Apprendimento Automatico

Libro di riferimento: Apprendimento Automatico, Tom Mitchell, McGraw Hill, 1998

Tutorial su SVM e Boosting

Lucidi

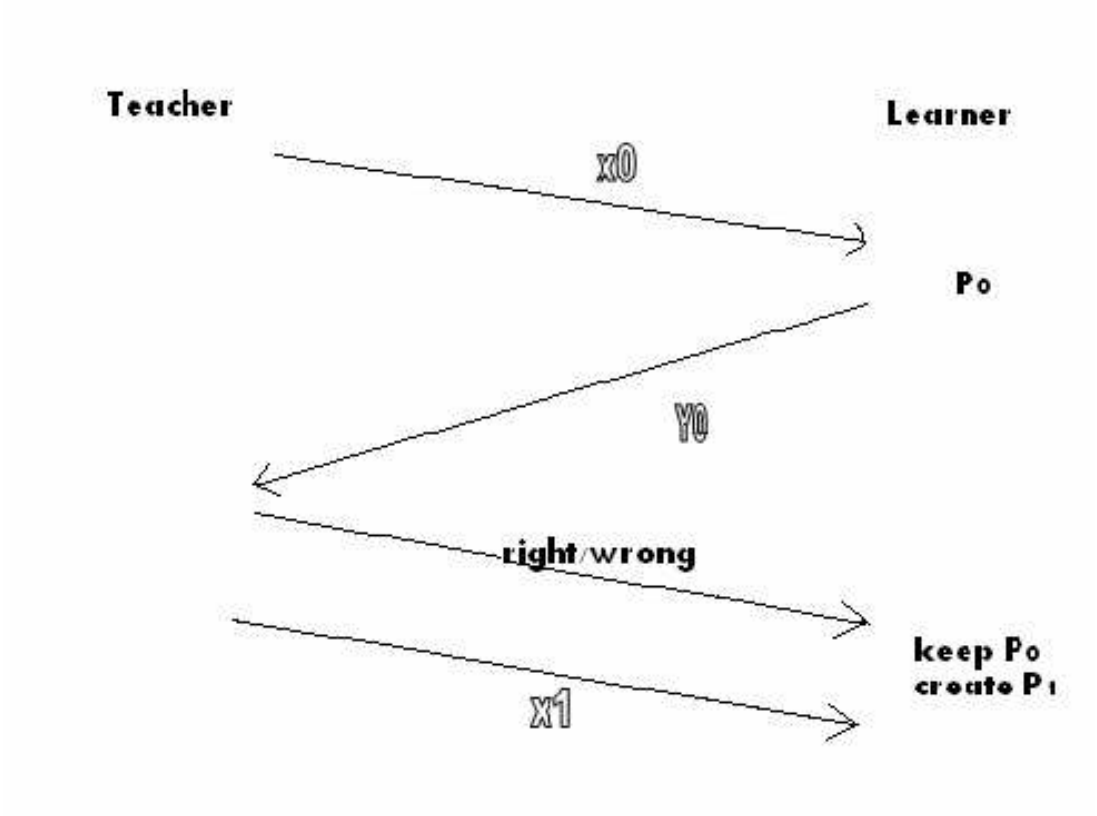
http://www.math.unipd.it/~sperduti/ml.html

Contenuti del corso

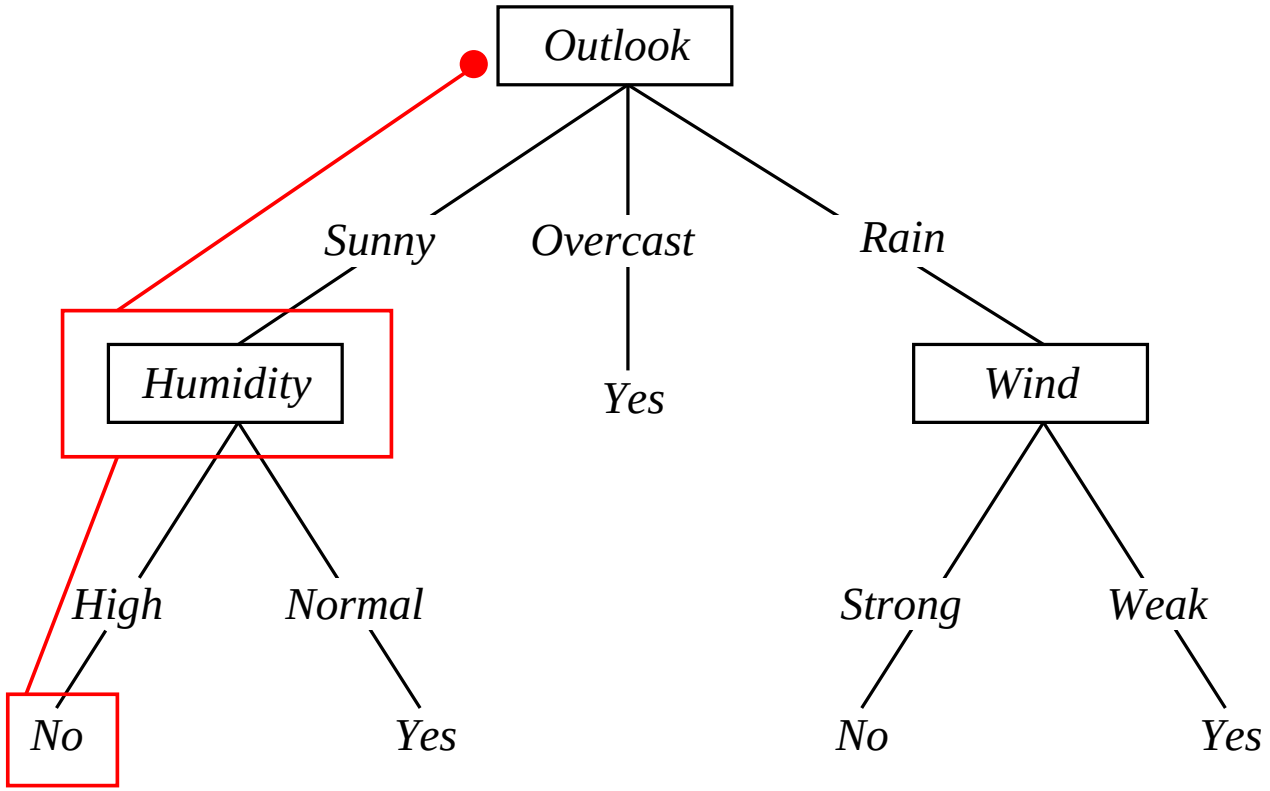
- Concetti fondamentali dell'apprendimento automatico
- Apprendimento on-line: alcuni semplici algoritmi e loro analisi teorica
- Apprendimento di alberi di decisione
- Boosting
- Reti Neurali
- Support Vector Machines
- Apprendimento probabilistico
- Apprendimento con rinforzo
- Esperienze pratiche in laboratorio

