

L'APPRENDIMENTO AUTOMATICO

MACHINE LEARNING

Alessandro Sperduti

Introduzione



Principali paradigmi di apprendimento



Ingredienti fondamentali



Errore Empirico ed Errore Ideale



Alcuni Modelli di Apprendimento

Introduzione



Introduzione



Principali paradigmi di apprendimento



Ingredienti fondamentali



Errore Empirico ed Errore Ideale





Principali paradigmi di apprendimento



Ingredienti fondamentali



Errore Empirico ed Errore I



Principali paradigmi di apprendimento



Ingredienti fondamentali



Errore Empirico ed Errore Ideale



Alcuni Modelli di Apprendimento

0



Errore Empirico ed Errore Ideale



Alcuni Modelli di Apprendimento

Quando è necessario l'Apprendimento Automatico ?

quando il sistema deve

- **adattarsi** all'ambiente in cui opera (anche personalizzazione automatica)
- **migliorare** le sue prestazioni rispetto ad un particolare compito
- **scoprire** regolarità e nuova informazione (conoscenza) a partire da dati empirici
- **acquisire** nuove capacità computazionali

Quando è necessario l'Apprendimento Automatico

quando il sistema deve

- **adattarsi** all'ambiente in cui opera (anche personalizzazione automatica)
- **migliorare** le sue prestazioni rispetto a un sistema tradizionale
- **scoprire** regolarità e nuove informazioni dai dati empirici

Quando è necessario l'Apprendimento Automatico ?

quando il sistema deve

- **adattarsi** all'ambiente in cui opera
(anche personalizzazione automatica)
- **migliorare** le sue prestazioni rispetto ad un particolare compito
- **scoprire** regolarità e nuova informazione (conoscenza) a partire da dati empirici
- **acquisire** nuove capacità computazionali

Quando è necessario l'Apprendimento Automatico ?

quando il sistema deve

- **adattarsi** all'ambiente in cui opera
(anche personalizzazione automatica)
- **migliorare** le sue prestazioni rispetto ad un particolare compito
- **scoprire** regolarità e nuova informazione (conoscenza) a partire da dati empirici
- **acquisire** nuove capacità computazionali

- **migliorare** le sue prestazioni rispetto ad
- **scoprire** regolarità e nuova informazione dati empirici
- **acquisire** nuove capacità computazionali

Perché non usare un approccio algoritmico tradizionale ?

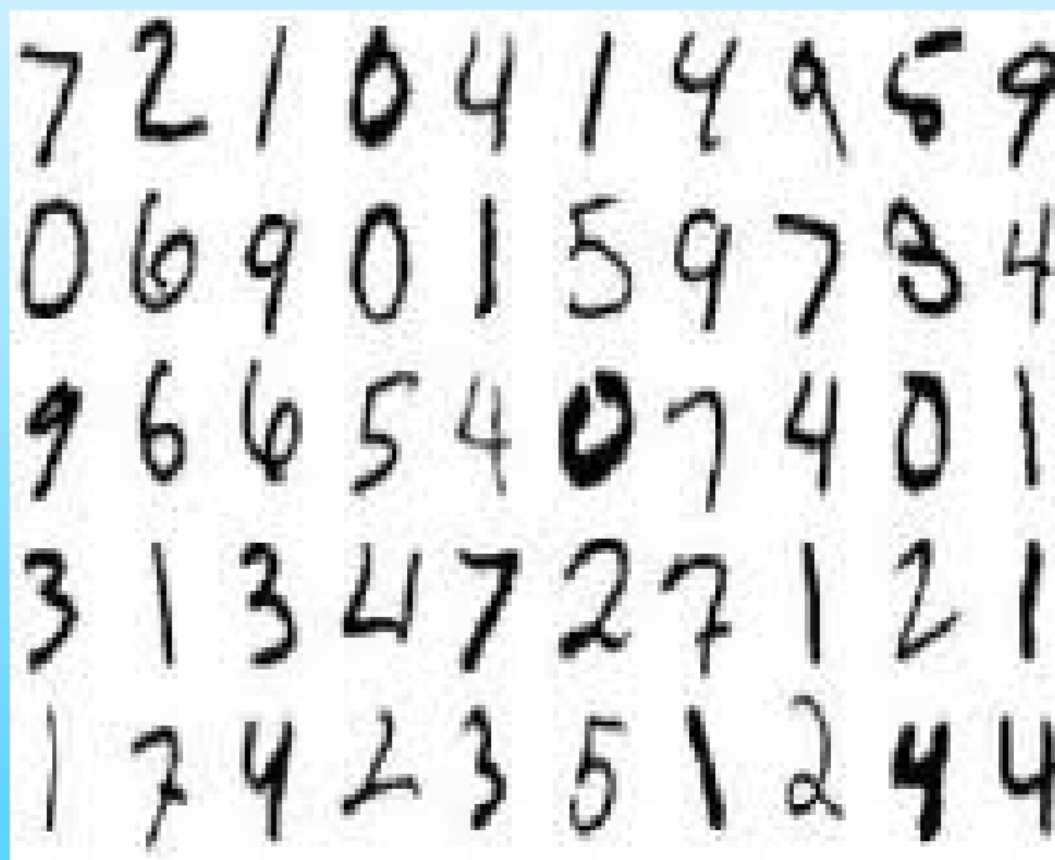
- impossibile formalizzare esattamente il problema (e quindi dare una soluzione algoritmica)
- presenza di rumore e/o incertezza
- complessità alta nel formulare una soluzione: impossibile "fare a mano"
- mancanza di conoscenza "compilata" rispetto al problema da risolvere

Rule 1: *If* Income (w_j) is *Very High* (VH) *then* the sensitivity to price (quality) parameter s_{ij} (s_{ij}) is *Very Insensitive* (VI) (*Very Sensitive* (VS)).

Perché non usare un approccio algoritmico tradizionale ?

- impossibile formalizzare esattamente il problema (e quindi dare una soluzione algoritmica)
- presenza di rumore e/o incertezza
- complessità alta nel formulare una soluzione: impossibile "fare a mano"
- mancanza di conoscenza "compilata" rispetto al problema da risolvere

Che cifre sono rappresentate in questa immagine?



7	2	1	0	4	1	4	9	5	9
0	6	9	0	1	5	9	7	8	4
9	6	6	5	4	0	7	4	0	1
3	1	3	4	7	2	7	1	2	1
1	7	4	2	3	5	1	2	4	4

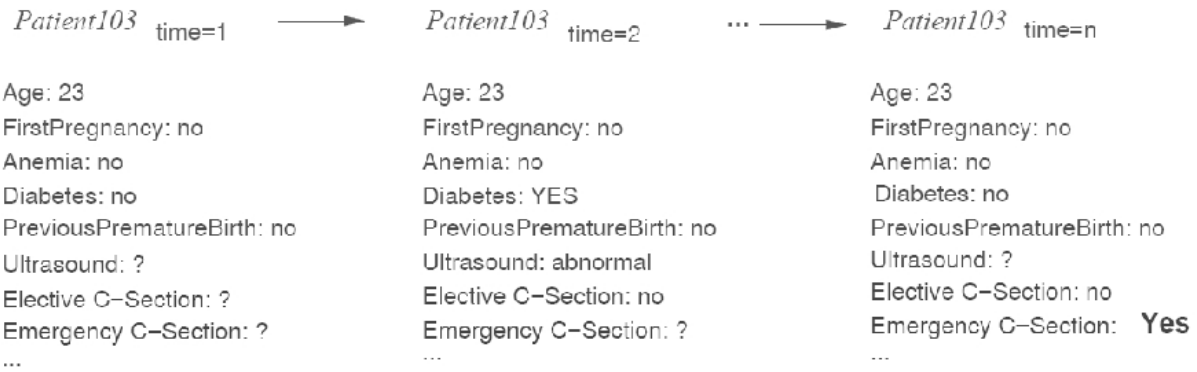
Perché non usare un approccio algoritmico

- impossibile formalizzare esattamente il problema e trovare una soluzione algoritmica
- presenza di rumore e/o incertezza
- complessità alta nel formulare una soluzione "a mano"
- mancanza di conoscenza "compilata" rispetto al problema da risolvere



Perché non usare un approccio algoritmico tradizionale ?

- impossibile **formalizzare** esattamente il problema (e quindi dare una soluzione algoritmica)
- presenza di **rumore** e/o **incertezza**
- **complessità alta** nel formulare una soluzione: impossibile "fare a mano"
- mancanza di **conoscenza "compilata"** rispetto al problema da risolvere



- impossibile formalizzare esattamente il problema (e quindi da una soluzione algoritmica)
- presenza di rumore e/o incertezza
- complessità alta nel formulare una soluzione: impossibile "fare a mano"
- mancanza di conoscenza "compilata" rispetto al problema da risolvere

e 1: If Income (w_j) is Very High (VH) then the sensitivity to price (quality) parameter s_{ij} (s_{ij}) is Very Insensitive (VI) (Very Sensitive (VS)).

Rule 1:	<i>If</i> Income (w_i) is <i>Very High</i> (VH) <i>then</i> the sensitivity to price (quality) parameter s_{ij} (s_{ij}) is <i>Very Insensitive</i> (VI) (<i>Very Sensitive</i> (VS)).
Rule 2:	<i>If</i> Income (w_i) is <i>High</i> (HG) <i>then</i> the sensitivity to price (quality) parameter s_{ij} (s_{ij}) is <i>Insensitive</i> (IS) (<i>Sensitive</i> (ST)).
Rule 3:	<i>If</i> Income (w_i) is <i>Medium</i> (MD) <i>then</i> the sensitivity to price (quality) parameter s_{ij} (s_{ij}) is <i>Medium Sensitive</i> (MS).
Rule 4:	<i>If</i> Income (w_i) is <i>Low</i> (LW) <i>then</i> the sensitivity to price (quality) parameter s_{ij} (s_{ij}) is <i>Sensitive</i> (<i>Insensitive</i> (IS)).
Rule 5:	<i>If</i> Income (w_i) is <i>Very Low</i> (VL) <i>then</i> the sensitivity to price (quality) parameter s_{ij} (s_{ij}) is <i>Very Sensitive</i> (<i>Very Insensitive</i> (VS)).

Patient10

Age: 23

FirstPregna

Anemia: no

Diabetes: n

PreviousPre

Ultrasound:

Elective C-

Emergency

...

Ruolo dei dati

Tipicamente

- si hanno a disposizione (molti?) **dati**
 - ottenuti una volta per tutte **batch learning**
 - acquisibili interagendo direttamente con l'ambiente **on-line learning**
- (forse) **conoscenza** del dominio applicativo, ma
 - incompleta
 - imprecisa (rumore, ambiguità, incertezza, errori, ...)

Desiderio: usare i dati per

- ottenere **nuova conoscenza**
- **raffinare** la conoscenza di cui si dispone
- **correggere** la conoscenza di cui si dispone

rse) **conoscenza** del dominio applicativo, ma
completa

imprecisa (rumore, ambiguità, incertezza, errori, ...)

