

Complessità dello Spazio
delle Ipotesi

Richiamo

Supervised Learning (apprendimento supervisionato)

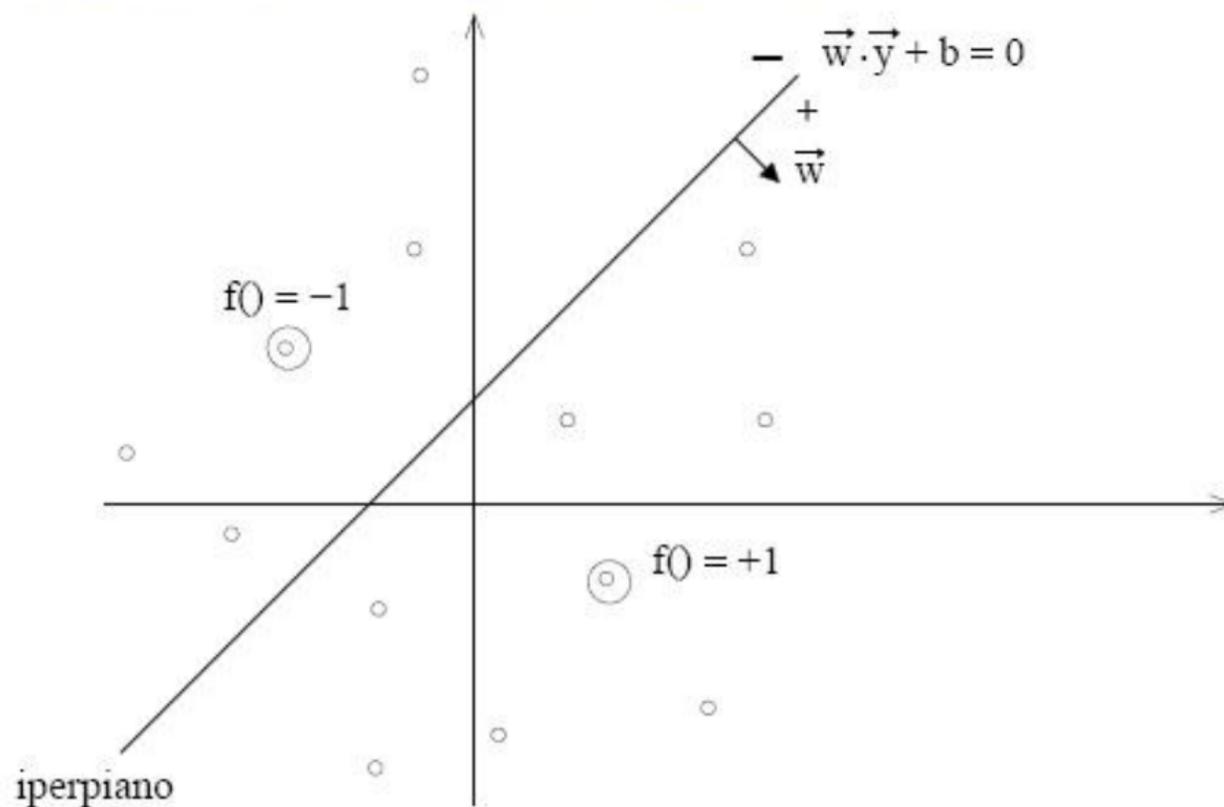
- dato un insieme di dati preclassificati (esempi di apprendimento) apprendere una descrizione generale che incapsula l'informazione contenuta negli esempi
- tale descrizione deve poter essere usata in modo predittivo dato un nuovo ingresso \tilde{x} , predire $f(\tilde{x})$
- si assume che un esperto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