Apprendimento on-line

Schema generale



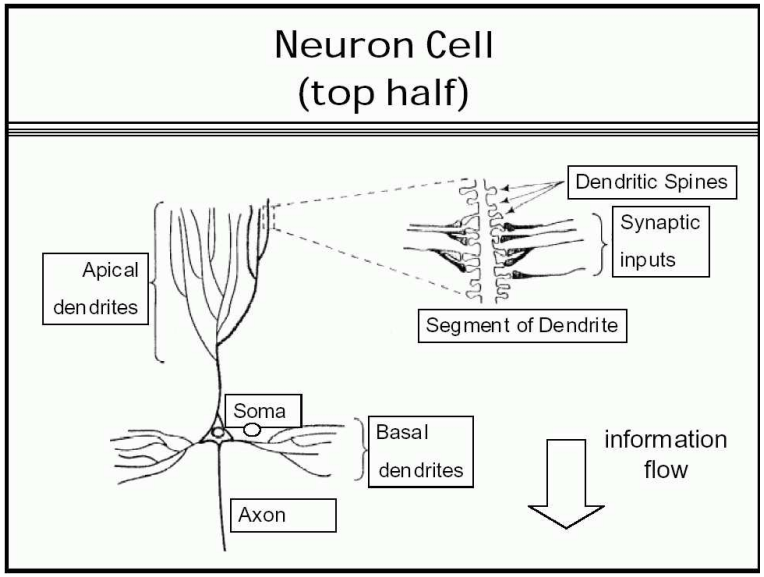
Apprendimento alberi di decisione



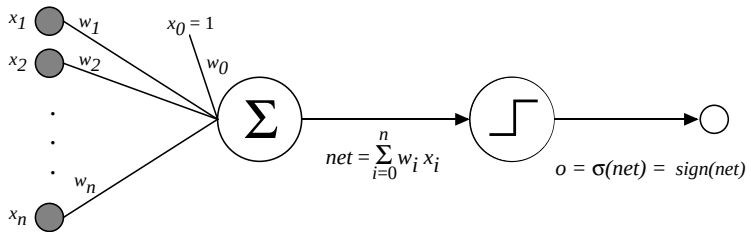
Boosting (di alberi di decisione)