Desiderio: usare i dati per

- ottenere **nuova conoscenza**
- **raffinare** la conoscenza di cui si dispone
- **correggere** la conoscenza di cui si dispone

ised Learning

di auto-supervisione)

Introduzione



Principali paradigmi di apprendimento



Ingredienti fondamentali



Errore Empirico ed Errore Ideale



Principali Paradigmi di Apprendimento

Un paradigma specifica

- quali dati sono disponibili per l'apprendimento
- con che modalità il sistema riceve i dati
- che tipo di output ci si aspetta dal sistema

Supervised Learning

Unsupervised Learning

Reinforcement Learning

Supervised Learning

Unsupervised Learning

Reinforcement Learning

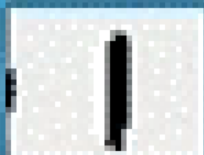
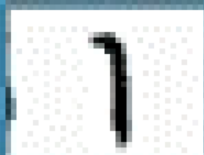
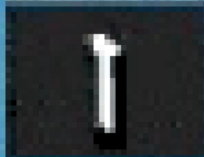
Supervised Learning (apprendimento supervisionato)

$\{(x_i, f(x_i))\}$

- dato un insieme di dati preclassificati (esempi di apprendimento) apprendere una descrizione generale che incapsula l'informazione contenuta negli esempi
- tale descrizione deve poter essere usata in modo predittivo dato un nuovo ingresso \tilde{x} , predire $f(\tilde{x})$
- si assume che un esperto (o maestro) ci fornisca la supervisione i valori $f(x_i)$



Face recognition technologies



AT&T

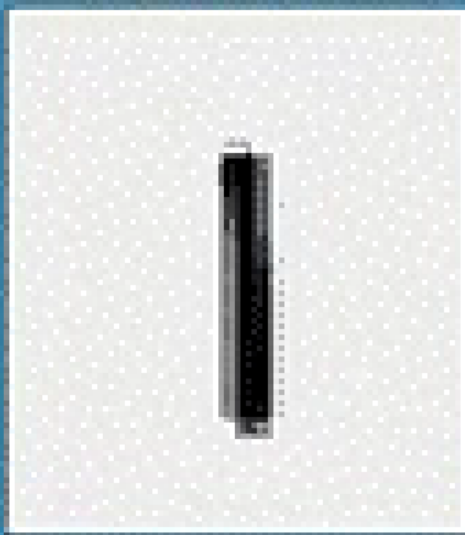
LeNet 5

RESEARCH

answer: 1

1

111

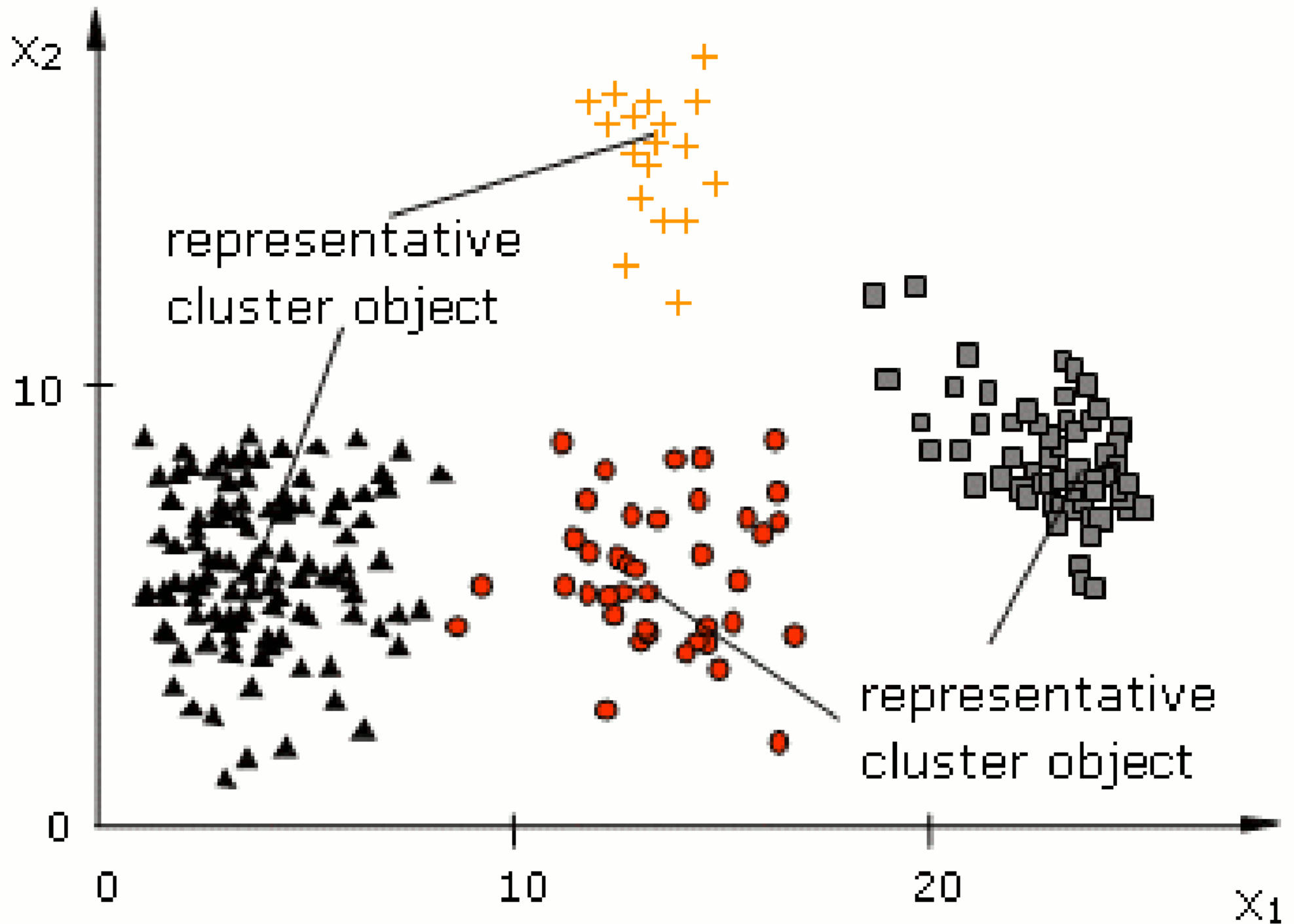


- si assume che un esperto (o maestro) ci fornisca supervisione i valori $f(x_i)$

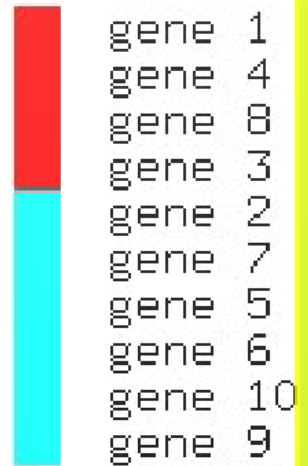
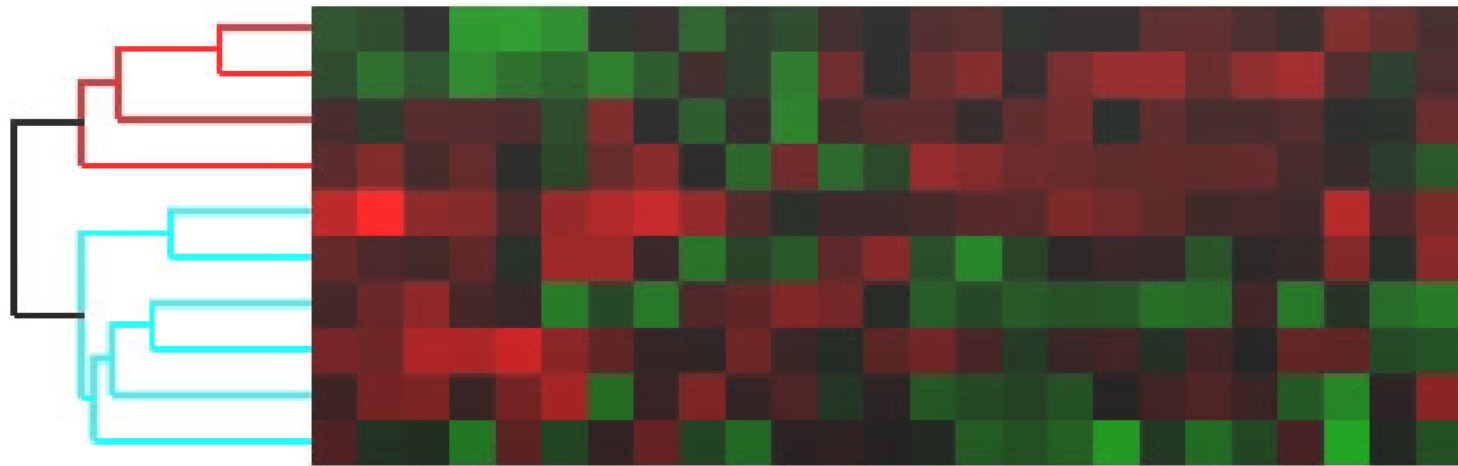
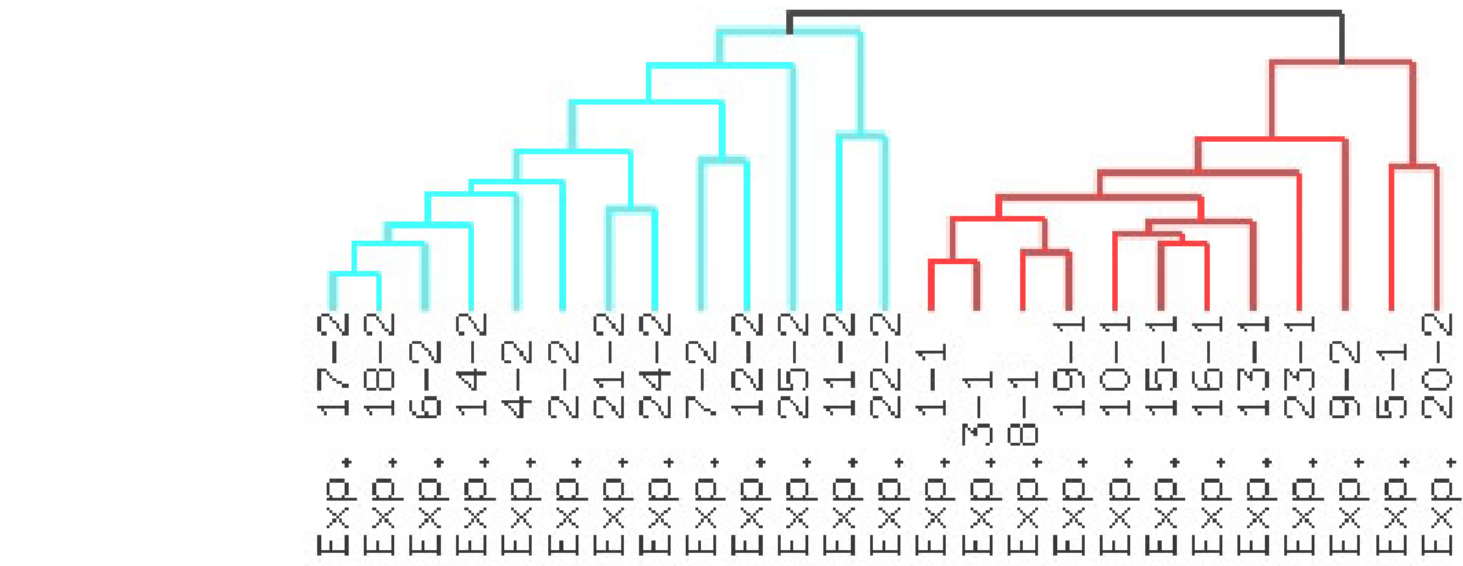
Unsupervised Learning (apprendimento non-supervisionato)

- dato un insieme di dati $\{x_i\}$, estrarre regolarità e/o pattern valide su tutto il dominio di ingresso
- non esiste un maestro che ci fornisca un aiuto

learning
(con rinforzo)



number of clusters: 4



Reinforcement Learning (apprendimento con rinforzo)

