ipotesi più probabile”
 - Ci basta la migliore classificazione di x , dato D
- **Metodi**
 - Troviamo la migliore (MAP) h , la usiamo per classificare
 - *Non è detto che la classificazione sia ottimale*
 - *Esempio: l’albero con l’errore minimo può sbagliare la classificazione su un’istanza specifica*
- **Obiettivo (raffinato)**
 - Vogliamo determinare la più probabile classificazione
 - Combiniamo la predizione di tutte le ipotesi
 - Le predizioni sono pesati dalle probabilità condizionali
 - Result: Bayes Optimal Classifier

Classificazione Bayesiana

- **Struttura**

- Si trovi la più probabile classificazione
- $f: X \rightarrow V$ (dominio \equiv spazio delle istanze, codominio \equiv insieme finito di valori)
- $x \in X$ espresso come collezione di attributi $x \equiv (x_1, x_2, \dots, x_n)$
- classificatore Bayesiano
 - Dato x_i
 - Si determini: il più probabile valore $v_j \in V$

$$\begin{aligned} v_{MAP} &= \arg \max_{v_j \in V} P(v_j | x) = \arg \max_{v_j \in V} P(v_j | x_1, x_2, \dots, x_n) \\ &= \arg \max_{v_j \in V} P(x_1, x_2, \dots, x_n | v_j) P(v_j) \end{aligned}$$

- **Problematiche**

- Stimare $P(v_j)$ è semplice: per ogni valore v_j , contiamo la sua frequenza in $D = \{ \langle x, f(x) \rangle \}$
- Tuttavia, è problematica la stima di $P(x_1, x_2, \dots, x_n | v_j)$ senza assunzioni a priori

Classificazione Bayesiana (con ipotesi)

- **Idea**

- $h_{MAP}(x)$ non fornisce necessariamente la classificazione più probabile

- **Esempio**

- Tre ipotesi: $P(h_1 | D) = 0.4$, $P(h_2 | D) = 0.3$, $P(h_3 | D) = 0.3$
- Su una nuova istanza x , $h_1(x) = +$, $h_2(x) = -$, $h_3(x) = -$
- Qual'è la migliore classificazione per x ?

- **Bayes Optimal Classification (BOC)**

$$v^* = v_{BOC} = \arg \max_{v_j \in V} \sum_{h_i \in H} [P(v_j | h_i) \cdot P(h_i | D)]$$

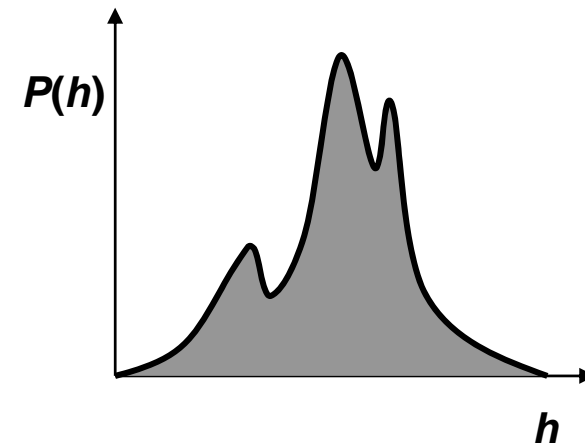
- **Esempio**

- $P(h_1 | D) = 0.4$, $P(- | h_1) = 0$, $P(+ | h_1) = 1$
- $P(h_2 | D) = 0.3$, $P(- | h_2) = 1$, $P(+ | h_2) = 0$
- $P(h_3 | D) = 0.3$, $P(- | h_3) = 1$, $P(+ | h_3) = 0$

- $\sum_{h_i \in H} [P(+ | h_i) \cdot P(h_i | D)] = 0.4$

$$\sum_{h_i \in H} [P(- | h_i) \cdot P(h_i | D)] = 0.6$$

- **Risultato:** $v^* = v_{BOC} = \arg \max_{v_j \in V} \sum_{h_i \in H} [P(v_j | h_i) \cdot P(h_i | D)] = -$



