L'APPRENDIMENTO AUTOMATICO

MACHINE LEARNING

Alessandro Sperduti
Fabio Aiolli

Orario

Lunedì, Martedì, Mercoledì dalle 15:30 alle 17:30
Aula 1BC50, Torre Archimede

Lezioni tenute in inglese (Pierre Baldi):
6, 7, 8 Maggio + 9 Maggio (13:30-15:30, 1BC/45)

Lezioni in laboratorio P036:
15 Aprile (intro Python); 22, 28 Maggio; 3, 5 Giugno

16 e 17 Aprile
NON si terrà lezione

Il 16 e 17 Aprile
NON ci sarà lezione

Esame: orale

possibilità di compitini, se lo desiderate

- 13 Maggio: I compitino
- 10 Giugno: II compitino

Introduzione

↓

Principali paradigmi di apprendimento

↓

Ingredienti fondamentali

↓

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

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_{pj} (s_{qj}) is *Very Insensitive* (VI) (*Very Sensitive* (VS)).

Che cifre sono rappresentate in questa immagine ?

STANFORD RACING
TEAM #F127
c/o Michael Montemerlo
Gates Hall 136
Computer Science Department
Stanford University
Stanford, CA 94305-9010

Patient103 time=1 → *Patient103* time=2 ··· → *Patient103* time=n

Patient103 time=1	*Patient103* time=2	*Patient103* time=n
Age: 23	Age: 23	Age: 23
FirstPregnancy: no	FirstPregnancy: no	FirstPregnancy: no
Anemia: no	Anemia: no	Anemia: no
Diabetes: no	Diabetes: YES	Diabetes: no
PreviousPrematureBirth: no	PreviousPrematureBirth: no	PreviousPrematureBirth: no
Ultrasound: ?	Ultrasound: abnormal	Ultrasound: ?
Elective C−Section: ?	Elective C−Section: no	Elective C−Section: no
Emergency C−Section: ?	Emergency C−Section: ?	Emergency C−Section: **Yes**
...	...	...

Rule 1:	***If*** Income (w_j) is *Very High* (**VH**) ***then*** the sensitivity to price (quality) parameter s_{pj} (s_{qj}) is *Very Insensitive* (**VI**) (*Very Sensitive* (**VS**)).
Rule 2:	***If*** Income (w_j) is *High* (**HG**) ***then*** the sensitivity to price (quality) parameter s_{pj} (s_{qj}) is *Insensitive* (**IS**) (*Sensitive* (**ST**)).
Rule 3:	***If*** Income (w_j) is *Medium* (**MD**) ***then*** the sensitivity to price (quality) parameter s_{pj} (s_{qj}) is *Medium Sensitive* (**MS**).
Rule 4:	***If*** Income (w_j) is *Low* (**LW**) ***then*** the sensitivity to price (quality) parameter s_{pj} (s_{qj}) is *Sensitive* (*Insensitive* (**IS**)).
Rule 5:	***If*** Income (w_j) is *Very Low* (**VL**) ***then*** the sensitivity to price (quality) parameter s_{pj} (s_{qj}) is *Very Sensitive* (*Very Insensitive* (**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 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
nprecisa (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

Principali Paradigmi di Apprendimento

Supervised Learning

Unsupervised Learning

Reinforcement Learning

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
(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 $f(x_i)$