Reti Neurali

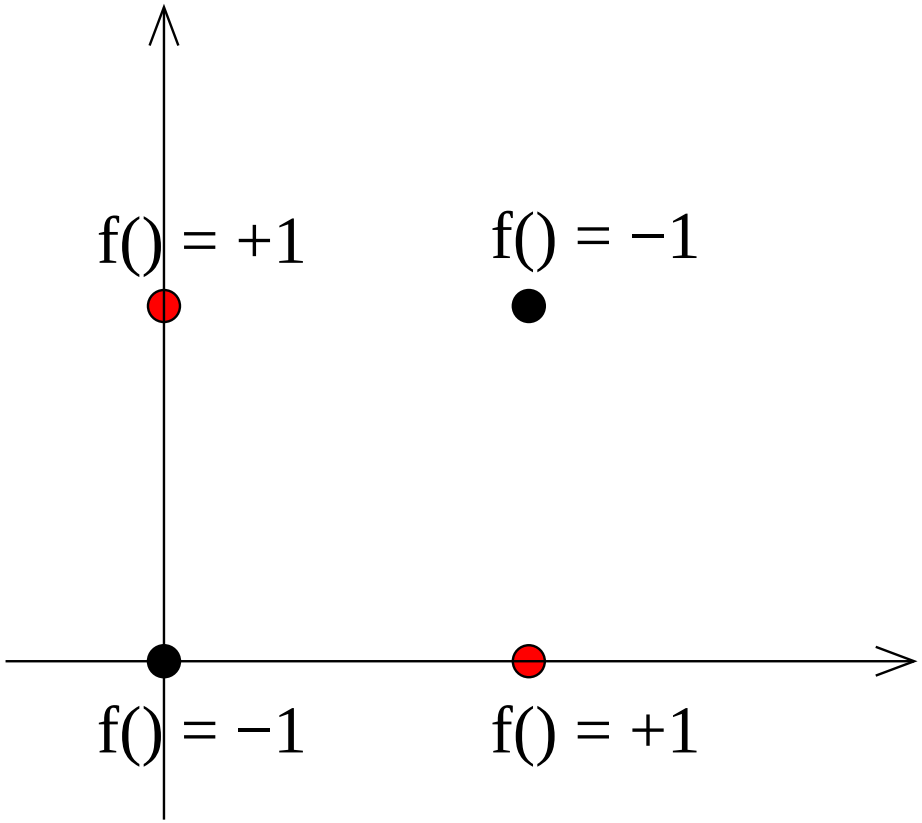
Neurone biologico

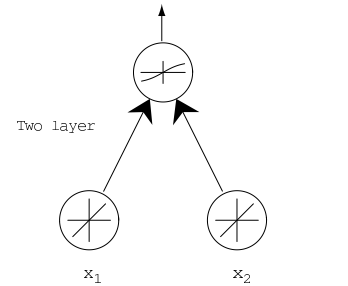


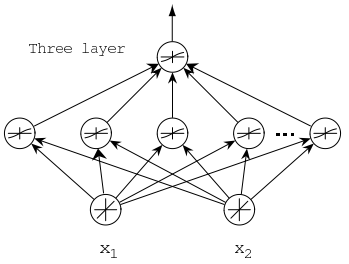
Neurone artificiale

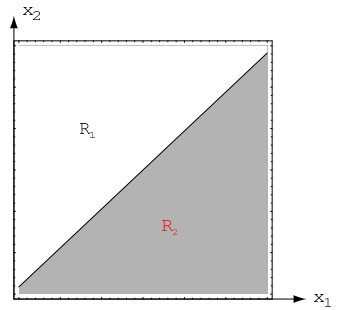


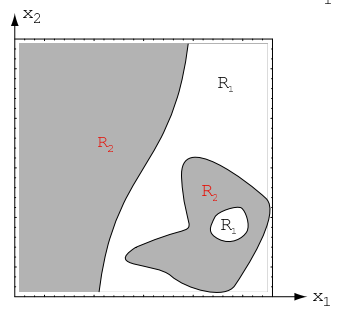
Reti Neurali



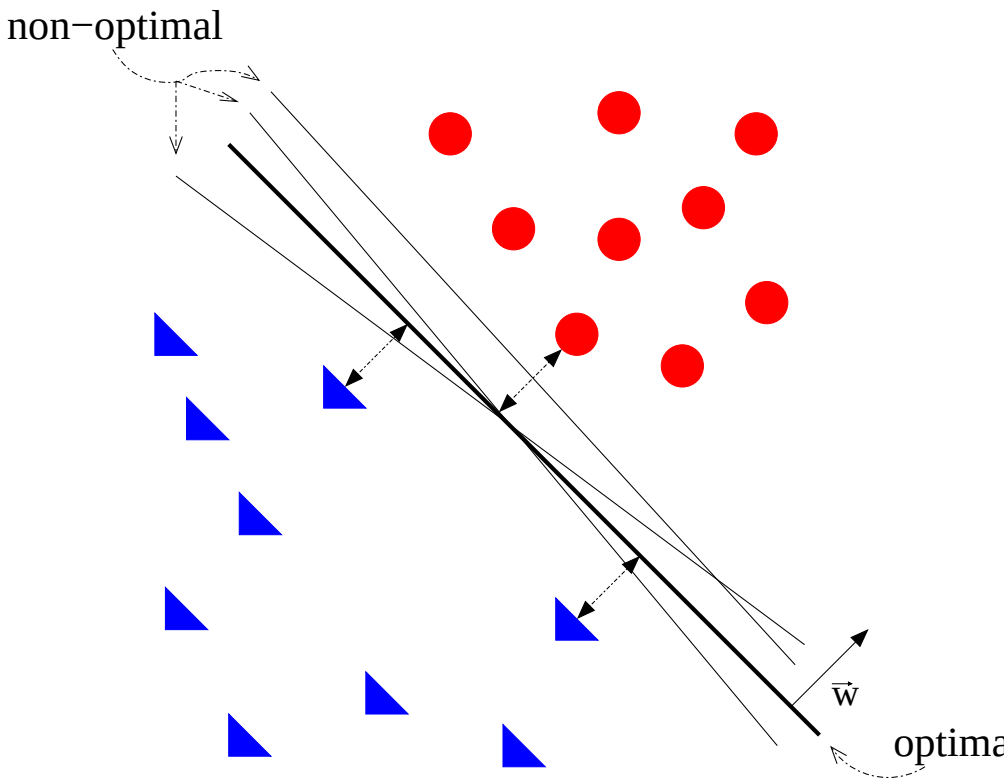




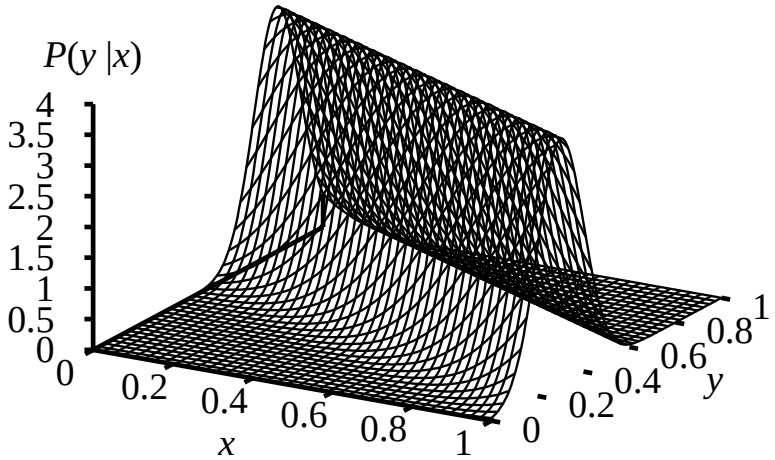


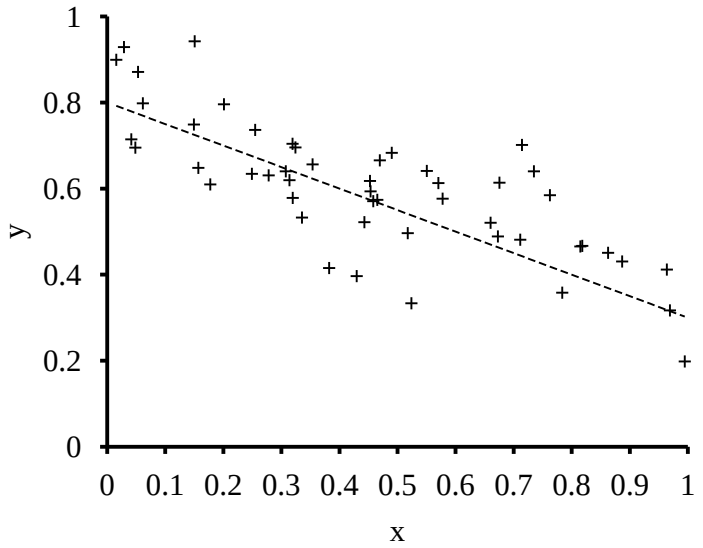


Support Vector Machines

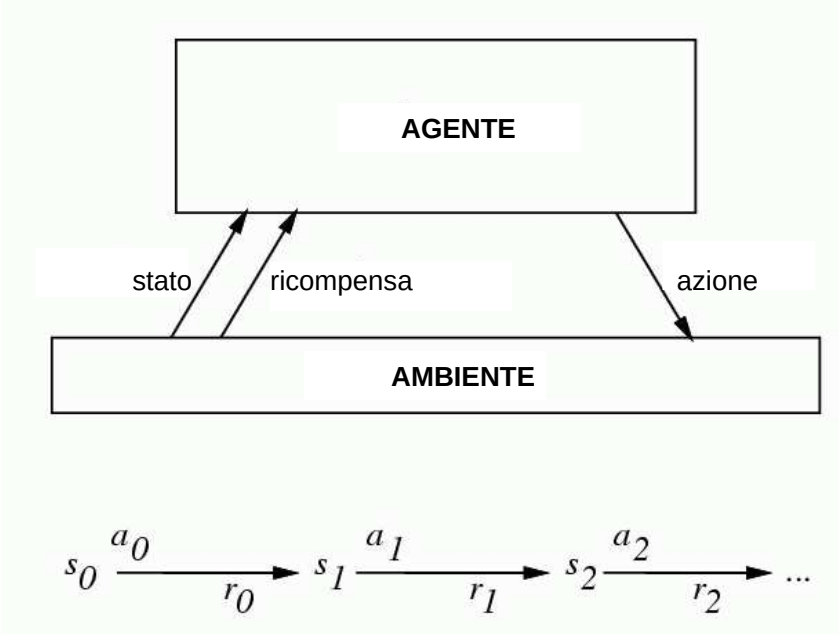


Apprendimento Probabilistico





Apprendimento con rinforzo



Goal: apprendere le azioni che massimizzano

$r_0 + \gamma r_1 + \gamma^2 r_2 + \ldots$, dove $0 \leq \gamma < 1$

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.