Naïve Bayes Classifier

- **Classificatore MAP**
$$\mathbf{v}_{MAP} = \arg \max_{v_j \in V} P(v_j | \mathbf{x}) = \arg \max_{v_j \in V} P(v_j | x_1, x_2, \dots, x_n)$$
$$= \arg \max_{v_j \in V} P(x_1, x_2, \dots, x_n | v_j) P(v_j)$$
- **Naive Bayes**
 - Uno dei metodi più pratici di apprendimento
 - Assunzione di base: gli attributi di x sono indipendenti
- **Quando si può usare**
 - Il training set è grande
 - Gli attributi che descrivono x sono (sostanzialmente) indipendenti
- **Applicazione più di successo**
 - Classificazione di testi
- **Assunzione Naïve**
$$P(x_1, x_2, \dots, x_n | v_j) = \prod_i P(x_i | v_j)$$
- **Classificatore (Naïve) Bayes**
$$\mathbf{v}_{NB} = \arg \max_{v_j \in V} P(v_j) \prod_i P(x_i | v_j)$$

Probabilità per PlayTennis

Outlook			Temperature			Humidity			Wind			Play	
Yes	No		Yes	No		Yes	No		Yes	No	Yes	No	
Sunny	2	3	Hot	2	2	High	3	4	Light	6	2	9	5
Overcast	4	0	Mild	4	2	Normal	6	1	Strong	3	3		
Rainy	3	2	Cool	3	1								
Sunny	2/9	3/5	Hot	2/9	2/5	High	3/9	4/5	Light	6/9	2/5	9/14	5/14
Overcast	4/9	0/5	Mild	4/9	2/5	Normal	6/9	1/5	Strong	3/9	3/5		
Rainy	3/9	2/5	Cool	3/9	1/5								

Outlook	Temp	Humidity	Wind	Play
Sunny	Hot	High	Light	No
Sunny	Hot	High	Strong	No
Overcast	Hot	High	Light	Yes
Rainy	Mild	High	Light	Yes
Rainy	Cool	Normal	Light	Yes
Rainy	Cool	Normal	Strong	No
Overcast	Cool	Normal	Strong	Yes
Sunny	Mild	High	Light	No
Sunny	Cool	Normal	Light	Yes
Rainy	Mild	Normal	Light	Yes
Sunny	Mild	Normal	Strong	Yes
Overcast	Mild	High	Strong	Yes
Overcast	Hot	Normal	Light	Yes
Rainy	Mild	High	Strong	No

Probabilità per PlayTennis

Outlook			Temperature			Humidity			Windy			Play	
Yes	No		Yes	No		Yes	No		Yes	No	Yes	No	
Sunny	2	3	Hot	2	2	High	3	4	Light	6	2	9	5
Overcast	4	0	Mild	4	2	Normal	6	1	Strong	3	3		
Rainy	3	2	Cool	3	1								
Sunny	2/9	3/5	Hot	2/9	2/5	High	3/9	4/5	Light	6/9	2/5	9/14	5/14
Overcast	4/9	0/5	Mild	4/9	2/5	Normal	6/9	1/5	Strong	3/9	3/5		
Rainy	3/9	2/5	Cool	3/9	1/5								

- Una nuova istanza:

Outlook	Temp.	Humidity	Windy	Play
Sunny	Cool	High	Strong	?

verosimiglianza delle due classi

$$\text{Per "yes"} = 2/9 \times 3/9 \times 3/9 \times 3/9 \times 9/14 = 0.0053$$

$$\text{Per "no"} = 3/5 \times 1/5 \times 4/5 \times 3/5 \times 5/14 = 0.0206$$

Probabilità associata:

$$P(\text{"yes"}) = 0.0053 / (0.0053 + 0.0206) = 0.205$$

$$P(\text{"no"}) = 0.0206 / (0.0053 + 0.0206) = 0.795$$

In pratica...

Outlook	Temp.	Humidity	Wind	Play
Sunny	Cool	High	Strong	?