ervised Learning

x_2

Face recognition technologies

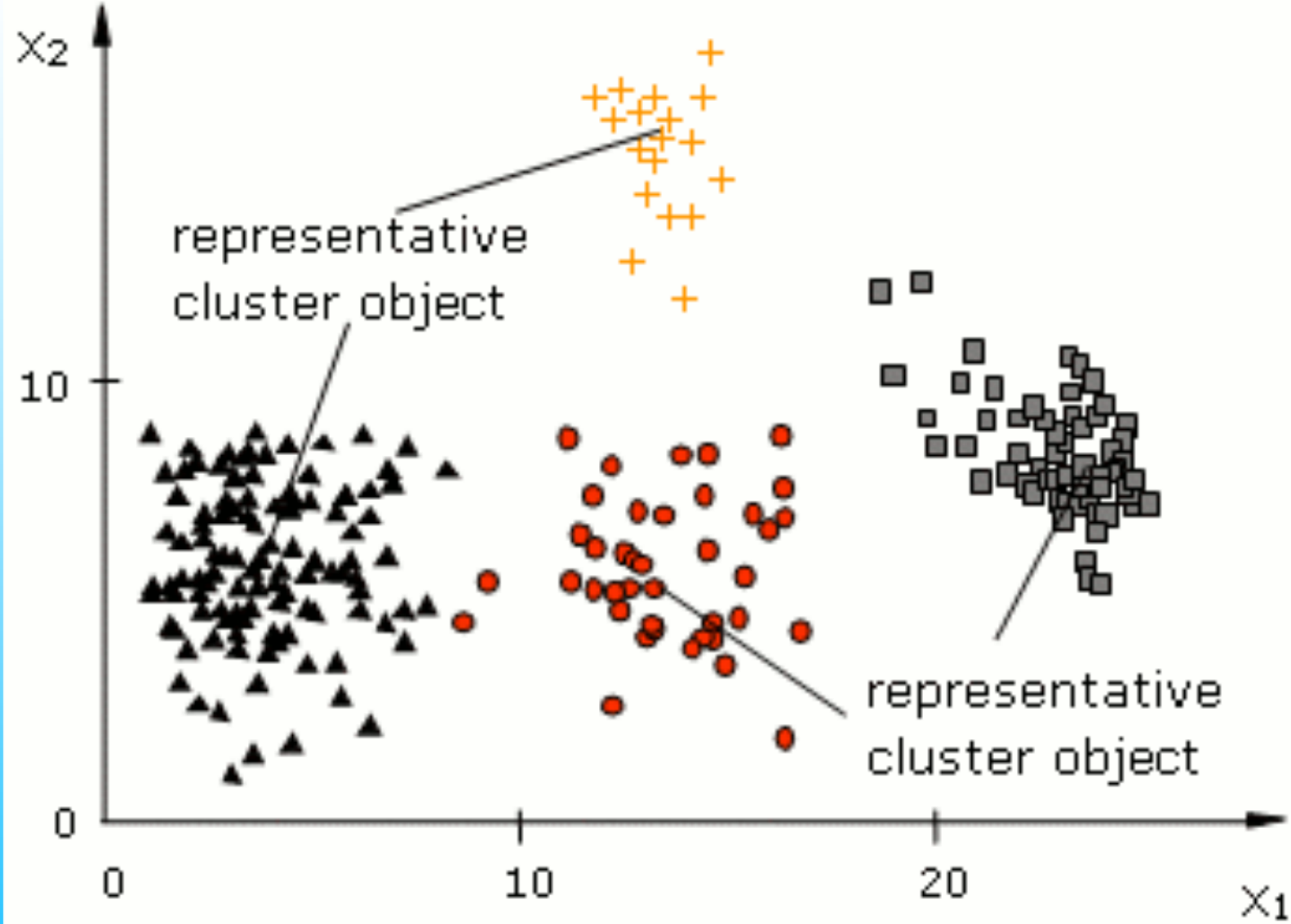
AT&T
LeNet 5
RESEARCH
answer: 0
0
103

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

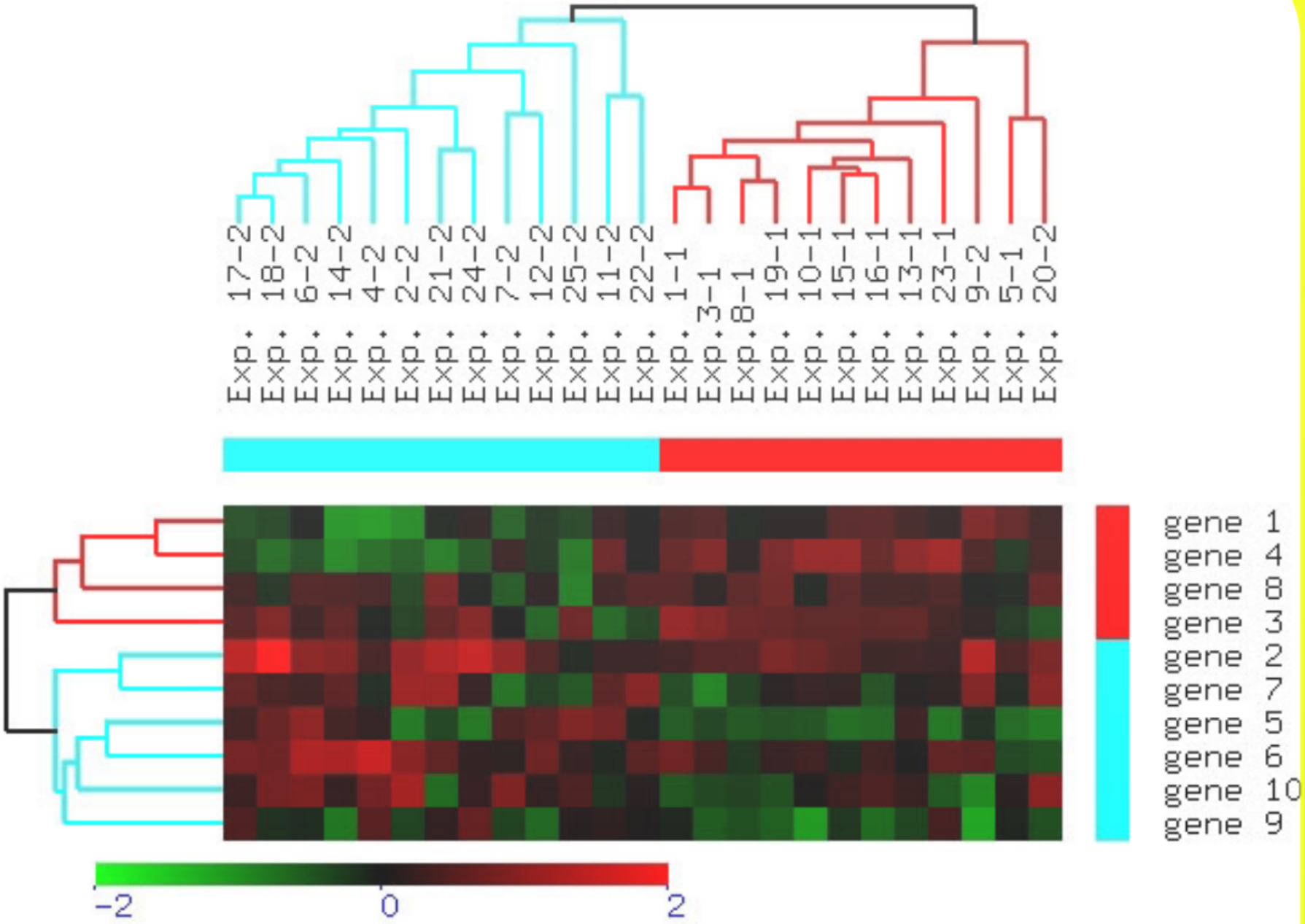
Unsupervised Learning
(apprendimento non-supervisionato)