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

Spazio delle Ipotesi: Esempio 1

Iperpiani in \mathbb{R}^2

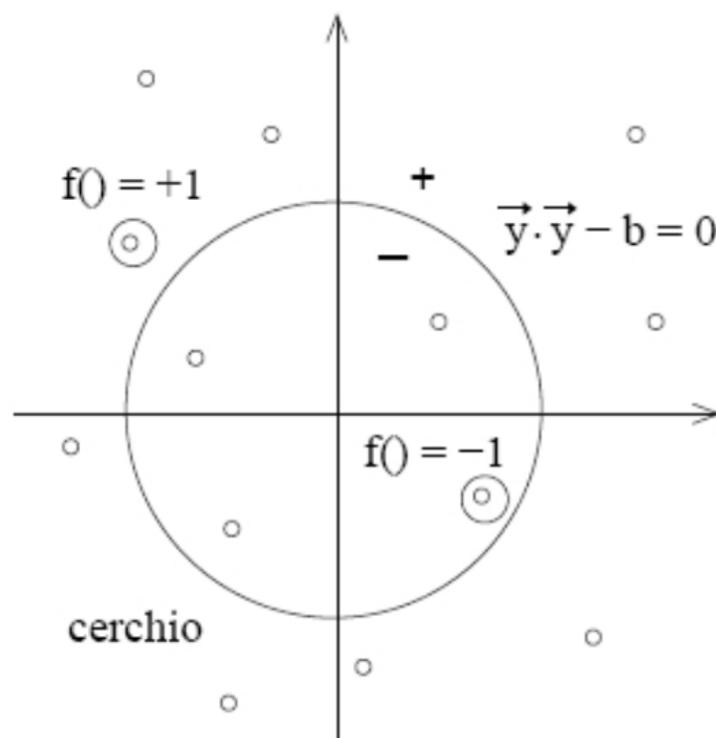
- Spazio delle Istanze \rightarrow punti nel piano: $X = \{\vec{y} \in \mathbb{R}^2\}$
- Spazio delle Ipotesi \rightarrow dicotomie indotte da iperpiani in \mathbb{R}^2 :
 $\mathcal{H} = \{f_{(\vec{w}, b)}(\vec{y}) \mid f_{(\vec{w}, b)}(\vec{y}) = \text{sign}(\vec{w} \cdot \vec{y} + b), \vec{w} \in \mathbb{R}^2,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



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

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

Dati

Consideriamo il paradigma di Apprendimento Supervisionato

Dati a nostra disposizione (off-line)

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

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

- **Training Set** = $\{(x^{(1)}, f(x^{(1)})), \dots, (x^{(N_{tr})}, f(x^{(N_{tr})}))\}$

usato direttamente dall'algoritmo di apprendimento;

- **Test Set** = $\{(x^{(1)}, f(x^{(1)})), \dots, (x^{(N_{ts})}, f(x^{(N_{ts})}))\}$

usato alla fine dell'apprendimento per **stimare** la bontà della soluzione.



Dati

Dati (cont.)

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

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



Dati

evitare l'overfitting!

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

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

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} definito su uno spazio delle istanze X è data dalla cardinalità del sottoinsieme più grande di X che è frammentato da \mathcal{H} :