← *Evidenza E*

↖
*Probabilità della
classe “yes”*

$$\begin{aligned}\Pr(\text{yes} | E) &= \Pr(\text{Outlook} = \text{Sunny} | \text{yes}) \\ &\quad \times \Pr(\text{Temperature} = \text{Cool} | \text{yes}) \\ &\quad \times \Pr(\text{Humidity} = \text{High} | \text{yes}) \\ &\quad \times \Pr(\text{Windy} = \text{True} | \text{yes}) \\ &\quad \times \frac{\Pr(\text{yes})}{\Pr(E)} \\ &= \frac{\frac{2}{9} \times \frac{3}{9} \times \frac{3}{9} \times \frac{3}{9} \times \frac{9}{14}}{\Pr(E)}\end{aligned}$$

Il problema della “frequenza-zero”

- Che succede se il valore di un attributo non compare con un valore di classe?

(esempio: “Humidity = high” per la classe “yes”)

- La probabilità è zero!

$$\Pr(\textit{Humidity} = \textit{High} \mid \textit{yes}) = 0$$

- La probabilità a posteriori diventa zero!

$$\Pr(\textit{yes} \mid E) = 0$$

- Rimedio: sommiamo 1 al conteggio di ogni combinazione attributo-classe (Laplace estimator)

Stime di probabilità

- Aggiungere una costante differente da 1 può risultare più appropriato
- Esempio su Outlook

$$\frac{2 + \mu/3}{9 + \mu}$$

Sunny

$$\frac{4 + \mu/3}{9 + \mu}$$

Overcast

$$\frac{3 + \mu/3}{9 + \mu}$$

Rainy

- I pesi non devono necessariamente essere uguali (ma la somma deve essere 1)

$$\frac{2 + \mu p_1}{9 + \mu}$$

$$\frac{4 + \mu p_2}{9 + \mu}$$

$$\frac{3 + \mu p_3}{9 + \mu}$$

Valori mancanti

- **Nel training: l'istanza non viene conteggiata nella combinazione attributo-valore**
- **Nella classificazione: l'attributo viene omesso dal calcolo**
- **esempio:**

Outlook	Temp.	Humidity	Wind	Play
?	Cool	High	Strong	?

verosimiglianza di "yes" = $3/9 \times 3/9 \times 3/9 \times 9/14 = 0.0238$

verosimiglianza di "no" = $1/5 \times 4/5 \times 3/5 \times 5/14 = 0.0343$

$P(\text{"yes"}) = 0.0238 / (0.0238 + 0.0343) = 41\%$

$P(\text{"no"}) = 0.0343 / (0.0238 + 0.0343) = 59\%$

Attributi numerici

- Assunzione tipica: gli attributi hanno una distribuzione normale (*Gaussiana*)
- La densità di probabilità è definita da due parametri
 - *Valor medio sul campione μ*

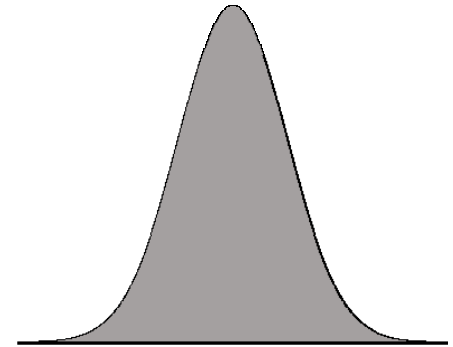
$$\mu = \frac{1}{n} \sum_{i=1}^n x_i$$

- *Devianza sul campione σ*

$$\sigma = \frac{1}{n-1} \sum_{i=1}^n (x_i - \mu)^2$$

- *densità $f(x)$:*

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



Ancora Playtennis

Outlook	Temperature		Humidity		Wind		Play				
	Yes	No	Yes	No	Yes	No	Yes	No			
Sunny	2	3	64, 68,	65, 71,	65, 70,	70, 85,	Light	6	2	9	5
Overcast	4	0	69, 70,	72, 80,	70, 75,	90, 91,	Strong	3	3		
Rainy	3	2	72, ...	85, ...	80, ...	95, ...					
Sunny	2/9	3/5	$\mu = 73$	$\mu = 75$	$\mu = 79$	$\mu = 86$	Light	6/9	2/5	9/14	5/14
Overcast	4/9	0/5	$\sigma = 6.2$	$\sigma = 7.9$	$\sigma = 10.2$	$\sigma = 9.7$	Strong	3/9	3/5		
Rainy	3/9	2/5									

- Valore di densità

$$f(\text{temperature} = 66 \mid \text{yes}) = \frac{1}{\sqrt{2\pi}6.2} e^{-\frac{(66-73)^2}{2*6.2^2}} = 0.0340$$

Classificazione su valori numerici

Outlook	Temp.	Humidity	Windy	Play
Sunny	66	90	true	?