- 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 (apprendimento con rinforzo)
REMO A LEZIONE



number of clusters: 4


Exp. 17-2
Exp. 18-2
Exp. 6-2
Exp. 14-2
Exp. 4-2
Exp. 2-2
Exp. 21-2
Exp. 24-2
Exp. 7-2
Exp. 12-2
Exp. 25-2
Exp. 11-2
Exp. 22-2
Exp. 1-1
Exp. 3-1
Exp. 8-1
Exp. 19-1
Exp. 10-1
Exp. 15-1
Exp. 16-1
Exp. 13-1
Exp. 23-1
Exp. 9-2
Exp. 5-1
Exp. 20-2
gene 1
gene 4
gene 8
gene 3
gene 2
gene 7
gene 5
gene 6
gene 10
gene 9
-2
0
2

Reinforcement Learning (apprendimento con rinforzo)
NON LO VEDREMO A LEZIONE

Sono dati

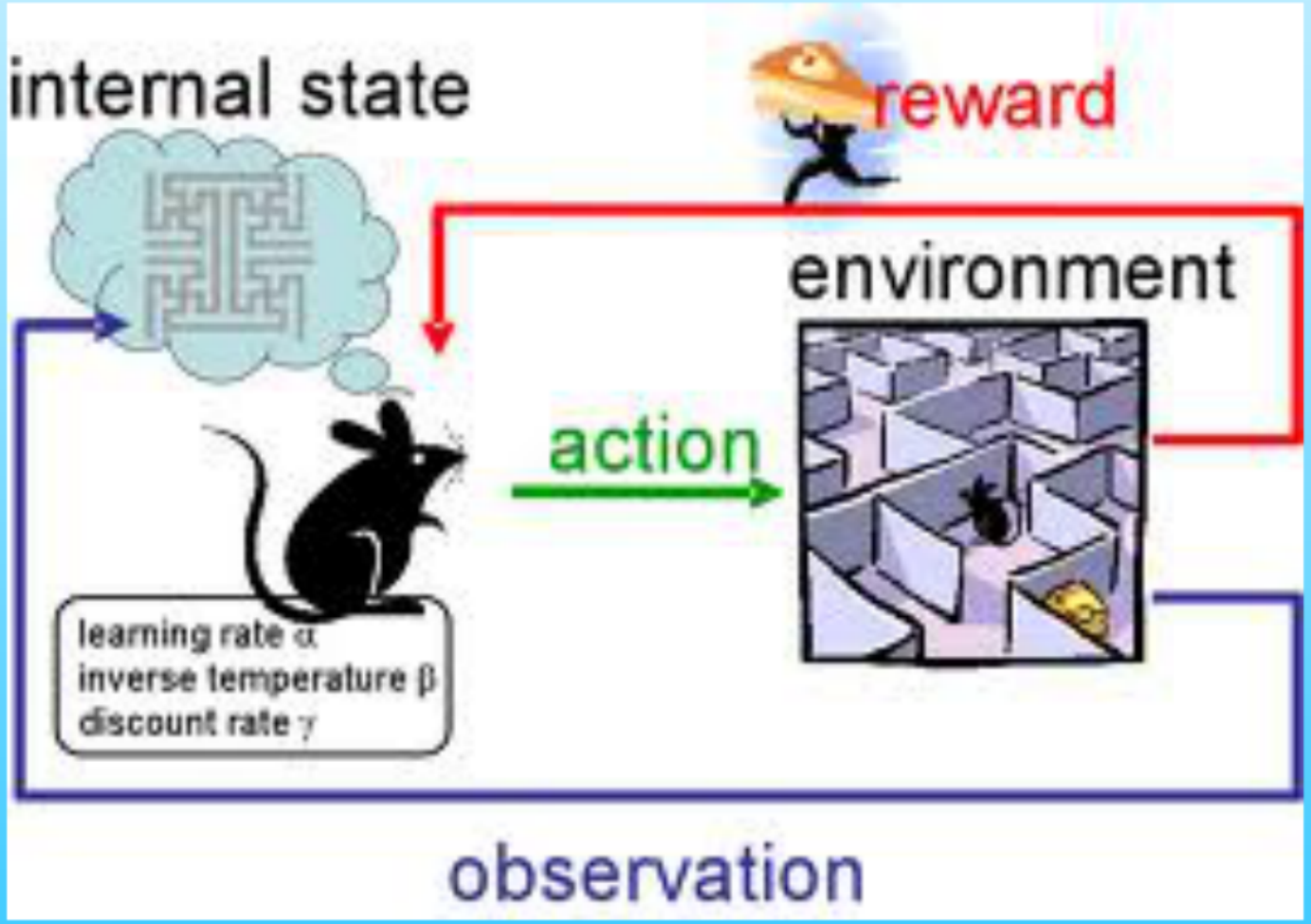
- un agente (intelligente ?), che può
  - trovarsi in uno stato s
  - eseguire una azione a (all'interno delle azioni possibili nello stato corrente)
- un ambiente tale che quando l'agente applica una azione a nello stato s restituisce
  - lo stato successivo
  - una ricompensa r, che può essere positiva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

Scopo dell'agente è quello di massimizzare la ricompensa

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

$\Downarrow$

scelta della strategia ottima



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

ATTENZIONE!!