Sono dati



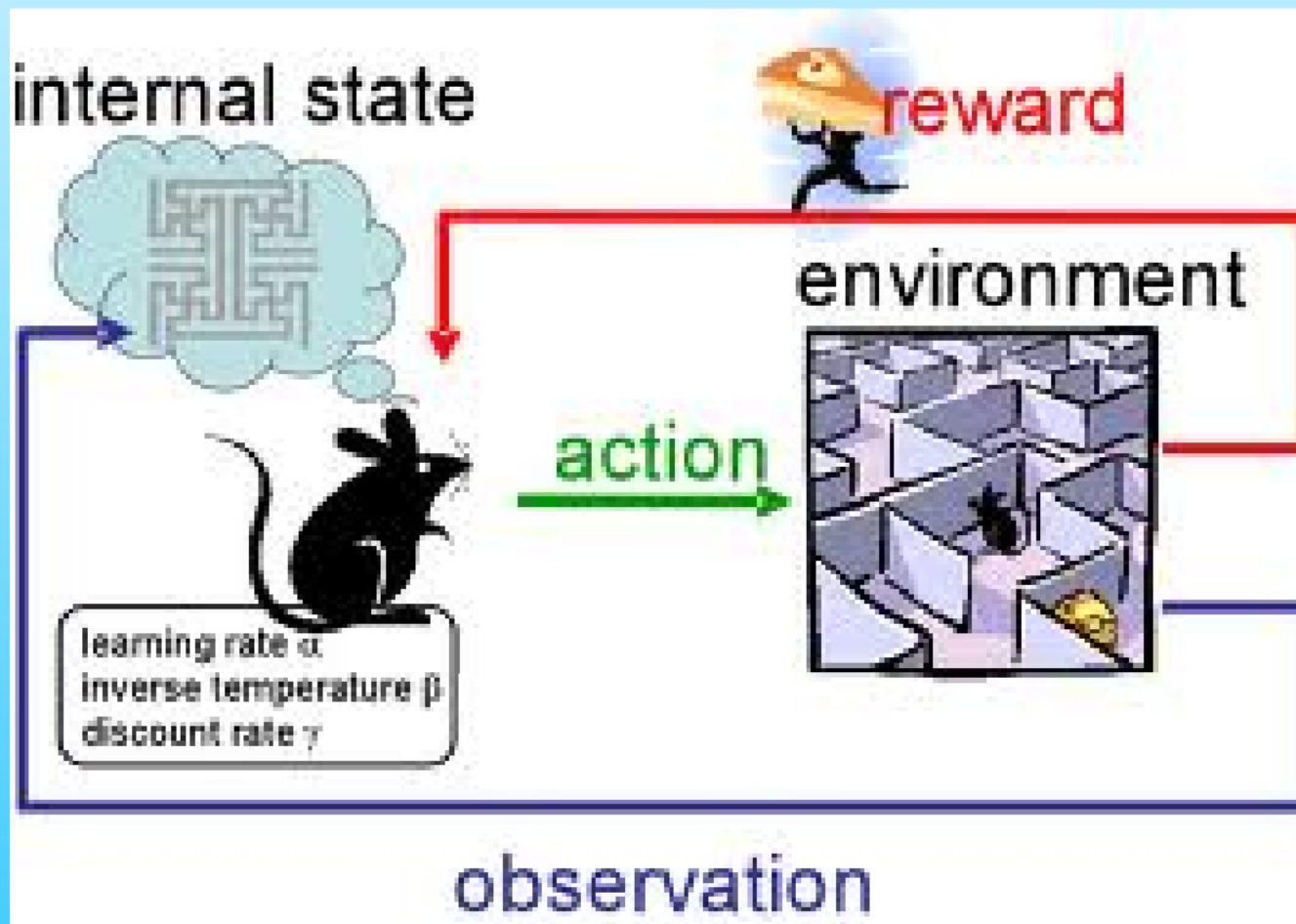
- un **agente** (intelligente?), che può
 - trovarsi in uno **stato s**
 - eseguire una **azione a** (all'interno delle azioni possibili nello stato corrente)
- un **ambiente** tale che quando l'agente applica una azione a nello stato s restituisce
 - lo **stato successivo**
 - una **ricompensa r** , che può essere positiva (+), negativa (-), o neutra (0)

Scopo dell'agente è quello di massimizzare la ricompensa

$$\sum_{t=0}^{\infty} \gamma^t r_{t+1} \text{ dove } 0 \leq \gamma < 1$$



scelta della strategia ottima





Principali paradigmi di apprendimento



Ingredienti fondamentali



Errore Empirico ed Errore I



Ingredienti Fondamentali Apprendimento

Dati

(spazio delle istanze)



Spazio delle Ipotesi H

- costituisce l'insieme delle funzioni che possono essere realizzate dal sistema di apprendimento
- si assume che la funzione da apprendere f possa essere rappresentata da una ipotesi

$h \in H$

(selezione di h attraverso i dati di apprendimento)

- o che almeno una ipotesi h sia simile a f
(approssimazione)



Algoritmo di ricerca
nello spazio delle Ipotesi
(algoritmo di apprendimento)



h "ottima"

ipotesi restituita
dall'apprendimento

Idati

pazio delle ista.

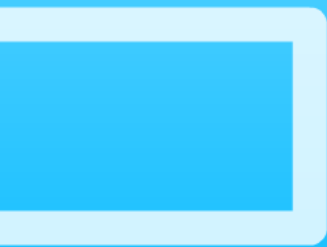
(selezione di h attr

- o che almeno un

(ap



Algoritmo di ricerca
nello spazio delle Ipotesi
(algoritmo di apprendimento)



Fondamentali Apprendimento

Spazio delle Ipotesi H

- costituisce l'insieme delle funzioni che possono essere realizzate dal sistema di apprendimento
- si assume che la funzione da apprendere f possa essere rappresentata da una ipotesi

$h \in H$

(selezione di h attraverso i dati di apprendimento)

- o che almeno una ipotesi h sia simile a f
(approssimazione)

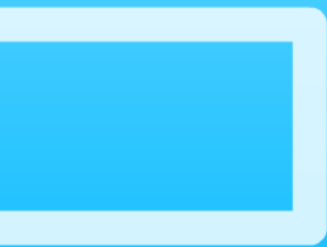
(selezione di h attr

- o che almeno un

(ap



Algoritmo di ricerca
nello spazio delle Ipotesi
(algoritmo di apprendimento)



L "ottima"

ipotesi restituita
dall'apprendimento

ATTENZIONE!!

azio delle ipotesi H no

ATTENZIONE!!

lo spazio delle ipotesi H non può coincidere con l'insieme di tutte le funzioni possibili e la ricerca essere esaustiva



essere esaustiva



Apprendimento Inutile !!

Bias Induttivo

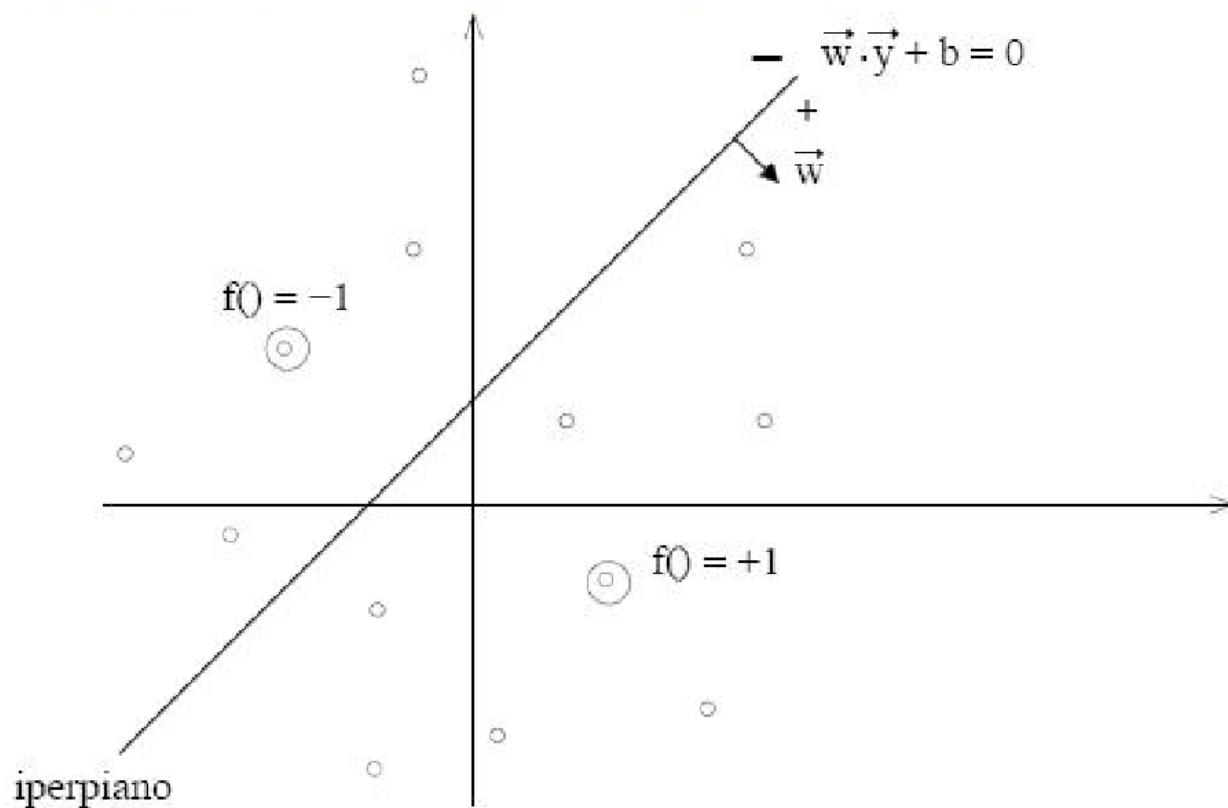
Bias Induttivo

- sulla rappresentazione (H)
- sulla ricerca (alg. apprendimento)

Spazio delle Ipotesi: Esempio 1

Iperpiani in \mathbb{R}^2

- Spazio delle Istanze \rightarrow punti nel piano: $X = \{\vec{y} \in \mathbb{R}^2\}$
- Spazio delle Ipotesi \rightarrow dicotomie indotte da iperpiani in \mathbb{R}^2 :
 $\mathcal{H} = \{f_{(\vec{w}, b)}(\vec{y}) \mid f_{(\vec{w}, b)}(\vec{y}) = \text{sign}(\vec{w} \cdot \vec{y} + b), \vec{w} \in \mathbb{R}^2, b \in \mathbb{R}\}$



Esempio 1

\mathbb{R}^2

in \mathbb{R}^2 :

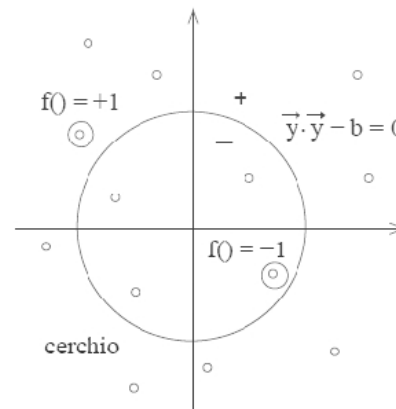
$\vec{w} \in \mathbb{R}^2, b \in \mathbb{R}$

$\vec{y} \cdot \vec{w} + b = 0$

Spazio delle Ipotesi: Esempio 2

Dischi in \mathbb{R}^2

- Spazio delle Istanze \rightarrow punti nel piano: $X = \{\vec{y} \in \mathbb{R}^2\}$
- Spazio delle Ipotesi \rightarrow dicotomie indotte da dischi in \mathbb{R}^2 centrati nell'origine:
 $\mathcal{H} = \{f_b(\vec{y}) \mid f_b(\vec{y}) = \text{sign}(\vec{y} \cdot \vec{y} - b), b \in \mathbb{R}\}$



Spazio

Congiunzione di m letterali po

- Spazio delle Istanze \rightarrow stringe
- Spazio delle Ipotesi \rightarrow tutte le stringhe l_1, \dots, l_m (l_1 è vero se il p...
contengono solo l'operatore