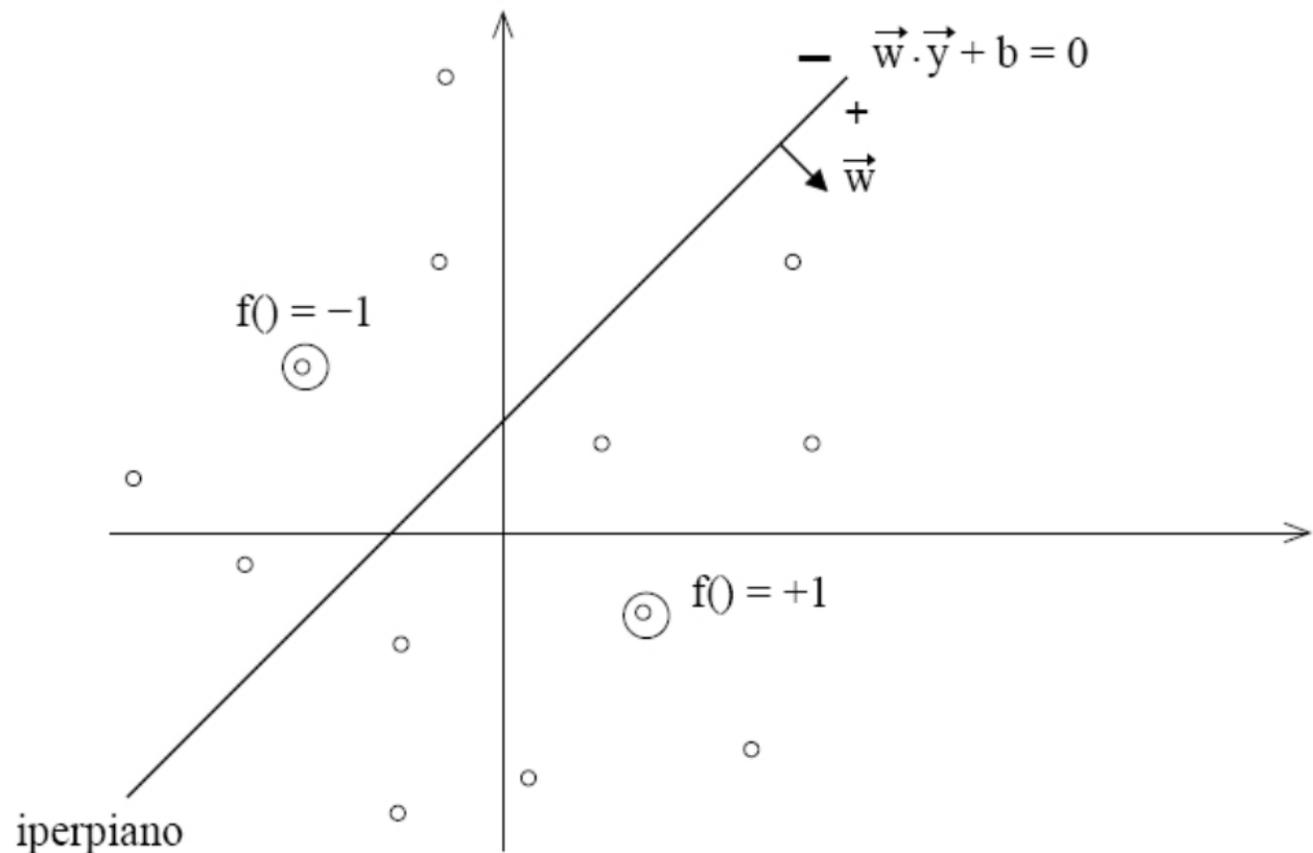
$$VC(\mathcal{H}) = \max_{S \subseteq X} |S| : \mathcal{H} \text{ frammenta } S$$