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

$\Downarrow$

essere esaustiva

Apprendimento Inutile !!

Bias Induttivo

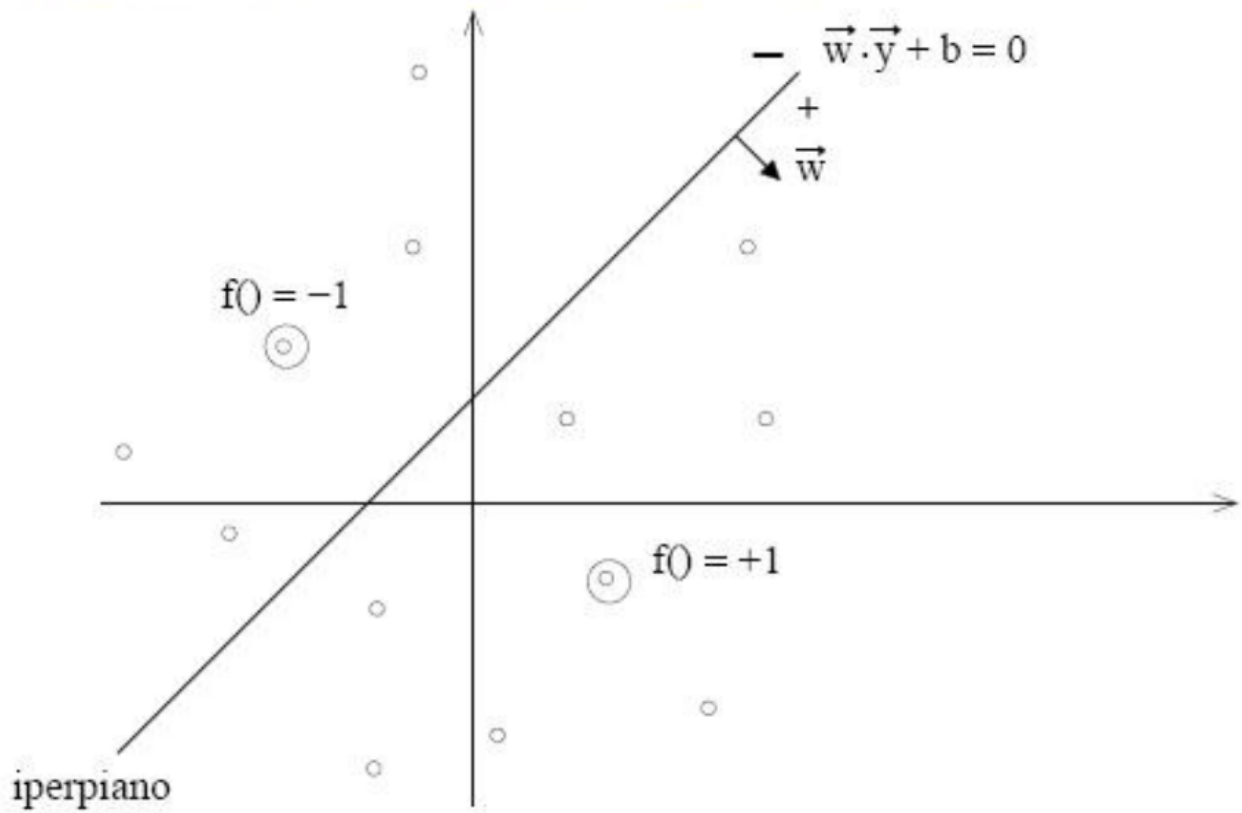
Bias Induttivo

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

Spazio delle Ipotesi: Esempio 1

Iperpiani in $\mathbb{R}^2$

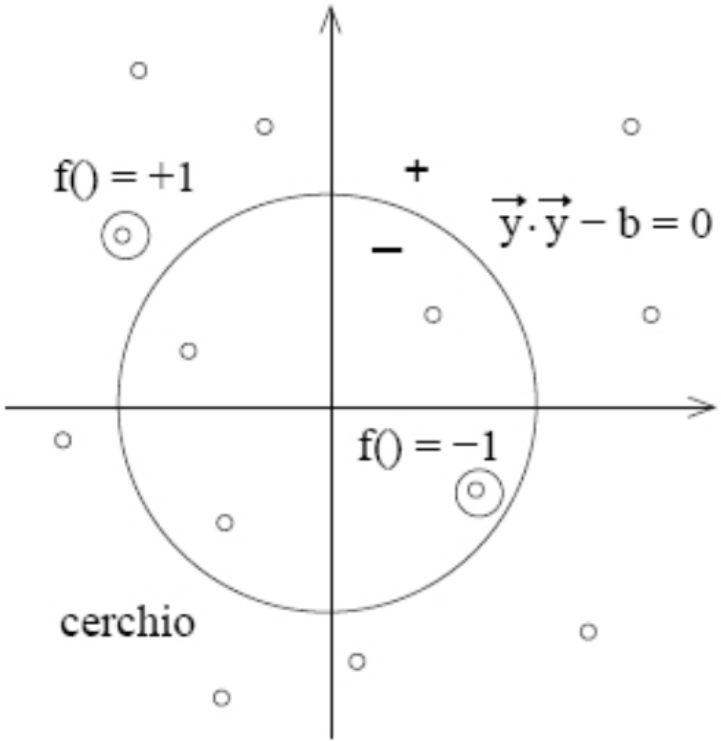
- Spazio delle Istanze $\rightarrow$ punti nel piano: $X = \{\vec{y} \in \mathbb{R}^2\}$
- Spazio delle Ipotesi $\rightarrow$ dicotomie indotte da iperpiani in $\mathbb{R}^2$:
  $\mathcal{H} = \{f_{(\vec{w},b)}(\vec{y}) | f_{(\vec{w},b)}(\vec{y}) = sign(\vec{w} \cdot \vec{y} + b), \vec{w} \in \mathbb{R}^2,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