Verosimiglianza di "yes" = $2/9 \times 0.0340 \times 0.0221 \times 3/9 \times 9/14 = 0.000036$

Verosimiglianza di "no" = $3/5 \times 0.0291 \times 0.0380 \times 3/5 \times 5/14 = 0.000136$

$P(\text{"yes"}) = 0.000036 / (0.000036 + 0.000136) = 20.9\%$

$P(\text{"no"}) = 0.000136 / (0.000036 + 0.000136) = 79.1\%$

Riassumendo...

- **Il Naïve Bayes funziona sorprendentemente bene (anche se l'assunzione di indipendenza è chiaramente violata)**
 - **Non servono stime accurate di probabilità finché la probabilità massima è assegnata alla classe corretta**
- **Tuttavia: troppi attributi ridondanti può causare problemi**
- **Inoltre: molti attributi numerici non seguono la distribuzione normale**
- **Miglioramenti:**
 - **Selezioniamo i migliori attributi**
 - **Reti Bayesiane**

Applicazioni: Spam?

From: "" <takworld@hotmail.com>

Subject: real estate is the only way... gem oalvgkay

Anyone can buy real estate with no money down

Stop paying rent TODAY !

There is no need to spend hundreds or even thousands for similar courses

I am 22 years old and I have already purchased 6 properties using the methods outlined in this truly INCREDIBLE ebook.

Change your life NOW !

=====
Click Below to order:

<http://www.wholesaledaily.com/sales/nmd.htm>
=====

Bayesian Learning e Spam

- **Msg1**
 - From a@hotmail.com
 - Eccezionale! Con l'acquisto di soli 3 coupons, potrai progettare un viaggio alle Maldive a costo 0!
- **Msg2**
 - From b@msnmail.de
 - Un'offerta d'eccezione! MS Office al prezzo di soli 49\$...
- **Msg3**
 - From c@email.it
 - Affrettati! L'offerta vale eccezionalmente per un periodo limitato! E' in corso una promozione per l'acquisto di ...
- **Msg4**
 - From instructor@univ.edu
 - L'analisi delle eccezioni è una tecnica molto utilizzata nell'analisi delle frodi
- **Msg5**
 - From folino@univ.edu
 - Le delucidazioni relative al progetto si forniscono esclusivamente il venerdì pomeriggio.
- **Msg6**
 - From knowledgeMiner@usa.com
 - Le offriamo il software di data mining più utile per i suoi progetti. Con l'acquisto del software, avrà a disposizione un corso per apprendere l'utilizzo
- **Msg7**
 - From student@univ.edu
 - Ti andrebbe di fare il progetto per il corso di data mining insieme a me? L'offerta formativa è troppo bassa per un impegno singolo...

Incidenza messaggio-termine

	progetto	eccezione	Data mining	offerta	corso	acquisto
msg1	1	1	0	0	0	1
msg2	0	1	0	1	0	0
msg3	0	1	0	1	1	1
msg4	0	1	0	0	0	0
msg5	1	0	0	0	0	0
msg6	1	0	1	1	1	1
msg7	1	0	1	1	1	0

1 se msg contiene **word**,
0 altrimenti

Bayesian Learning

- **Forma normale**
 - Rimozione di plurali, forme verbali, ecc.
- **Rimozione di stopwords**
 - Stopwords: parole troppo comuni
- **Lemmatizzazione**
 - Ricostruzione della “radice” della parola
- **Ricostruzione della struttura tabellare**
 - Binaria
 - Numerica
 - Frequenze delle parole
- **Analisi delle tabelle di contingenza e definizione del classificatore bayesiano**