$\mathcal{H} = \{f_{\{i_1, \dots, i_j\}}(s) \mid f_{\{i_1, \dots, i_j\}}(s) = \dots\}$

Es. $m = 3, X = \{0, 1\}^3$

Esempi di istanze $\rightarrow s_1 = 101$

Esempi di ipotesi $\rightarrow h_1 \equiv l_2, h_2 \equiv l_3$

Notare che: h_1, h_2 , e h_5 sono t

h_4 è vera per s_1 e s_4 ma falsa p

Esempio 2

\mathbb{R}^2

\mathbb{R}^2 centrati nell'origine:

o

b=0

o

→

Spazio delle Ipotesi: Esempio 3

Congiunzione di m letterali positivi

- Spazio delle Istanze \rightarrow stringhe di m bit: $X = \{s | s \in \{0, 1\}^m\}$
- Spazio delle Ipotesi \rightarrow tutte le sentenze logiche che riguardano i letterali positivi l_1, \dots, l_m (l_1 è vero se il primo bit vale 1, l_2 è vero se il secondo bit vale 1, etc.) e che contengono solo l'operatore \wedge (**and**):

$$\mathcal{H} = \{f_{\{i_1, \dots, i_j\}}(s) | f_{\{i_1, \dots, i_j\}}(s) \equiv l_{i_1} \wedge l_{i_2} \wedge \dots \wedge l_{i_j}, \{i_1, \dots, i_j\} \subseteq \{1, \dots, m\}\}$$

Es. $m = 3, X = \{0, 1\}^3$

Esempi di istanze $\rightarrow s_1 = 101, s_2 = 001, s_3 = 100, s_4 = 111$

Esempi di ipotesi $\rightarrow h_1 \equiv l_2, h_2 \equiv l_1 \wedge l_2, h_3 \equiv \text{true}, h_4 \equiv l_1 \wedge l_3, h_5 \equiv l_1 \wedge l_2 \wedge l_3$

Notare che: h_1, h_2 , e h_5 sono false per s_1, s_2 e s_3 e vere per s_4 ; h_3 è vera per ogni istanza;

h_4 è vera per s_1 e s_4 ma falsa per s_2 e s_3

Principali paradigmi di apprendimento



Ingredienti fondamentali

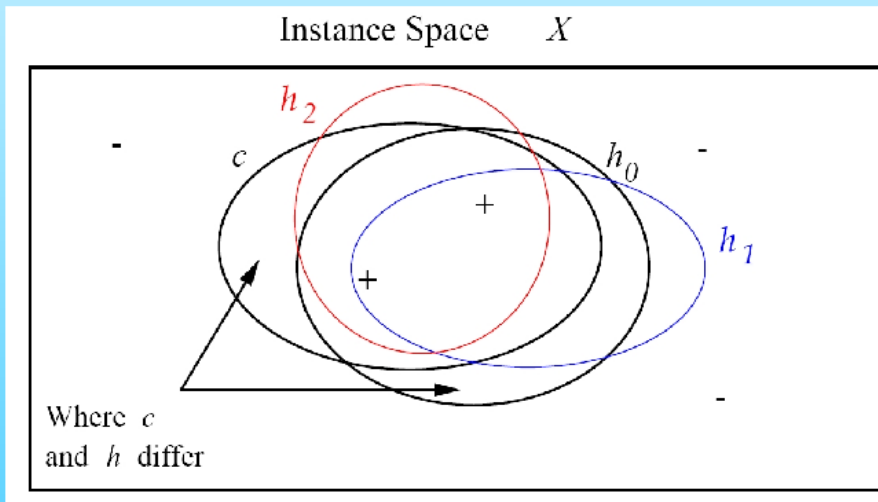


Errore Empirico ed Errore Ideale



Alcuni Modelli di Apprendimento

Errore Empirico ed Errore Ideale



Errore Ideale:
probabilità che h classifichi erroneamente un input selezionato dallo spazio delle istanze

(secondo la distribuzione di probabilità di occorrenza dell'input)

Errore Empirico:

(frazione del) numero di esempi classificati erroneamente da h



Overfitting

errore empirico(h_1) < errore empirico(h_2)

ma

errore ideale(h_1) > errore ideale(h_2)

Problema dell'overfitting

pochi dati



tante ipotesi con stesso errore empirico ...
... ma errore ideale maggiore o minore

Quale ipotesi scegliere ?

Problema dell'underfitting

poche ipotesi



nessuna ipotesi "spiega" bene i dati



errore empirico alto!

Come rimediare?

Soluzione

utilizzare uno spazio delle ipotesi che non sia

- né troppo semplice (underfitting)
- né troppo complesso (overfitting)

Occorre "misurare" la complessità dello spazio delle ipotesi

non è facile!!!

Occorre **misurare** la complessità dello spazio delle ipotesi

Bound sull'Errore Ideale

$$\text{errore ideale}(h) < \text{errore empirico}(h) + e(N, VC(H), d)$$

$e(N, VC(H), d)$ è

- inversamente proporzionale a N
- direttamente proporzionale a $VC(H)$

VC dimension

misura di complessità di H

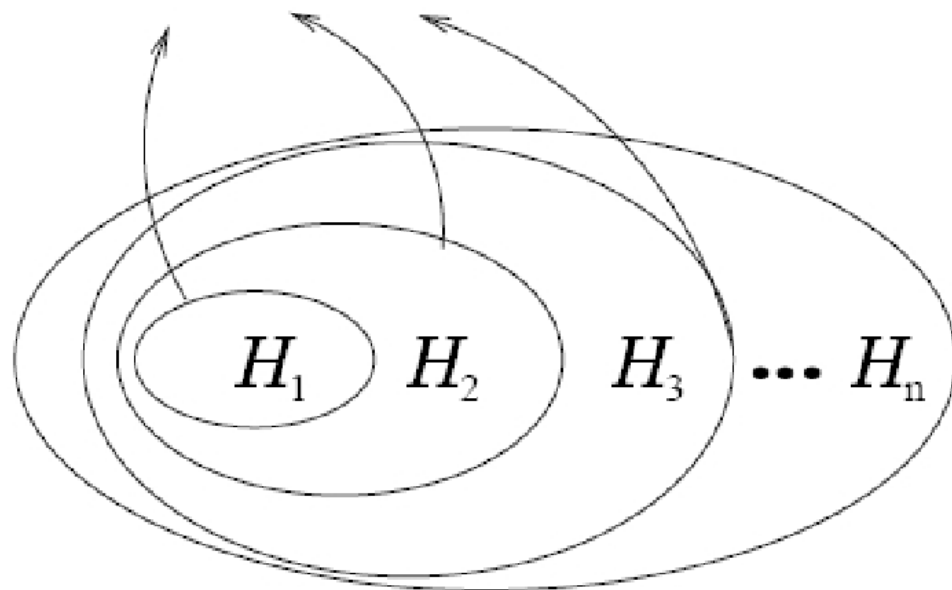
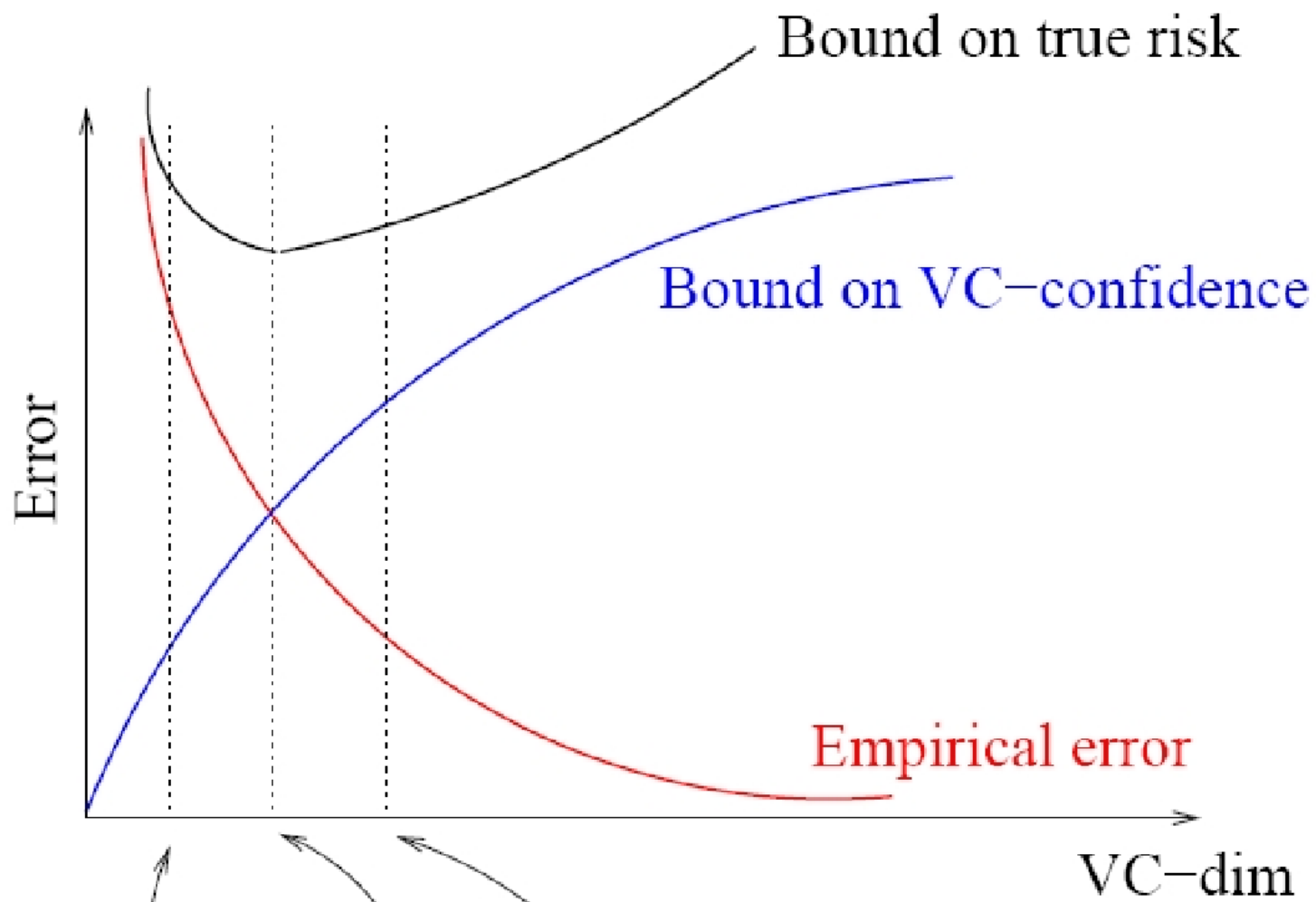


numero di esempi di apprendimento



il bound vale con probabilità $1-d$





0



Errore Empirico ed Errore Ideale



Alcuni Modelli di Apprendimento

Decision Trees

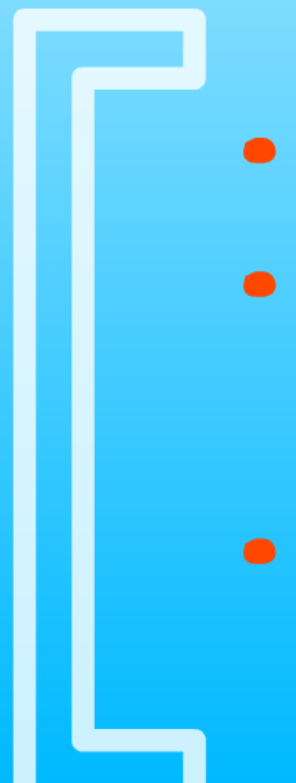


- istanze rappresentate da coppie attributo-valore
- classificazione multiclasse
- esempi di apprendimento possono contenere errori e/o valori mancanti