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: $\mathcal{H} = \{f_b(\vec{y}) | f_b(\vec{y}) = sign(\vec{y} \cdot \vec{y} - b), b \in \mathbb{R}\}$



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 $l_1, \ldots, l_m$ (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**):

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

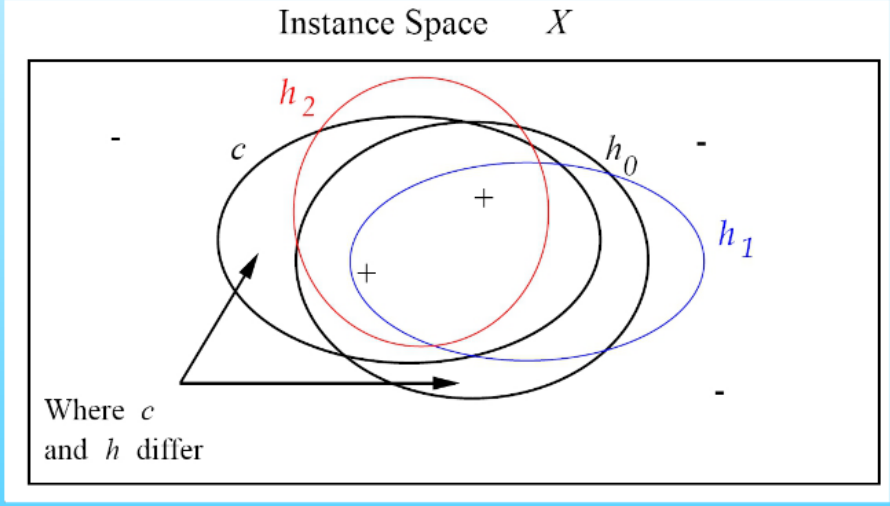
Es. $m = 3, X = \{0,1\}^3$

Esempi di istanze $\rightarrow s1 = 101, s2 = 001, s3 = 100, s4 = 111$

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

Notare che: h_1, h_2, e h_5 sono false per $s1$, $s2$ e $s3$ e vere per $s4$; h_3 è vera per ogni istanza; h_4 è vera per $s1$ e $s4$ ma falsa per $s2$ e $s3$

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

$\Downarrow$

tante ipotesi con stesso errore empirico ...

... ma errore ideale maggiore o minore

Quale ipotesi scegliere ?

Problema dell'underfitting

poche ipotesi

$\Downarrow$

nessuna ipotesi "spiega" bene i dati

$\Downarrow$

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

Occorre misurare la complessità dello spazio delle ipotesi

VC dimension
misura di complessità di H

Bound sull'Errore Ideale

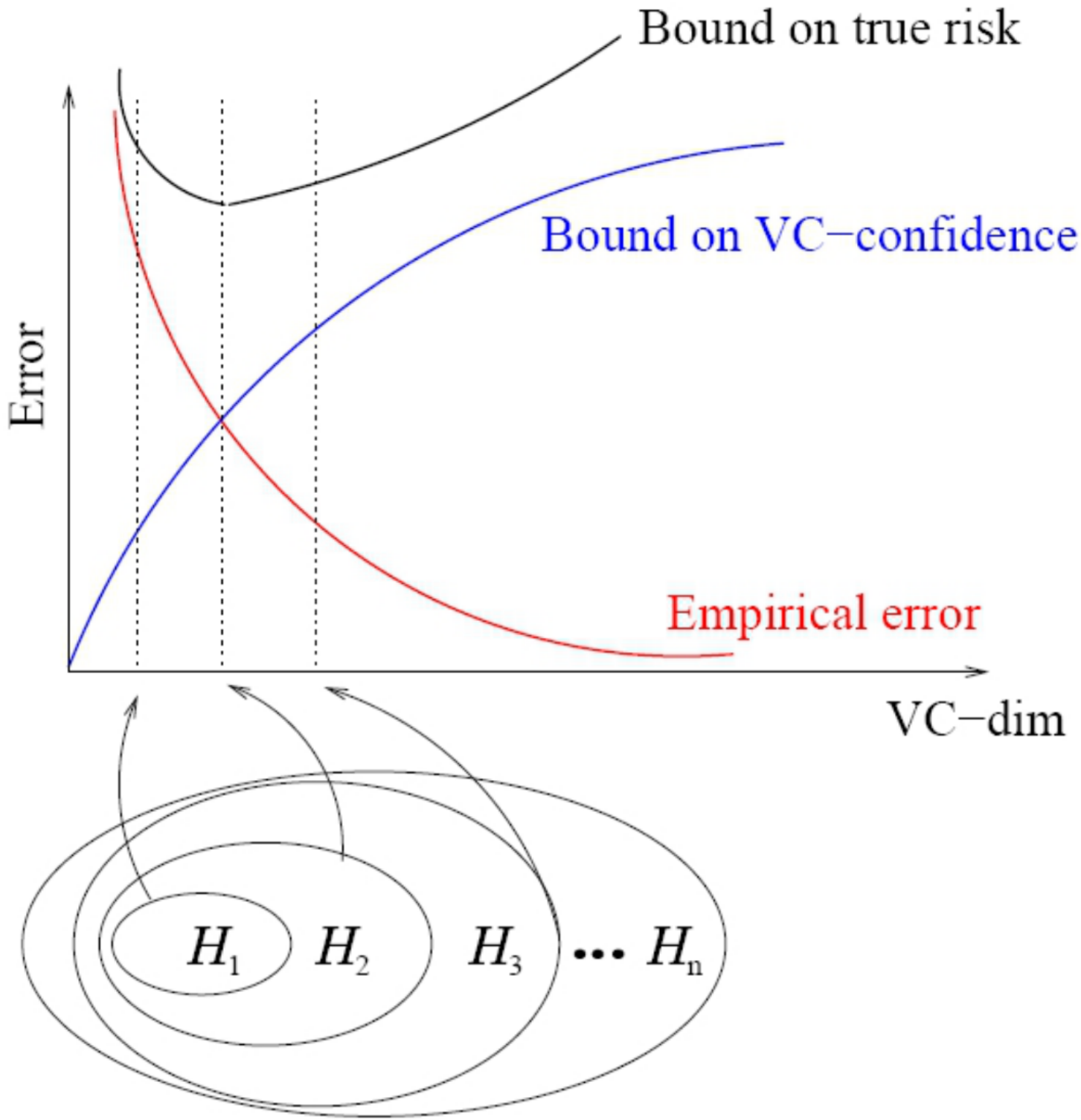
errore ideale(h) < errore empirico(h) + e(N,VC(H),d)