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

VC-dimension: Esempio

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

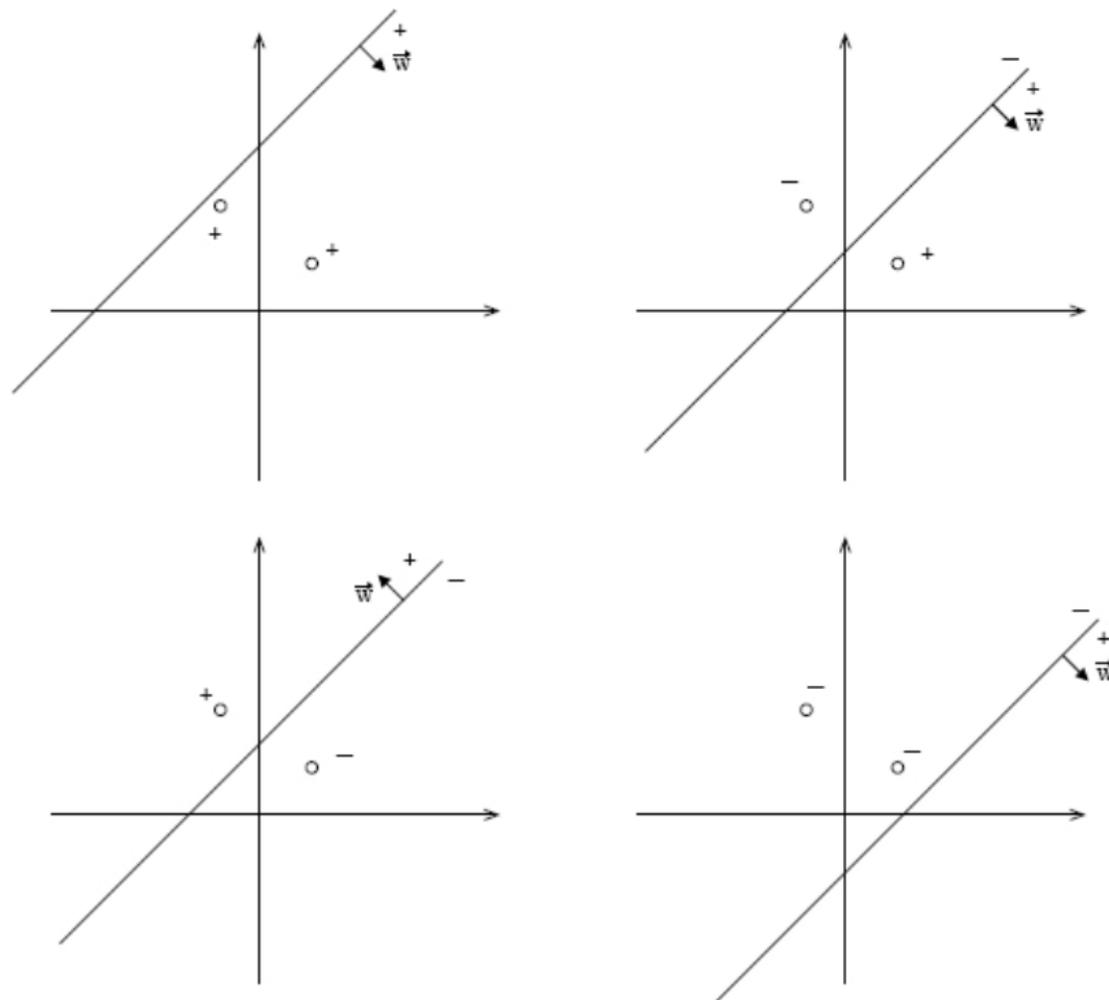
$$\mathcal{H}_1 = \{f_{(\vec{w}, b)}(\vec{y}) \mid f_{(\vec{w}, b)}(\vec{y}) = \text{sign}(\vec{w} \cdot \vec{y} + b), \vec{w} \in \mathbb{R}^2, b \in \mathbb{R}\}$$