- **nodo:** test su attributo
- **ramo:** corrisponde ad un possibile valore dell'attributo
- **foglia:** classificazione



- istan
- attr
- class
- esen
- cont



- esempi di apprendimento possono contenere errori e/o valori mancanti

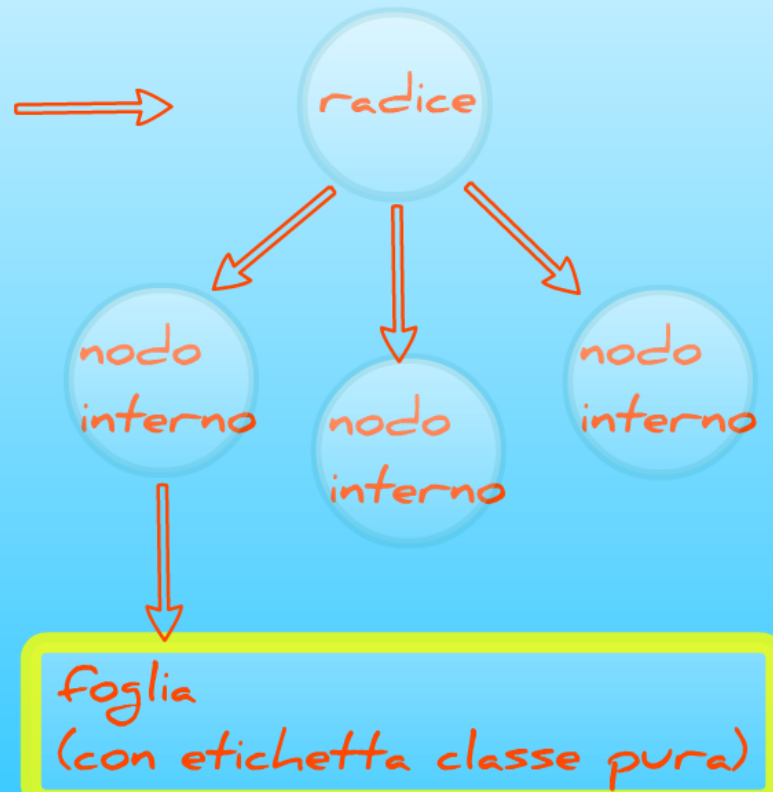
- nodo: test su attributo
- ramo: corrisponde ad un possibile valore dell'attributo
- foglia: classificazione

Algoritmo di apprendimento

training set \longrightarrow scegli attributo più "discriminante" \longrightarrow

\swarrow
partiziona training set usando i valori dell'attributo scelto

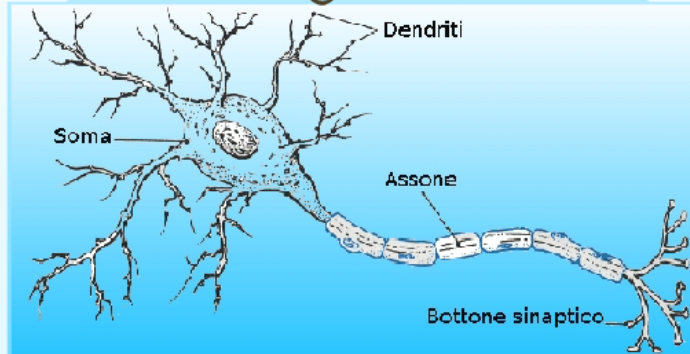
\downarrow
se un elemento della partizione contiene solo esempi della stessa classe (puro)



guadagno entropico

Neural Networks

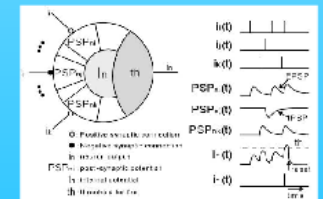
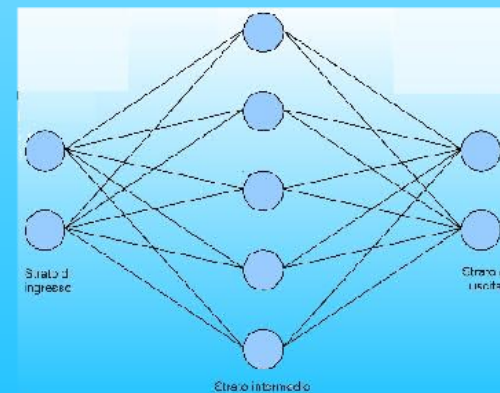
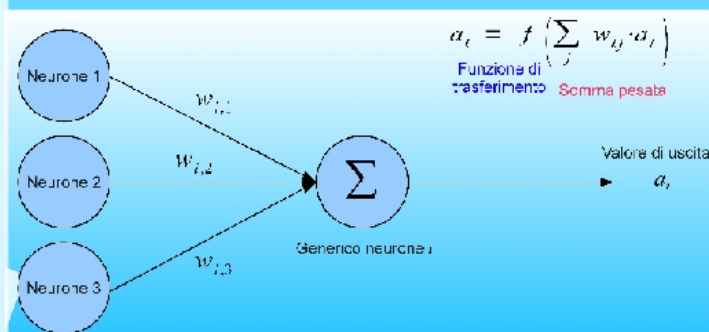
Neurone biologico



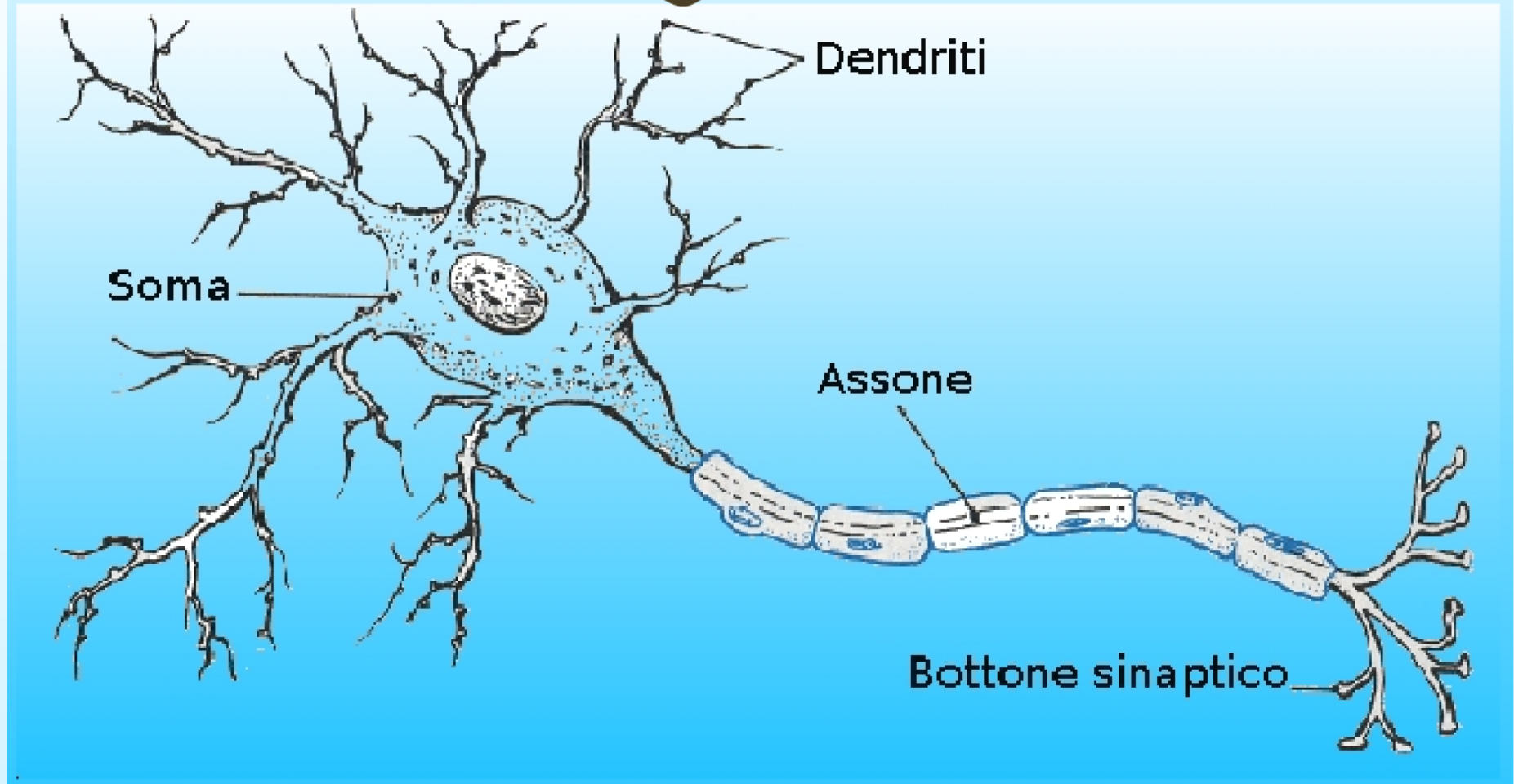
- si ispirano al cervello umano
- sia classificazione che regressione
- sia supervised che unsupervised
- adatte ad approssimare funzioni reali continue

Nikola Kasabov vi spiegherà anche la versione con spiking neurons!