numero di esempi di apprendimento

il bound vale con probabilità 1-d

e(N,VC(H),d) è

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


Bound on true risk
Bound on VC−confidence
Empirical error
Error
VC−dim
H_1
H_2
H_3
... H_n

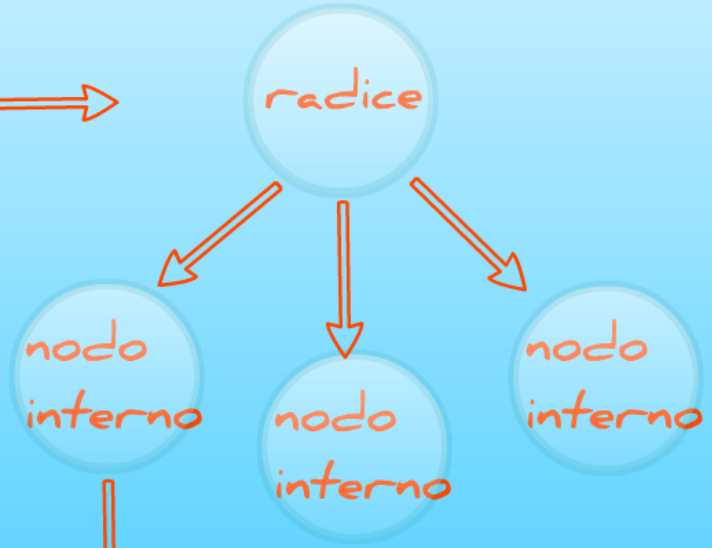
Decision Trees

- 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

Algoritmo di apprendimento

training set ⟹ scegli attributo più "discriminante" ⟹



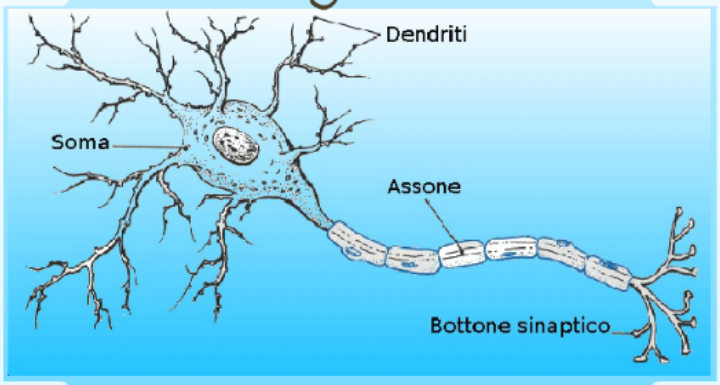
partiziona training set usando i valori dell'attributo scelto

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

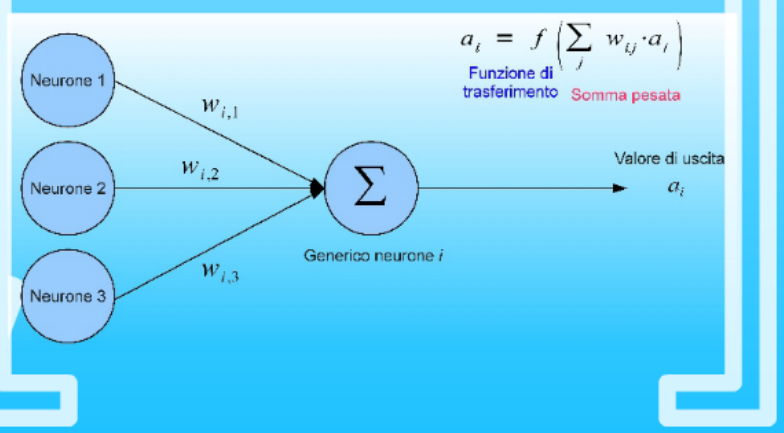
foglia
(con etichetta classe pura)