VC-dimension: Esempio

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

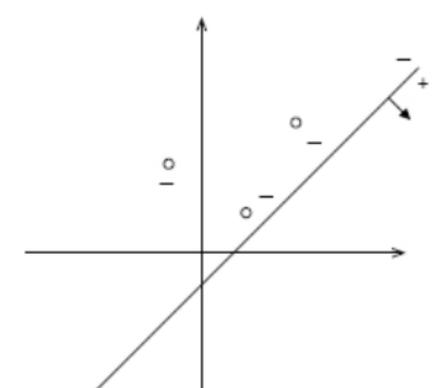
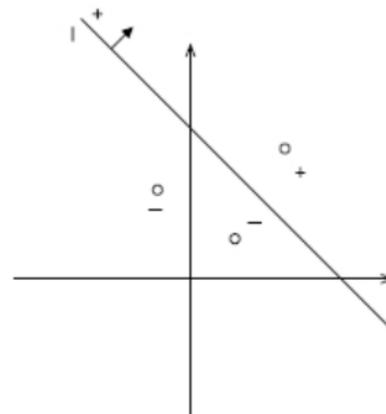
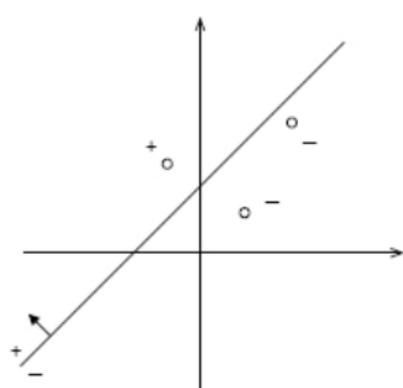
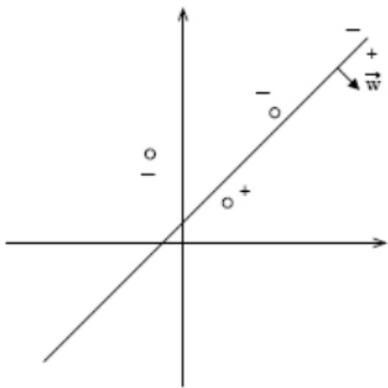
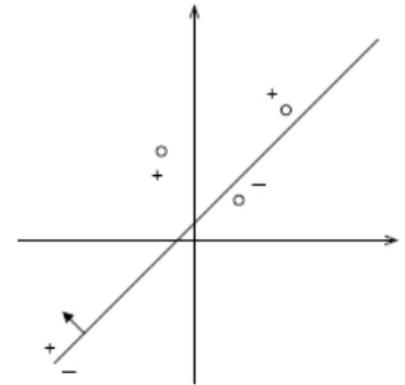
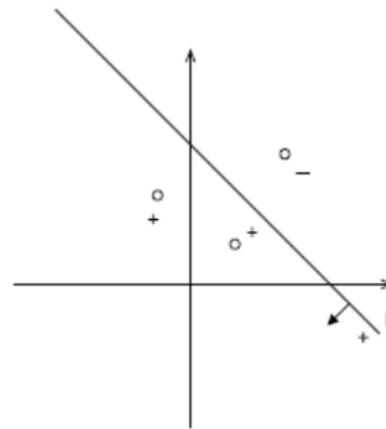
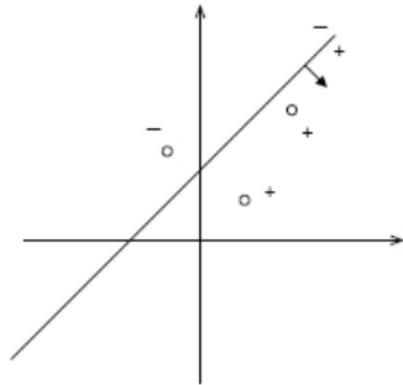
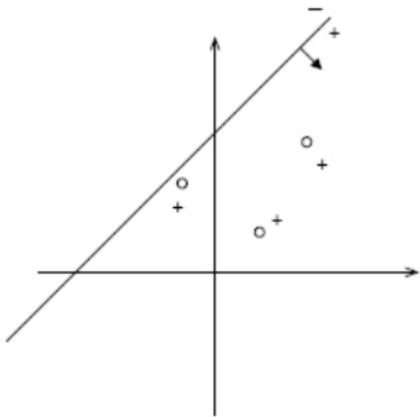
$VC(\mathcal{H}) \geq 1$ banale. Vediamo cosa succede con 2 punti:



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 ?

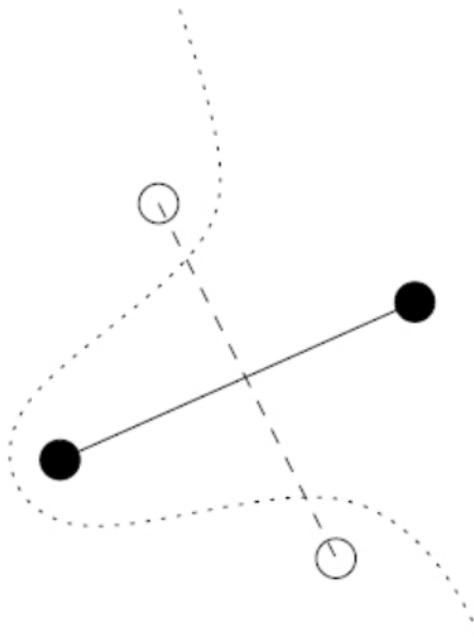


VC-dimension: Esempio

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 $VC(\mathcal{H}) = 3$



Bound sull'Errore Ideale per Classificazione Binaria

Consideriamo un problema di classificazione binario (i.e., apprendimento di concetti). Dati

- Training Set $Tr = \{(\mathbf{x}^{(1)}, f(\mathbf{x}^{(1)})), \dots, (\mathbf{x}^{(N_{tr})}, f(\mathbf{x}^{(N_{tr})}))\}$
- Spazio delle Ipotesi $\mathcal{H} = \{h_{\mathbf{w}}(\mathbf{x}) | \mathbf{w} \in \mathbb{R}^k\}$
- Algoritmo di Apprendimento L che restituisce l'ipotesi $h_{\mathbf{w}^*}(\mathbf{x})$, dove \mathbf{w}^* minimizza l'errore empirico $error_{Tr}(h_{\mathbf{w}}(\mathbf{x}))$