Neurone artificiale

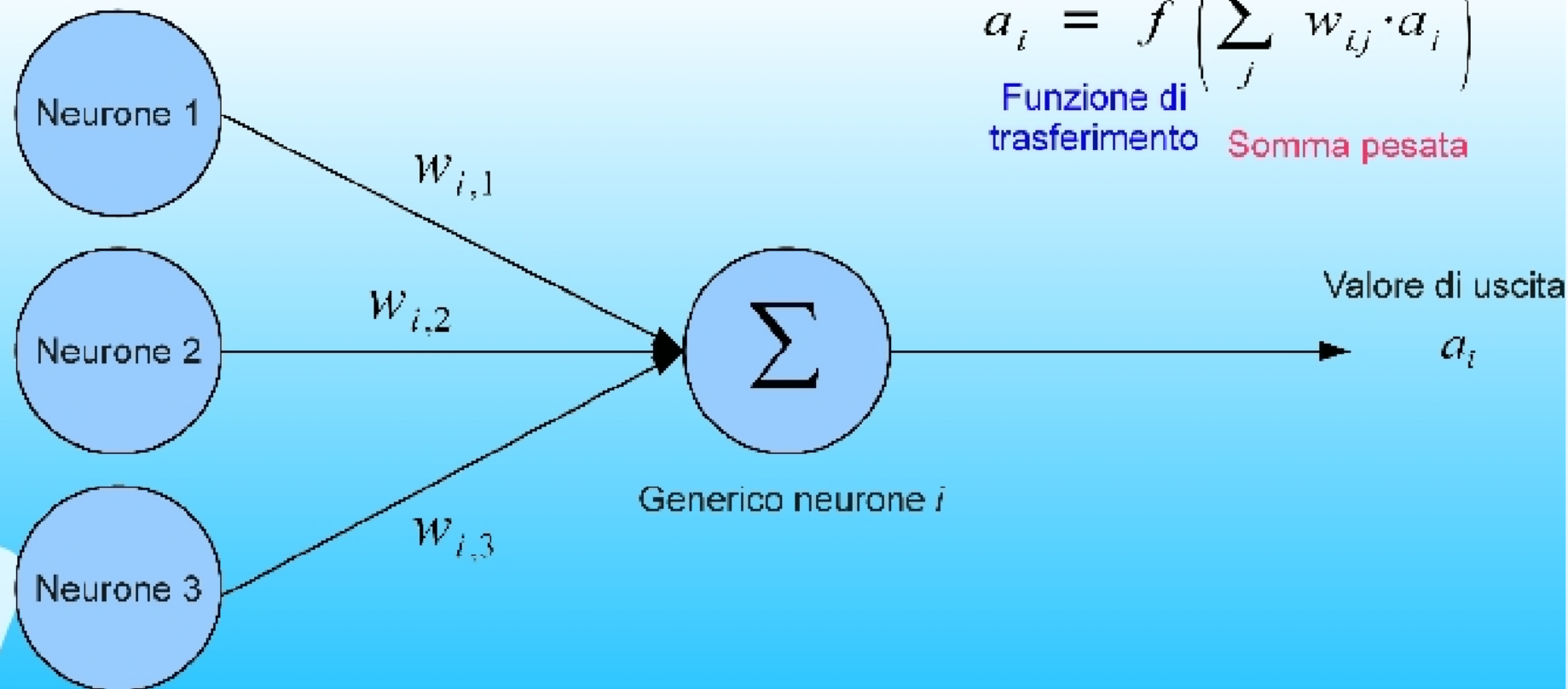


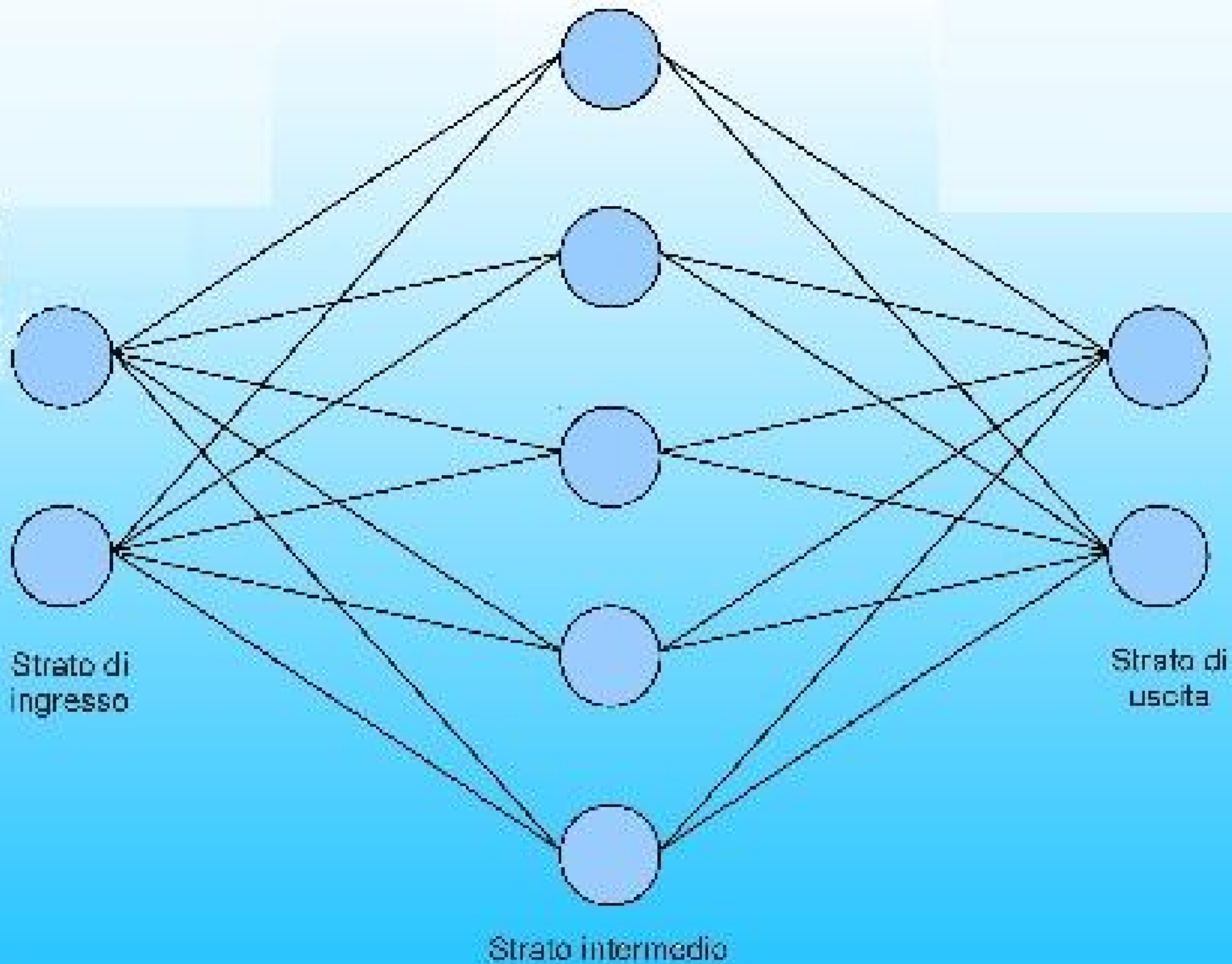
Neurone biologico



Neurone artificiale

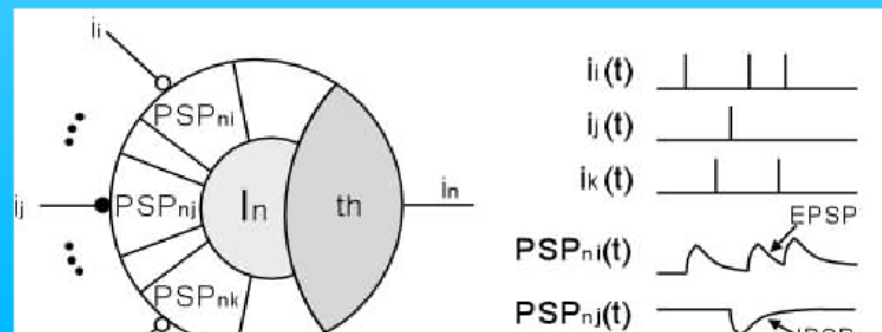
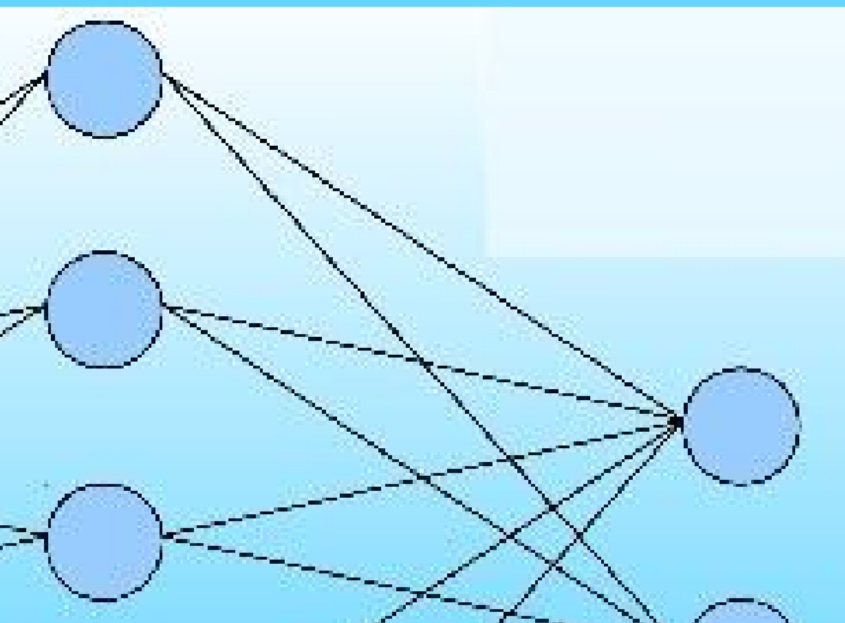
Neurone artificiale