Neural Networks

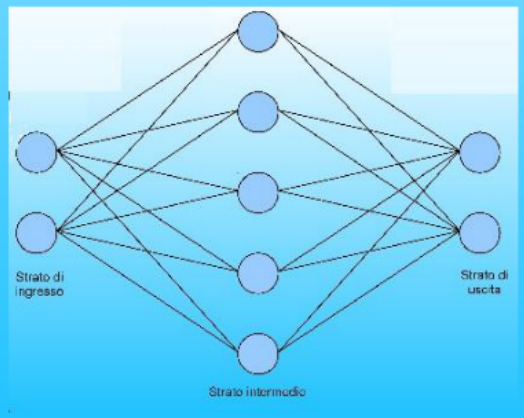
Neurone biologico



Neurone artificiale

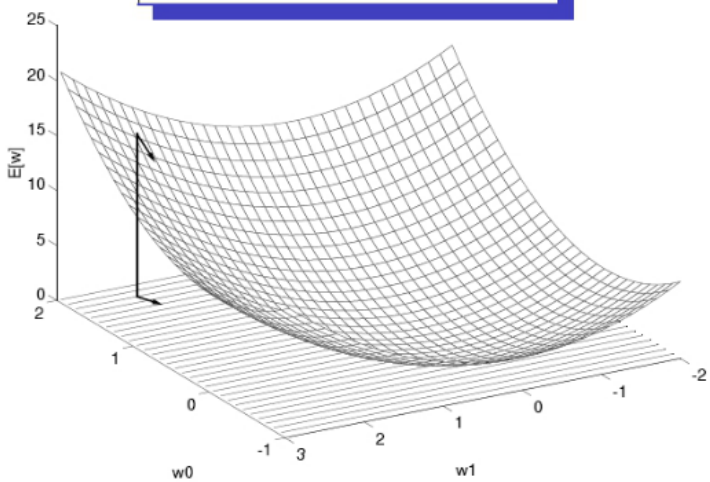


- si ispirano al cervello umano
- sia classificazione che regressione
- sia supervised che unsupervised
- adatte ad approssimare funzioni reali continue



Apprendimento

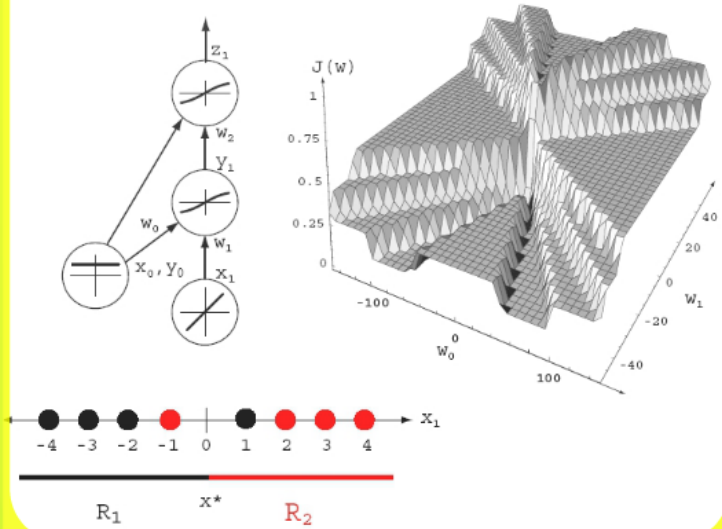
Discesa di Gradiente



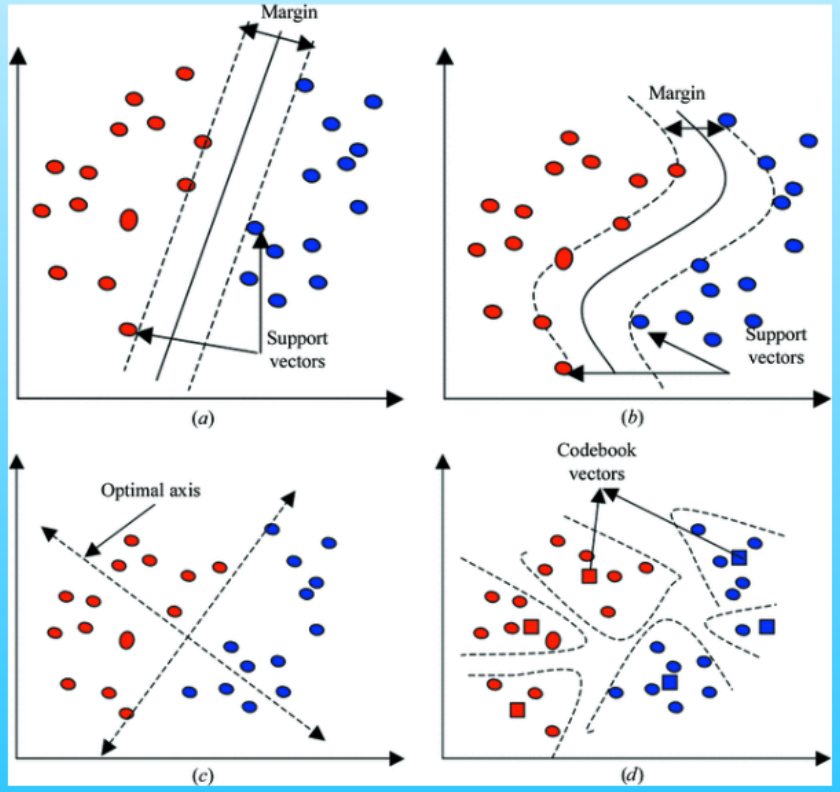
Idea base: partire da un $\vec{w}$ random e modifi carlo nella direzione contraria al gradiente (che indica la direzione di crescita di $E[\vec{w}]$)