Perchè 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: non si può fare a mano;
- mancanza di conoscenza "compilata" rispetto al problema da risolvere;

Ruolo dei Dati

Tipicamente...

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

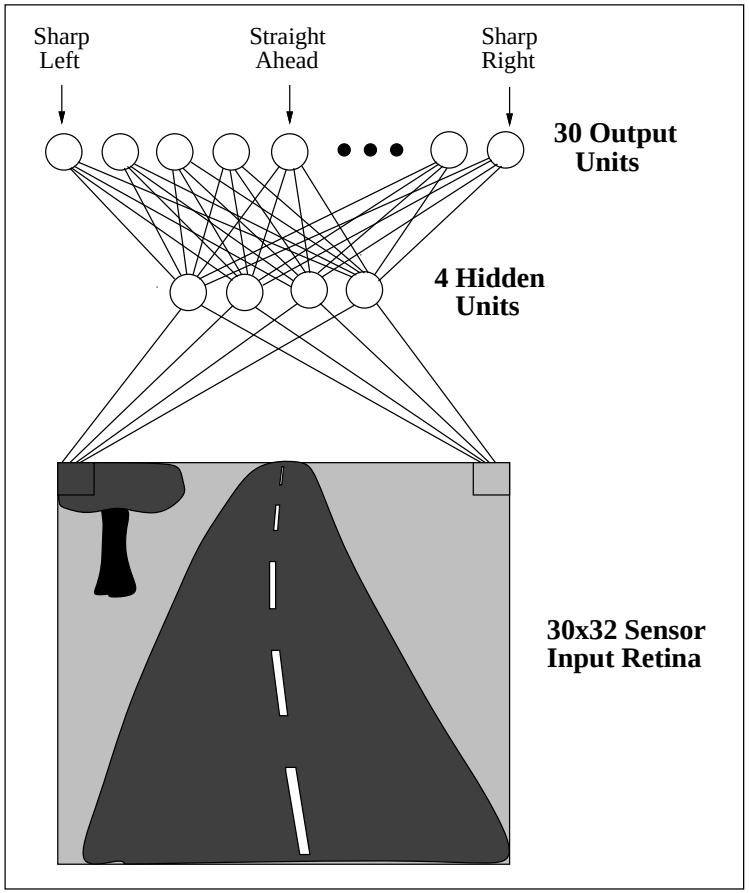
Desiderio: usare i dati per

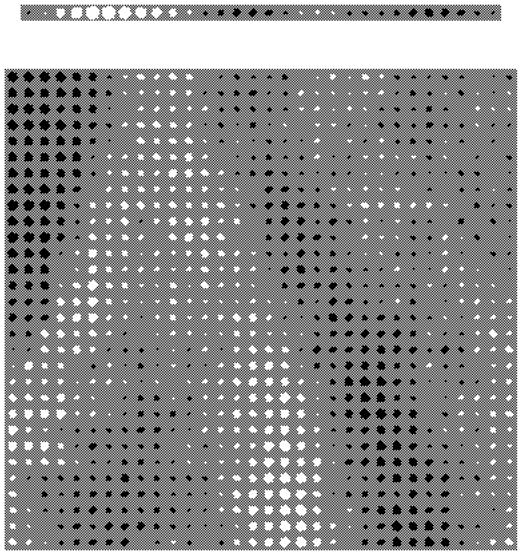
- ottenere nuova conoscenza;
- raffinare la conoscenza di cui si dispone;
- correggere la conoscenza di cui si dispone;

Es. - Riconoscimento di Cifre Manoscritte

- impossibile formalizzare esattamente il problema: disponibili solo esempi;
- possibile presenza di rumore e dati ambigui;

Es. - Guidare una Automobile





Es. - Estrarre Conoscenza Medica dai Dati

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

Age: 23
FirstPregnancy: no
Anemia: no
Diabetes: no
PreviousPrematureBirth: no
Ultrasound: ?
Elective C−Section: ?
Emergency C−Section: ?
...

Age: 23
FirstPregnancy: no
Anemia: no
Diabetes: YES
PreviousPrematureBirth: no
Ultrasound: abnormal
Elective C−Section: no
Emergency C−Section: ?
...

Age: 23
FirstPregnancy: no
Anemia: no
Diabetes: no
PreviousPrematureBirth: no
Ultrasound: ?
Elective C−Section: no
Emergency C−Section: **Yes**
...

Linee di Ricerca all'interno dell' Apprendimento Automatico

- induzione di regole/alberi di decisione,
- algoritmi connessionisti (reti neurali),
- "clustering" & "discovery",
- apprendimento basato sulle istanze
- apprendimento Bayesiano,
- apprendimento basato sulla spiegazione,
- apprendimento con rinforzo,
- apprendimento induttivo guidato dalla conoscenza,
- ragionamento per analogia & basato sui casi,
- algoritmi genetici,
- programmazione logica induttiva, . . .