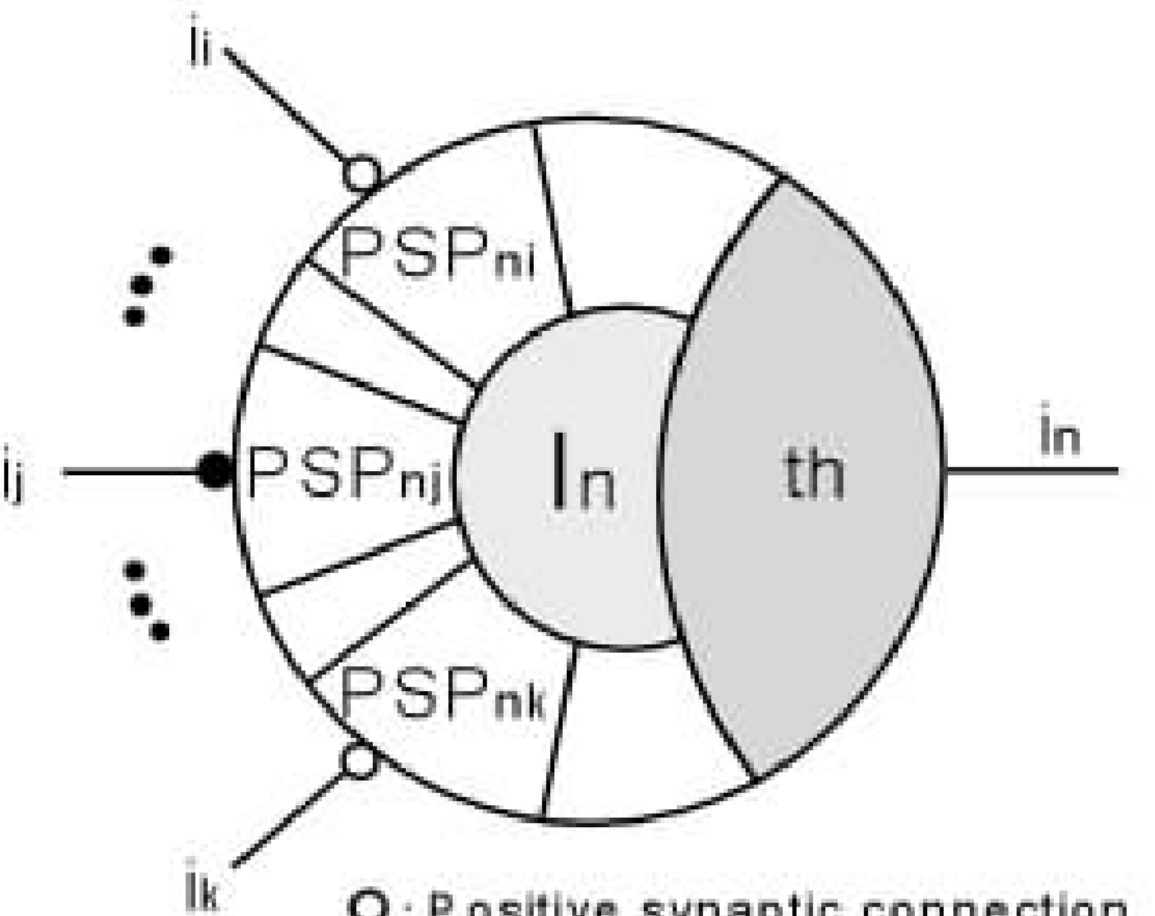




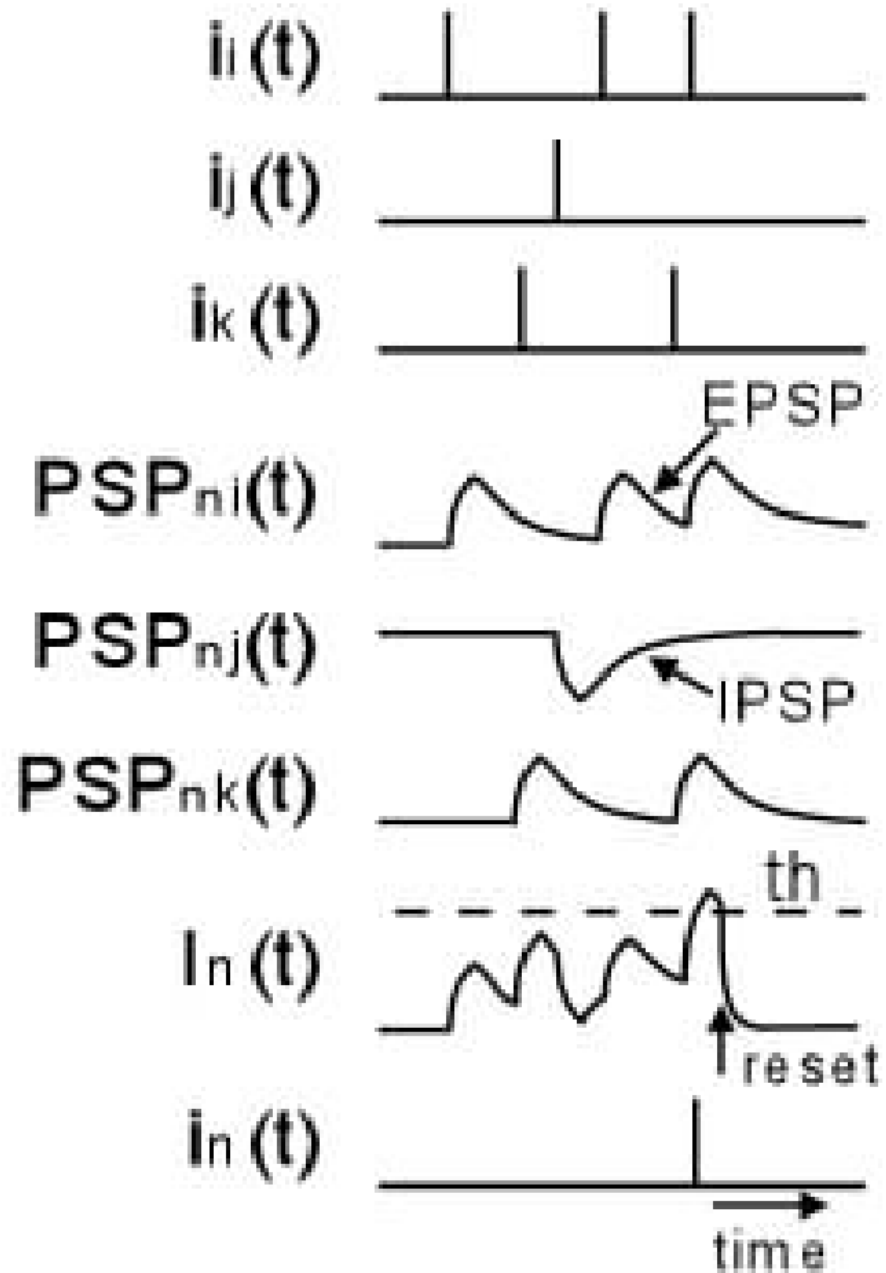
che unsupervised prossimare funzioni

Nikola Kasabov vi spiegherà anche la
versione con spiking neurons!



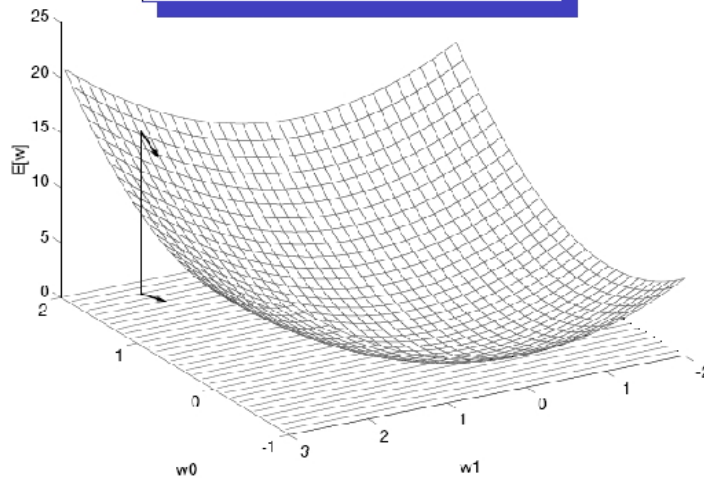


○ : Positive synaptic connection
 ● : Negative synaptic connection
 i_n : neuron output
 PSP_{ni} : post-synaptic potential
 I_n : internal potential
 th : threshold for fire



Apprendimento

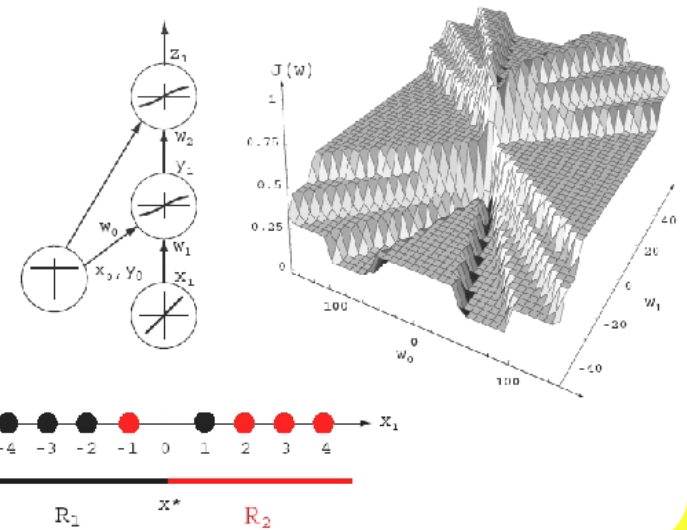
Discesa di Gradiente



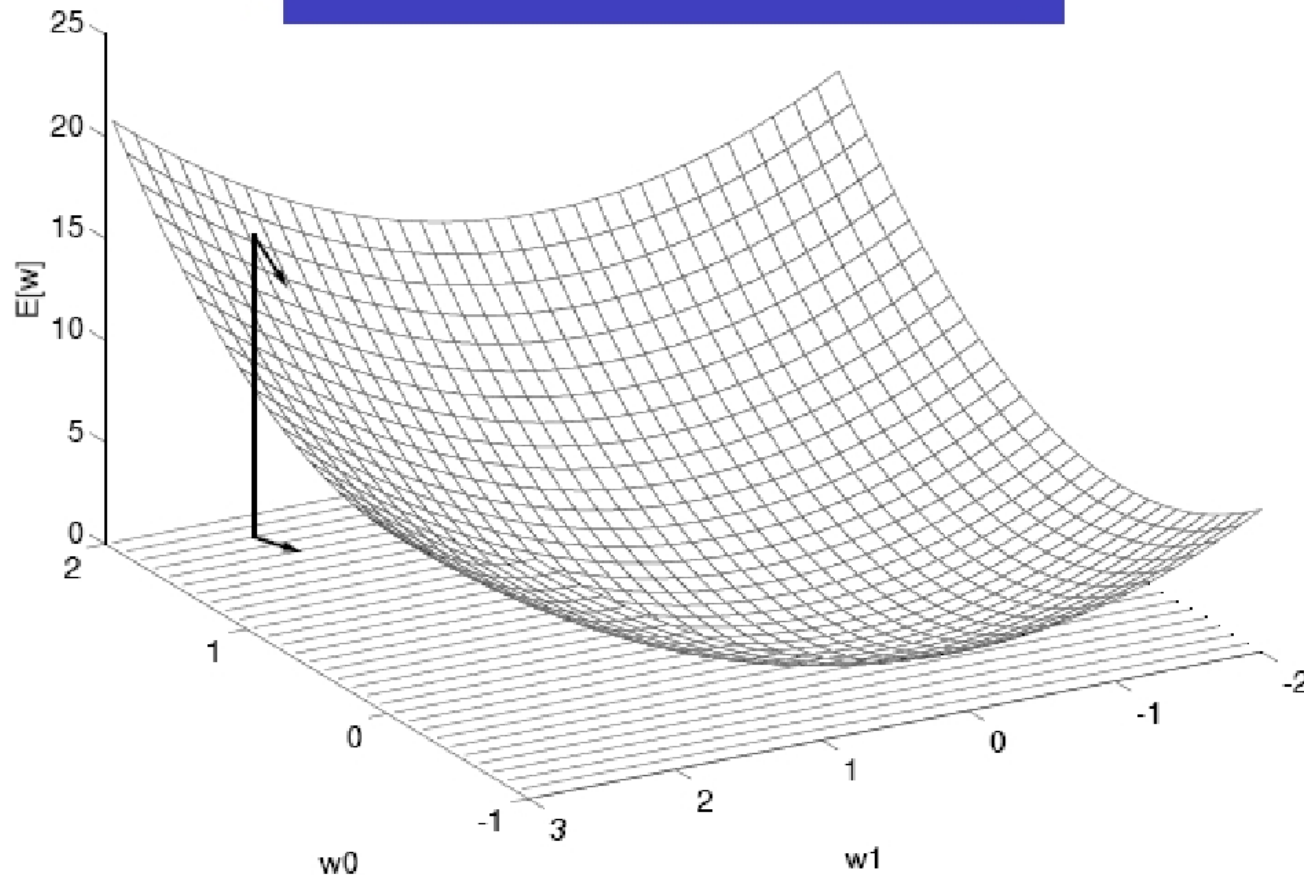
Idea base: partire da un \vec{w} random e modifi carlo nella direzione contraria al gradiente (che indica la direzione di crescita di $E[\vec{w}]$)

$$\underbrace{\nabla E[\vec{w}]}_{\text{gradiente}} \equiv \left[\frac{\partial E}{\partial w_0}, \frac{\partial E}{\partial w_1}, \dots, \frac{\partial E}{\partial w_n} \right], \quad \Delta \vec{w} = -\eta \nabla E[\vec{w}], \quad \Delta w_i = -\eta \frac{\partial E}{\partial w_i}$$