è possibile derivare dei bound sull'errore ideale (detto anche errore di generalizzazione), validi con probabilità $1 - \delta$, che hanno una forma del tipo

$$error_{\mathcal{D}}(h_{\mathbf{w}^*}(\mathbf{x})) \leq error_{Tr}(h_{\mathbf{w}^*}(\mathbf{x})) + \epsilon(N_{tr}, VC(\mathcal{H}), \delta)$$

Esempio:

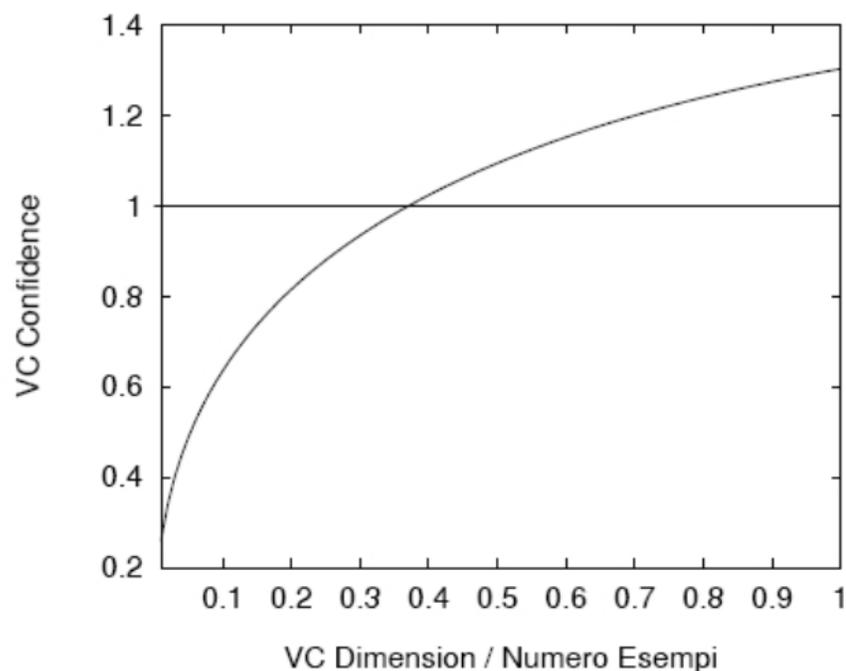
$$error_{\mathcal{D}}(h_{\mathbf{w}^*}(\mathbf{x})) \leq \underbrace{error_{Tr}(h_{\mathbf{w}^*}(\mathbf{x}))}_A + \underbrace{\sqrt{\frac{VC(\mathcal{H})}{N_{tr}} (\log(\frac{2N_{tr}}{VC(\mathcal{H})}) + 1) - \frac{1}{N_{tr}} \log(\delta)}}_B$$

Bound sull'Errore Ideale per Classificazione Binaria

Si noti che

- il termine **A** DIPENDE SOLO dalla ipotesi restituita dall'algoritmo di apprendimento L ;
- il termine **B** è INDIPENDENTE dalla ipotesi restituita dall'algoritmo di apprendimento L ;
in particolare dipende dal rapporto fra VC-dimension dello spazio delle ipotesi \mathcal{H} e il numero di esempi di apprendimento (N_{tr}), oltre ovviamente che dalla confidenza $(1 - \delta)$ con cui il bound è valido.

Il termine **B** è usualmente chiamato VC-confidence e risulta essere monotono rispetto al rapporto $\frac{VC(\mathcal{H})}{N_{tr}}$; fissato N_{tr} aumenta all'aumentare di $VC(\mathcal{H})$.