Principali Paradigmi di Apprendimento

Apprendimento Supervisionato:

- dato in insieme di esempi pre-classificati, $Tr = \{(x^{(i)}, f(x^{(i)}))\}$, apprendere una descrizione generale che incapsula l'informazione contenuta negli esempi (regole valide su tutto il dominio di ingresso)
- tale descrizione deve poter essere usata in modo predittivo (dato un nuovo ingresso $\tilde{x}$ predire l'output associato $f(\tilde{x})$)
- si assume che un esperto (o maestro) ci fornisca la supervisone (cioè i valori della $f()$ per le istanze x dell'insieme di apprendimento)

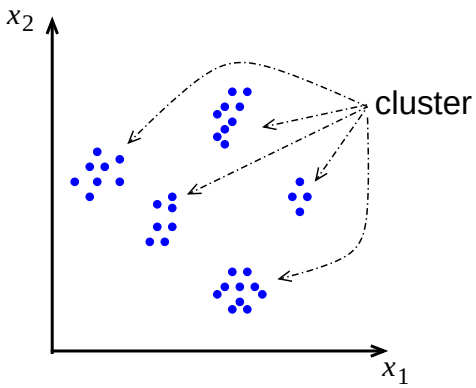
Esempio di applicazione: classificazione di caratteri manoscritti

Principali Paradigmi di Apprendimento

Apprendimento Non-supervisionato:

- dato in insieme di esempi $Tr = \{x^{(i)}\}$, estrarre regolarità e/o pattern (valide(i) su tutto il dominio di ingresso)
- non esiste nessun esperto (o maestro) che ci fornisca un aiuto

  – Clustering



  – Scoperta di Regole (Discovery)

Esempio di applicazione: data mining su database strutturati

Principali Paradigmi di Apprendimento

Apprendimento con Rinforzo:

- Sono dati:
  - agente (intelligente ?), che può
    - trovarsi in uno stato s, ed
    - eseguire una azione a (all'interno delle azioni possibili nello stato corrente)
  - ed opera in un ambiente e, che applicando una azione a nello stato s

    restituisce
    - lo stato successivo, e
    - una ricompensa r, che può essere positiva (+), negativa (-), o neutra (0).
- Scopo dell'agente è quello di massimizzare una funzione delle ricompense

  (es. ricompensa scontata: $\sum_{t=0}^{\infty} \gamma^t r_{t+1}$ dove $0 \leq \gamma < 1$)

Esempio di applicazione: navigare sul Web alla ricerca di informazione focalizzata

Ingredienti Fondamentali Apprendimento Automatico

- Dati di Allenamento (estratti dallo Spazio delle Istanze, X)
- Spazio delle Ipotesi, $\mathcal{H}$
  - 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 $h \in \mathcal{H}$... (selezione di h attraverso i dati di apprendimento)
  - o che almeno una ipotesi $h \in \mathcal{H}$ sia simile a f (approssimazione);
- Algoritmo di Ricerca nello Spazio delle Ipotesi, alg. di apprendimento

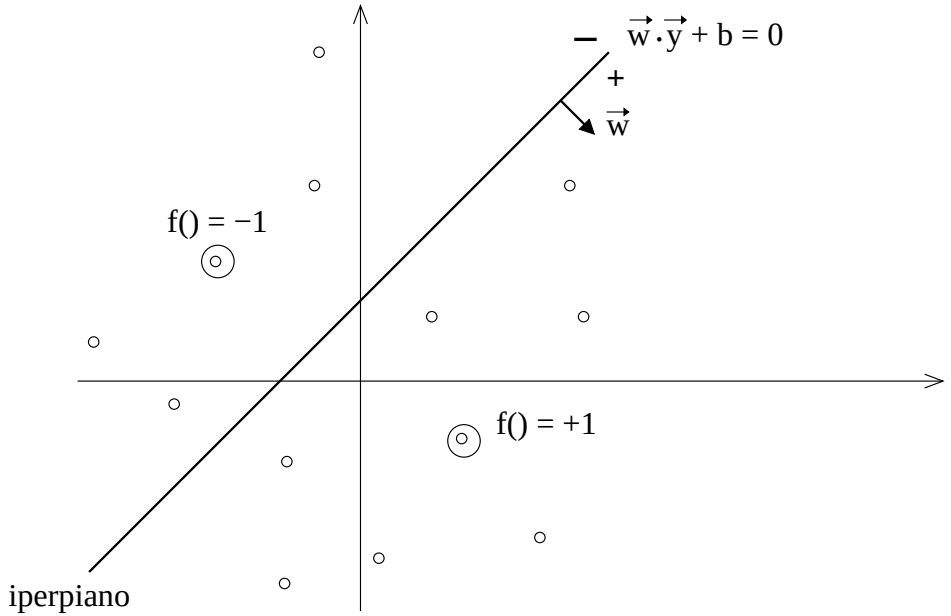
ATTENZIONE: $\mathcal{H}$ non può coincidere con l'insieme di tutte le funzioni possibili e la ricerca essere esaustiva $\rightarrow$ Apprendimento è inutile!!!

Si parla di **Bias Induttivo**: sulla rappresentazione ($\mathcal{H}$) e/o sulla ricerca (alg. di apprendimento)

Spazio delle Ipotesi: Esempio 1

Iperpiani in $\mathbb{R}^2$

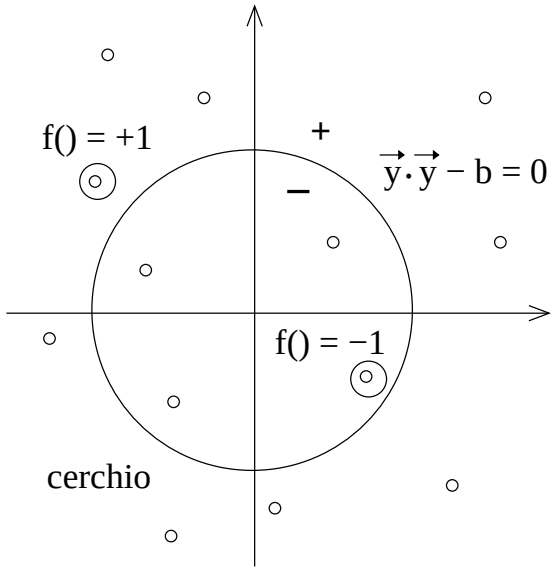
- Spazio delle Istanze $\rightarrow$ punti nel piano: $X = \{\vec{y} \in \mathbb{R}^2\}$
- Spazio delle Ipotesi $\rightarrow$ dicotomie indotte da iperpiani in $\mathbb{R}^2$:
  $\mathcal{H} = \{f_{(\vec{w},b)}(\vec{y}) | f_{(\vec{w},b)}(\vec{y}) = sign(\vec{w} \cdot \vec{y} + b), \vec{w} \in \mathbb{R}^2,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



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 $s3$

Principali Paradigmi di Apprendimento: Richiamo

Apprendimento Supervisionato:

- dato in insieme di esempi pre-classificati, $Tr = \{(x^{(i)}, f(x^{(i)}))\}$, apprendere una descrizione generale che incapsula l'informazione contenuta negli esempi
  (regole valide su tutto il dominio di ingresso)
- tale descrizione deve poter essere usata in modo predittivo
  (dato un nuovo ingresso $\tilde{x}$ predire l'output associato $f(\tilde{x})$)
- si assume che un esperto (o maestro) ci fornisca la supervisone
  (cioè i valori della $f()$ per le istanze x dell'insieme di apprendimento)