Esempio di Funzione Errore



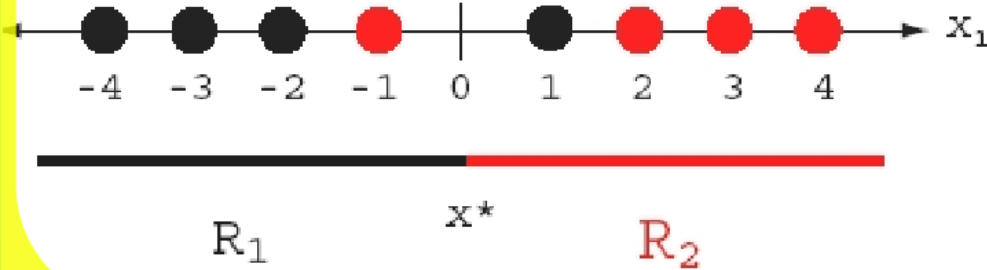
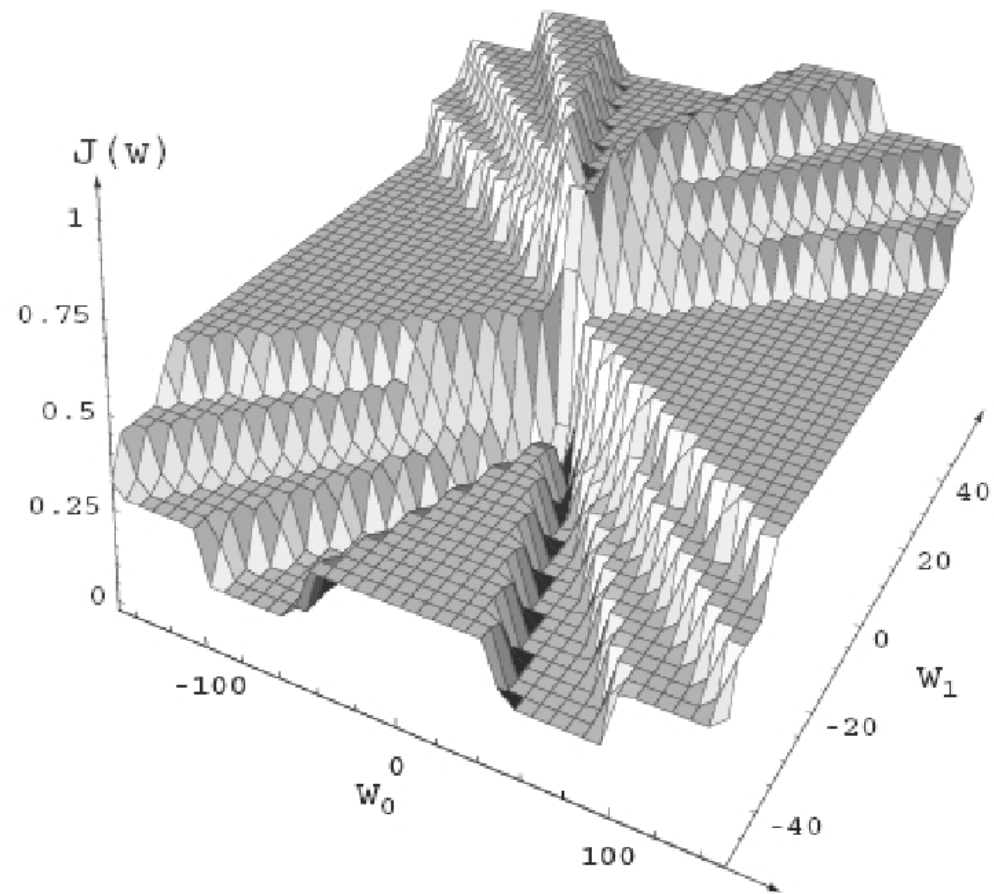
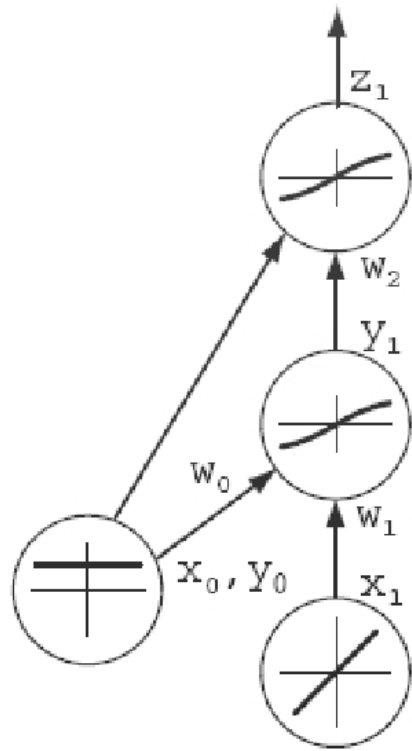
Discesa di Gradiente



Idea base: partire da un \vec{w} random e modificarlo nella direzione contraria al gradiente (che indica la direzione di crescita di $E[\vec{w}]$)

$$\underbrace{\nabla E[\vec{w}]}_{\text{gradiente}} \equiv \left[\frac{\partial E}{\partial w_0}, \frac{\partial E}{\partial w_1}, \dots, \frac{\partial E}{\partial w_n} \right], \quad \Delta \vec{w} = -\eta \nabla E[\vec{w}], \quad \Delta w_i = -\eta \frac{\partial E}{\partial w_i}$$

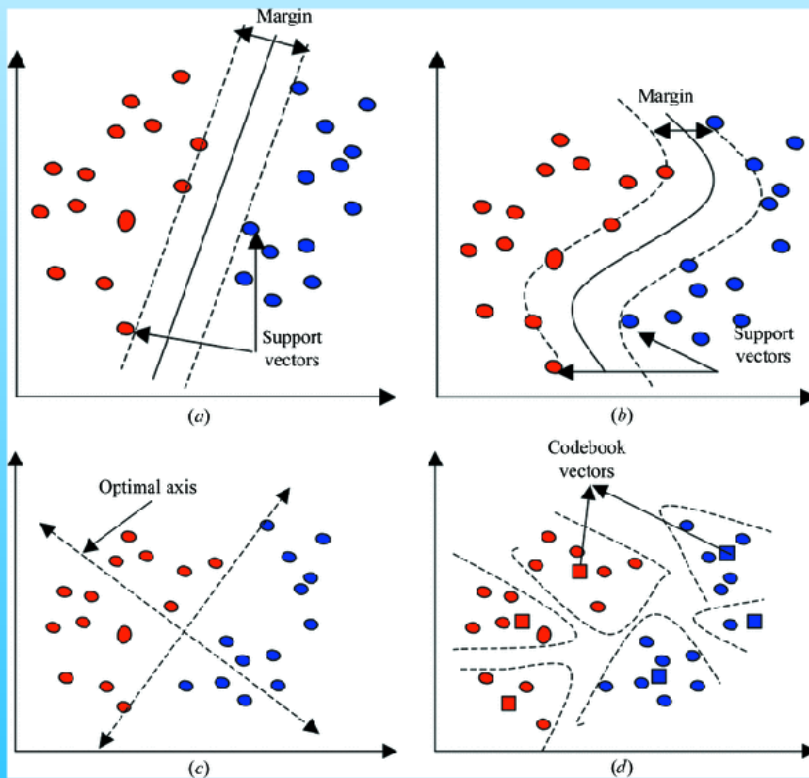
Esempio di Funzione Errore



he

$\frac{\partial E}{\partial w_i}$

Kernel Methods



- sia supervised che unsupervised
- sfruttano le funzioni kernel
- possono trattare direttamente dati strutturati
- fra le migliori tecniche di apprendimento automatico

Apprendimento di Support Vector Machines

l'apprendimento consiste nel risolvere un problema di ottimizzazione vincolata

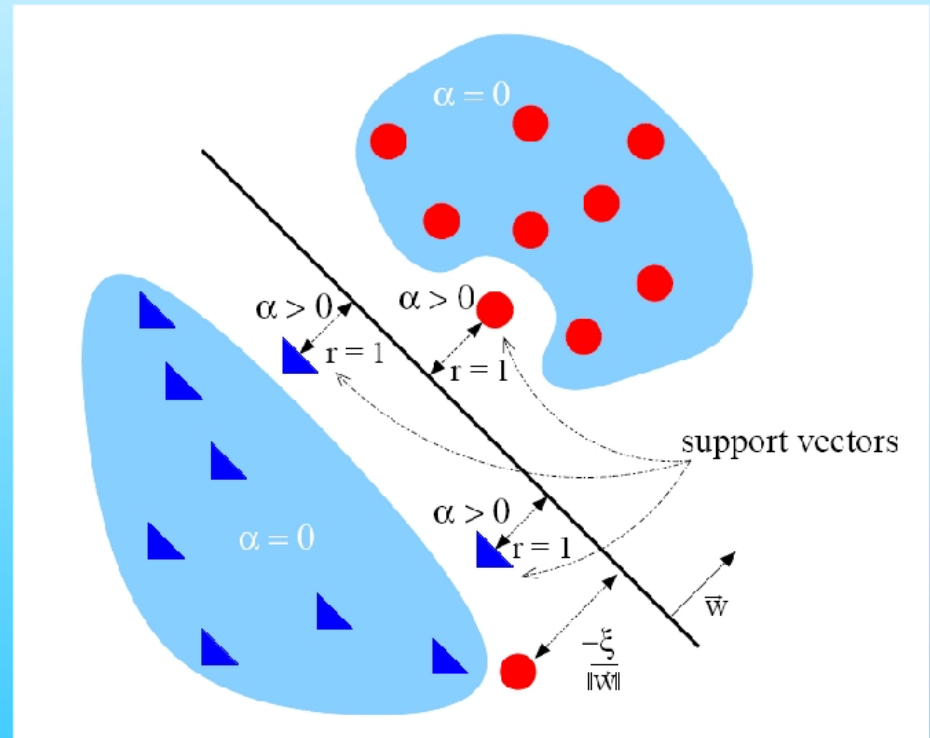
formulazione duale

$$\max_{\alpha} \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i,j=1}^n y_i y_j \alpha_i \alpha_j K(\vec{x}_i, \vec{x}_j)$$

soggetto a: $\forall i \in \{1, \dots, n\} : 0 \leq \alpha_i \leq C$ e $\sum_{i=1}^n y_i \alpha_i = 0$.

$K(x,y)$ funzione kernel:

- calcola il grado di "similarità" fra x e y
- si può definire anche su input strutturati

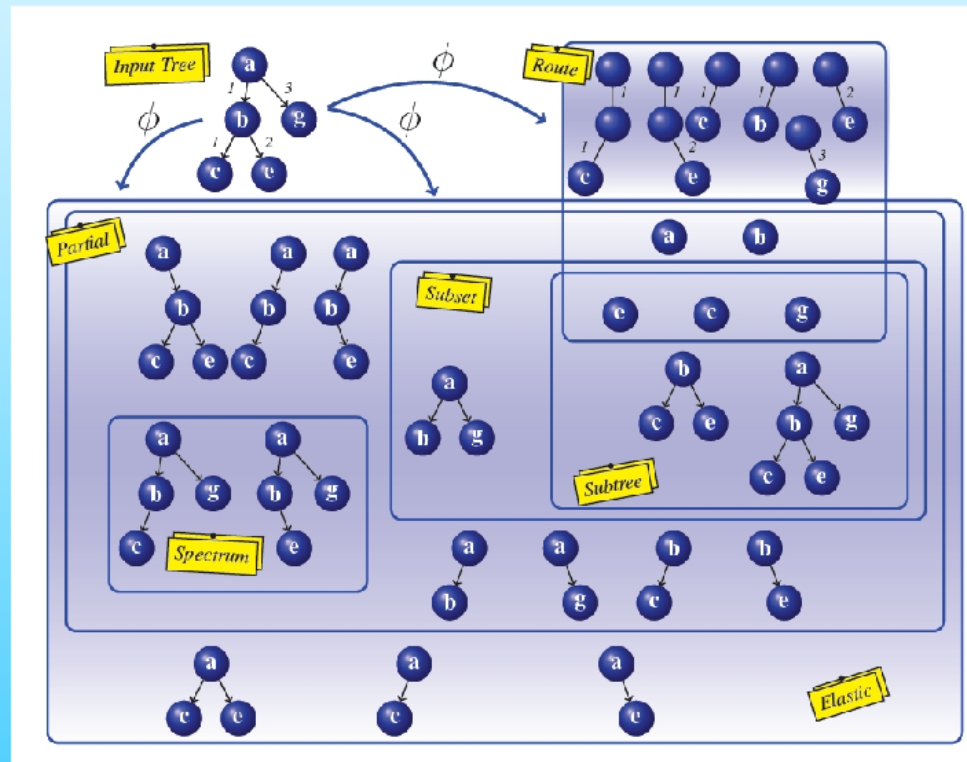


soggetto a: $\forall i \in \{1, \dots, n\} : 0 \leq \alpha_i \leq C$ e $\sum_{i=1}^n y_i \alpha_i$

$K(x, y)$ funzione kernel:

- calcola il grado di "similarità" fra x e y
- si può definire anche su input strutturati

$$K(\text{albero}_1, \text{albero}_2) = \text{Phi}(\text{albero}_1) \cdot \text{Phi}(\text{albero}_2)$$



Vedremo anche...

- Apprendimento Probabilistico
- Evolutionary Computation (Nikola Kasabov)
- Swarm Intelligence (Nikola Kasabov)
- Artificial Immune Systems (Nikola Kasabov)
- Mining Spatio-Temporal Data (Nikola Kasabov)
- On-line learning



e non finisce qui...