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

Esempio di Funzione Errore



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

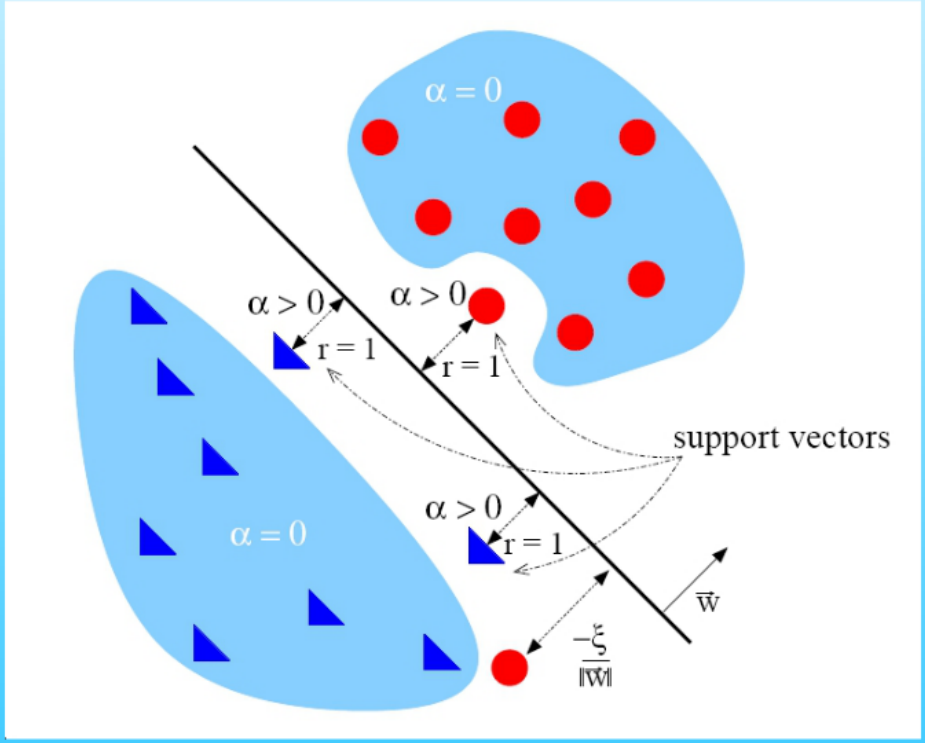
formulazione duale

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

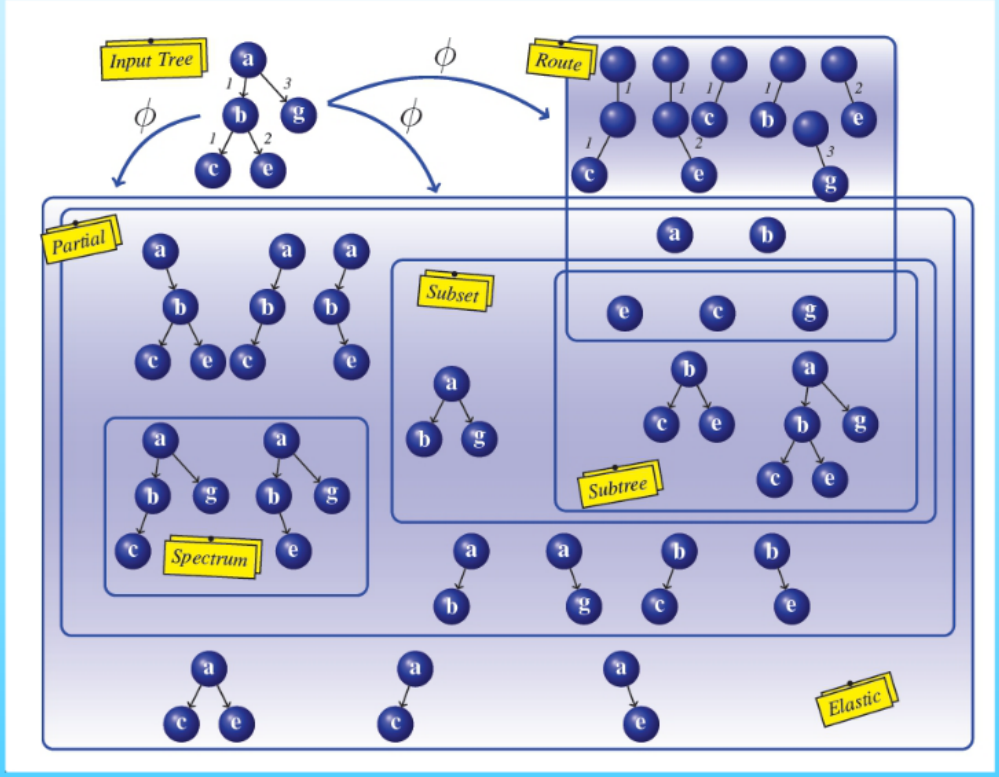
soggetto a: $\forall i \in \{1,\ldots,n\} : 0 \leq \alpha_i \leq C$ e $\sum_{i=1}^n y_i \alpha_i = 0.$

K(x,y) funzione kernel:

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



$$K(albero_1, albero_2) = Phi(albero_1) \cdot Phi(albero_2)$$



Vedremo anche...

- Rappresentazione delle istanze ("feature selection")
- Apprendimento Probabilistico (bayesiano)
- Algoritmi "semplici" di clustering

e non finisce qui...