Find-S è un algoritmo di apprendimento supervisionato

Dati

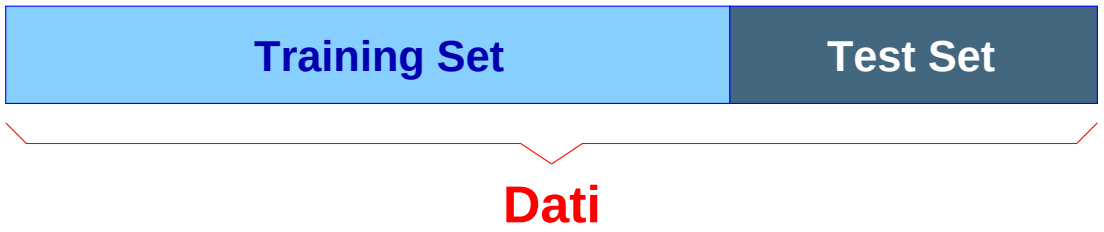
Consideriamo il paradigma di Apprendimento Supervisonato

Dati a nostra disposizione (off-line)

$$\text{Dati} = \{(x^{(1)}, f(x^{(1)})), \ldots, (x^{(N)}, f(x^{(N)}))\}$$

Suddivisione tipica ($N = N_{tr} + N_{ts}$) :

- $\text{Training Set} = \{(x^{(1)}, f(x^{(1)})), \ldots, (x^{(N_{tr})}, f(x^{(N_{tr})}))\}$
  usato direttamente dall'algoritmo di apprendimento;
- $\text{Test Set} = \{(x^{(1)}, f(x^{(1)})), \ldots, (x^{(N_{ts})}, f(x^{(N_{ts})}))\}$
  usato alla fine dell'apprendimento per stimare la bontà della soluzione.

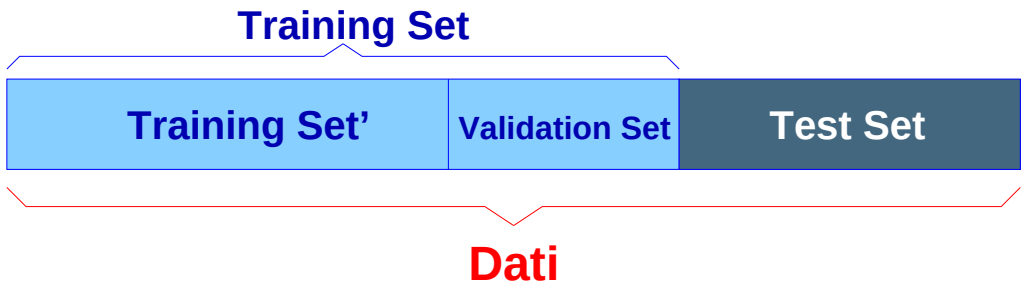


Dati (cont.)

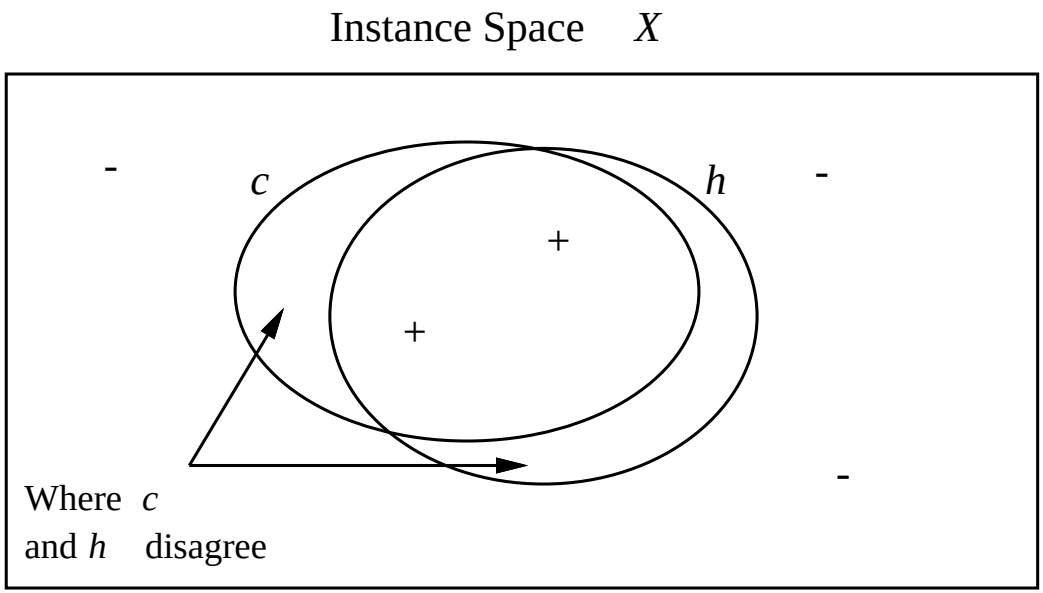
Se N abbastanza grande il Training Set è ulteriormente suddiviso in due sottoinsiemi ($N_{tr} = N_{\widehat{tr}} + N_{val}$) :

- Training Set' $= \{(x^{(1)}, f(x^{(1)})), \ldots, (x^{(N_{\widehat{tr}})}, f(x^{(N_{\widehat{tr}})})\}$
  usato direttamente dall'algoritmo di apprendimento;
- Validation Set $= \{(x^{(1)}, f(x^{(1)})), \ldots, (x^{(N_{val})}, f(x^{(N_{val})})\}$
  usato indirettamente dall'algoritmo di apprendimento.