Structural Risk Minimization

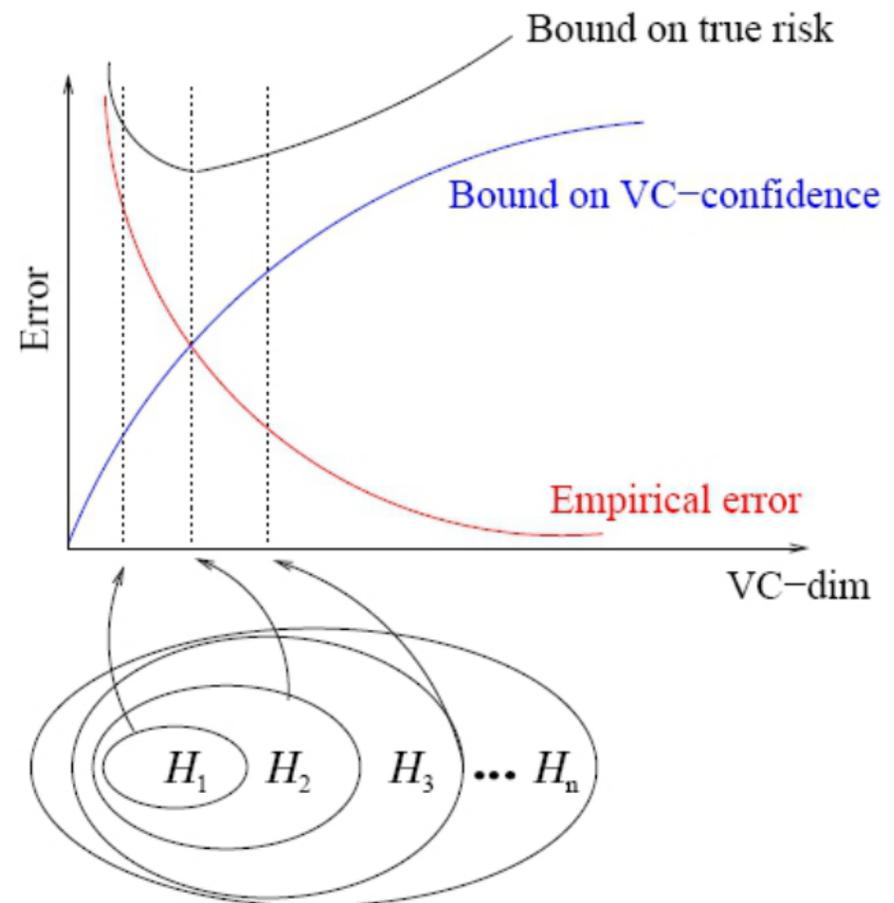
Problema: all'aumentare della VC-dimension diminuisce l'errore empirico (termine A), ma aumenta la VC confidence (termine B)!

L'approccio **Structural Risk Minimization** tenta di trovare un compromesso tra i due termini:

Si considerano \mathcal{H}_i tali che

- $\mathcal{H}_1 \subseteq \mathcal{H}_2 \subseteq \dots \subseteq \mathcal{H}_n$
- $VC(\mathcal{H}_1) \leq \dots \leq VC(\mathcal{H}_n)$
- si seleziona l'ipotesi che ha il bound sull'errore ideale più basso

Esempio: Reti neurali con un numero crescente di neuroni nascosti



Se la cardinalità dello spazio delle ipotesi è finito

$$VC(\mathcal{H}) \leq \log_2(|\mathcal{H}|)$$

$$\mathcal{H}) \leq \log_2(|\mathcal{H}|)$$

Se la cardinalità dello spazio delle ipotesi è finito

$$VC(\mathcal{H}) \leq \log_2(|\mathcal{H}|)$$

Mostriamo che $VC(\mathcal{H}) \leq \log_2(|\mathcal{H}|)$

- per ogni S tale che \mathcal{H} frammenta S abbiamo $|\mathcal{H}| \geq 2^{|S|}$, infatti \mathcal{H} può implementare tutte le possibili dicotomie di S , che sono esattamente $2^{|S|}$.
- scegliendo un S tale che $|S| = VC(\mathcal{H})$, otteniamo $|\mathcal{H}| \geq 2^{VC(\mathcal{H})}$

Quindi, applicando \log_2 ad entrambi i lati della disuguaglianza, possiamo concludere che $\log_2(|\mathcal{H}|) \geq VC(\mathcal{H})$