Il Validation Set serve per scegliere l'ipotesi $h \in \mathcal{H}$ migliore fra quelle consistenti con il Training Set'



Errore Ideale

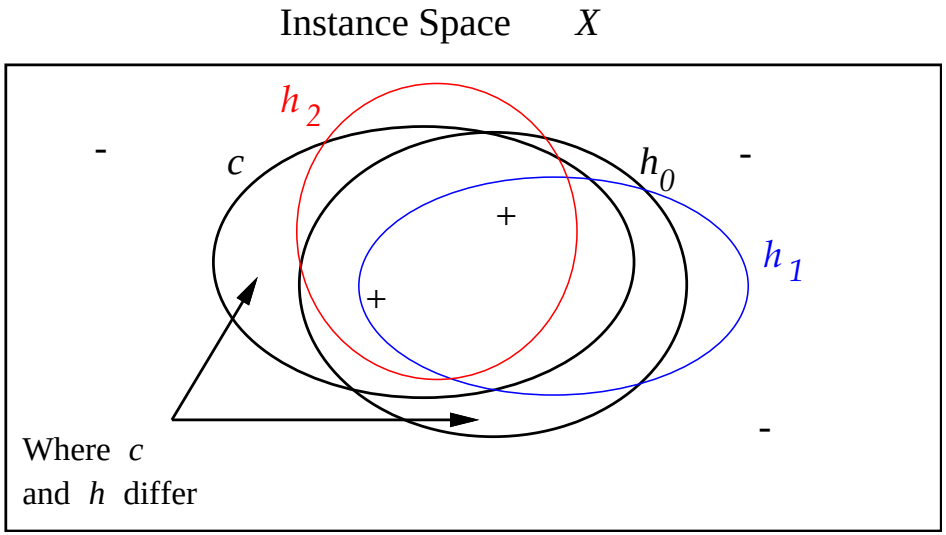


Supponiamo che la funzione f da apprendere sia una funzione booleana (concetto):

$$f : X \to \{0, 1\} \ (\{-, +\})$$

Def: L'**Errore Ideale** ($error_{\mathcal{D}}(h)$) di una ipotesi h rispetto al concetto f e la distribuzione di probabilità $\mathcal{D}$ (probabilità di osservare l'ingresso $x \in X$) è la probabilità che h classifi chi erroneamente un input selezionato a caso secondo $\mathcal{D}$: $error_{\mathcal{D}}(h) \equiv \underset{x \in \mathcal{D}}{Pr}[f(x) \neq h(x)]$

Errore di Apprendimento



Dato Tr = Training Set', più ipotesi possono essere consistenti: h_0, h_1, h_2 quale scegliere ?

Def: L'**Errore Empirico** ($error_{Tr}(h)$) di una ipotesi h rispetto a Tr è il numero di esempi che h classifi ca erroneamente: $error_{Tr}(h) \equiv \#\{(x, f(x)) \in Tr | f(x) \neq h(x)\}$

Def: Una ipotesi $h \in \mathcal{H}$ è sovraspecializzata (**overfit**) Tr se $\exists h' \in \mathcal{H}$ tale che $error_{Tr}(h) < error_{Tr}(h')$, ma $error_{\mathcal{D}}(h) > error_{\mathcal{D}}(h')$.

Il Validation Set serve per cercare di selezionare l'ipotesi migliore (evitare **overfit**).

VC-dimension

Definizione: Frammentazione (Shattering)

Dato $S \subset X$, S è frammentato (shattered) dallo spazio delle ipotesi $\mathcal{H}$ se e solo se

$$\forall S' \subseteq S,\ \exists h \in \mathcal{H},\ \text{tale che } \forall x \in S,\ h(x) = 1 \Leftrightarrow x \in S'$$

($\mathcal{H}$ realizza tutte le possibili dicotomie di S)

Definizione: VC-dimension

La VC-dimension di uno spazio delle ipotesi $\mathcal{H}$ defi nito su uno spazio delle istanze X è data dalla cardinalità del sottoinsieme più grande di X che è frammentato da $\mathcal{H}$:

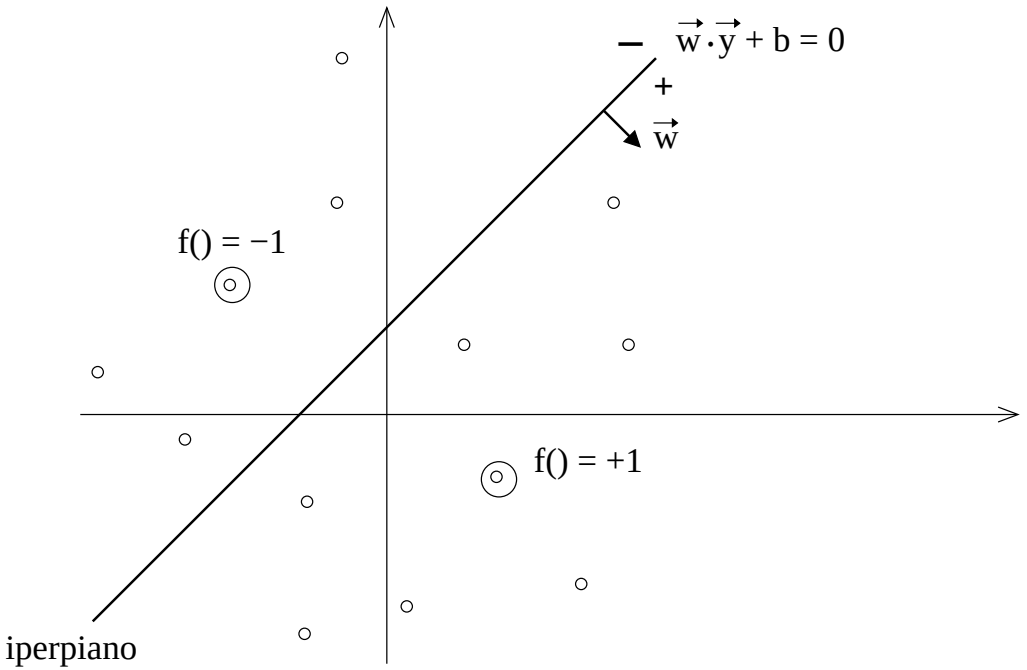
$$VC(\mathcal{H}) = \max_{S \subseteq X} |S| :\ \mathcal{H} \text{ frammenta } S$$

$VC(\mathcal{H}) = \infty$ se S non è limitato

VC-dimension: Esempio

Quale è la VC-dimension di $\mathcal{H}_1$?

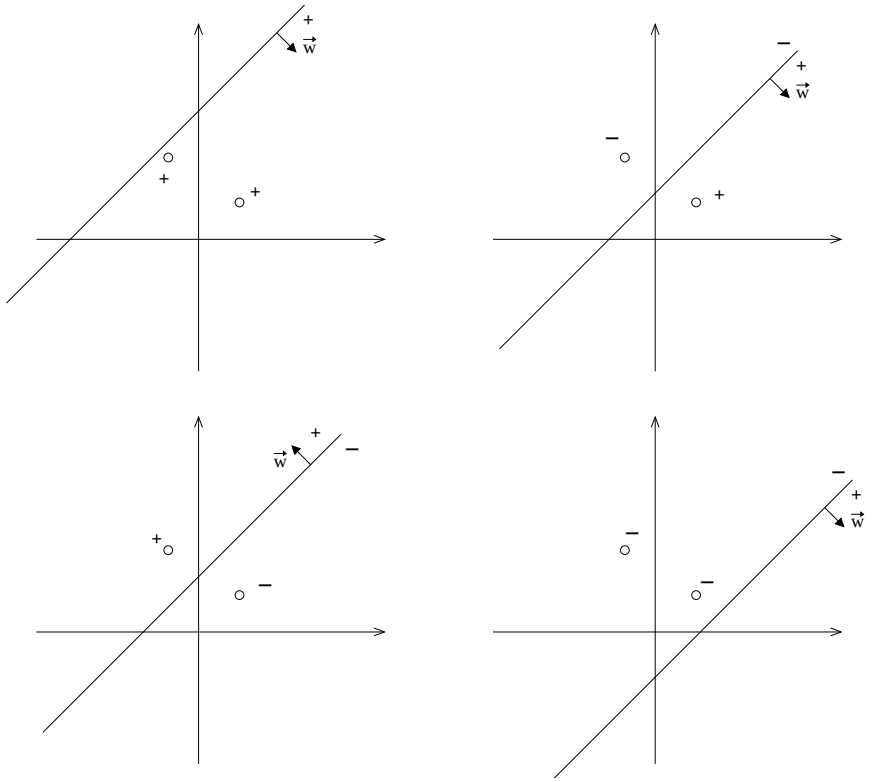
$$\mathcal{H}_1 = \{f_{(\vec{w},b)}(\vec{y}) | f_{(\vec{w},b)}(\vec{y}) = sign(\vec{w} \cdot \vec{y} + b), \vec{w} \in \mathbb{R}^2, b \in \mathbb{R}\}$$



VC-dimension: Esempio

Quale è la VC-dimension di $\mathcal{H}_1$?

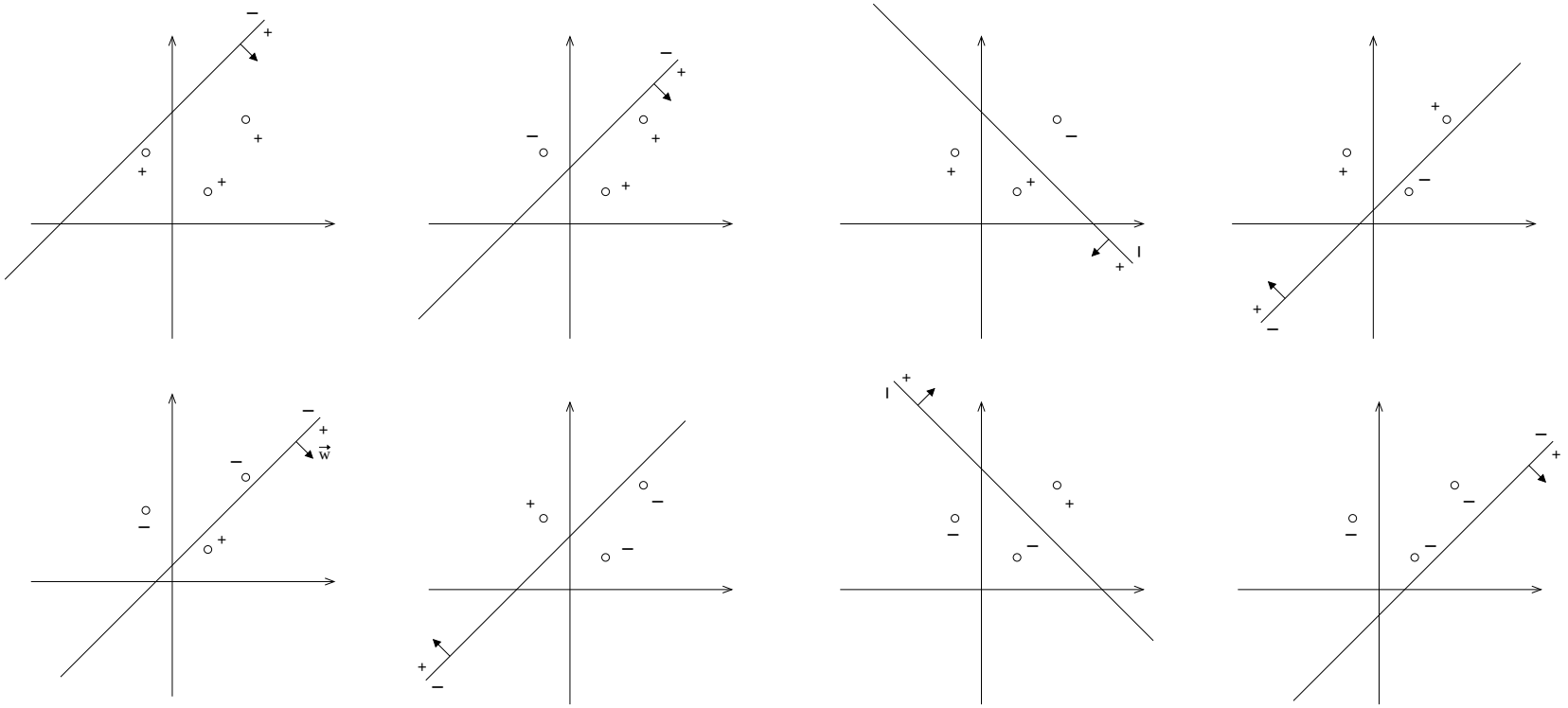
$VC(\mathcal{H}) \geq 1$ banale. Vediamo cosa succede con 2 punti:



VC-dimension: Esempio

Quale è la VC-dimension di $\mathcal{H}_1$?

Quindi $VC(\mathcal{H}) \geq 2$. Vediamo cosa succede con 3 punti:



VC-dimension: Esempio

Quale è la VC-dimension di $\mathcal{H}_1$?

Quindi $VC(\mathcal{H}) \geq 3$. Cosa succede con 4 punti ?

VC-dimension: Esempio

Quale è la VC-dimension di $\mathcal{H}_1$?

Quindi $VC(\mathcal{H}) \geq 3$. Cosa succede con 4 punti ? Non si riesce a frammentare 4 punti!!

Infatti esisteranno sempre due coppie di punti che se unite con un segmento provocano una intersezione fra i due segmenti e quindi, ponendo ogni coppia di punti in classi diverse, per separarli non basta una retta, ma occorre una curva. Quindi $VC(\mathcal{H}) = 3$