(F94) Metodi statistici per l'apprendimento

Rischio statistico e sua analisi

Docente: *Nicolò Cesa-Bianchi* versione 7 aprile 2018

Per analizzare un algoritmo di apprendimento dobbiamo costruire un modello formale di generazione dei dati etichettati, ovvero un modello di generazione degli esempi $(\boldsymbol{x}, y)$. Nel modello statistico di apprendimento assumiamo che ogni esempio $(\boldsymbol{x}, y)$ sia ottenuto tramite un'estrazione indipendente da una distribuzione di probabilità su $\mathcal{X} \times \mathcal{Y}$ fissata ma ignota. Per evidenziare il fatto che $\boldsymbol{x}$ e y sono variabili casuali spesso scriveremo $(\boldsymbol{X}, Y)$. È abbastanza plausibile modellare con una distribuzione di probabilità il fatto che non tutti i dati $\boldsymbol{x}$ sono osservati con la stessa frequenza (pensate al caso in cui $\boldsymbol{x}$ è un'immagine o un testo). Analogamente, dato che come abbiamo detto le etichette sono per varie ragioni affette da rumore, è pure plausibile modellare con una distribuzione di probabilità la varietà di etichette a cui un dato può essere associato.

Dato che nel modello statistico ogni esempio $(\boldsymbol{X}, Y)$ è ottenuto tramite un'estrazione indipendente dalla stessa distribuzione di probabilità congiunta, ogni dataset (p.es., training set o test set) sarà un **campione casuale** nel senso statistico del termine. In generale, l'assunzione di indipendenza è comoda dal punto di vista della trattabilità analitica del problema, ma poco plausibile in realtà. Si pensi ad esempio al problema di categorizzare le notizie di prodotte da un'agenzia giornalistica. È ovvio che, queste non saranno statisticamente indipendenti, ma seguiranno i diversi temi che di volta in volta si alterneranno.

Le prestazioni di un predittore $h : \mathcal{X} \to \mathcal{Y}$ rispetto ad un modello statistico dato ed a una funzione di perdita $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ data vengono valutate con il **rischio statistico**, definito da

$$\mathrm{er}(h) = \mathbb{E}\big[\ell(Y, h(\boldsymbol{X}))\big]$$

ovvero, il valore atteso della funzione di perdita rispetto ad un esempio $(\boldsymbol{X}, Y)$ generato dal modello statistico. Ipotizzando di conoscere il modello statistico, possiamo costruire il **predittore Bayesiano ottimo** $f^* : \mathcal{X} \to \mathcal{Y}$. Esso è definito come

$$f^*(\boldsymbol{x}) = \operatorname*{argmin}_{\widehat{y} \in \mathcal{Y}} \mathbb{E}\big[\ell(Y, \widehat{y}) \,\big|\, \boldsymbol{X} = \boldsymbol{x}\big]$$

ovvero la predizione $\widehat{y}$ che minimizza il rischio condizionato, cioè la perdita attesa rispetto alla distribuzione di Y condizionata su $\boldsymbol{X} = \boldsymbol{x}$. Si noti che, per definizione di f^*, vale

$$\mathbb{E}\Big[\ell\big(Y, f^*(x)\big) \,\Big|\, \boldsymbol{X} = \boldsymbol{x}\Big] \leq \mathbb{E}\Big[\ell\big(Y, h(x)\big) \,\Big|\, \boldsymbol{X} = \boldsymbol{x}\Big]$$

per ogni classificatore $h : \mathcal{X} \to \mathcal{Y}$. Dato che la disuguaglianza sopra vale per ogni $\boldsymbol{x} \in \mathcal{X}$, vale anche in media rispetto all'estrazione di $\boldsymbol{X}$. Ma siccome la media del rischio condizionato coincide col rischio, abbiamo che $\mathrm{er}(f^*) \leq \mathrm{er}(h)$. Il rischio $\mathrm{er}(f^*)$ del predittore Bayesiano ottimo è detto **Bayes error**. In generale il Bayes error è maggiore di zero in quanto i predittori sono deterministici mentre le etichette sono probabilistiche.

Iniziamo ora a calcolare il predittore Bayesiano ottimo per la funzione di perdita quadratica $\ell(y, \widehat{y}) = (y - \widehat{y})^2$ nel caso $\mathcal{Y} \equiv \mathbb{R}$,

$$\begin{aligned}
f^*(\boldsymbol{x}) &= \operatorname*{argmin}_{\widehat{y} \in \mathbb{R}} \mathbb{E}\big[(Y - \widehat{y})^2 \,\big|\, \boldsymbol{X} = \boldsymbol{x}\big] \\
&= \operatorname*{argmin}_{\widehat{y} \in \mathbb{R}} \Big(\mathbb{E}\big[Y^2 \,\big|\, \boldsymbol{X} = \boldsymbol{x}\big] + \widehat{y}^2 - 2\widehat{y}\,\mathbb{E}\big[Y \,\big|\, \boldsymbol{X} = \boldsymbol{x}\big] \Big) \\
&= \operatorname*{argmin}_{\widehat{y} \in \mathbb{R}} \Big(\widehat{y}^2 - 2\widehat{y}\,\mathbb{E}\big[Y \,\big|\, \boldsymbol{X} = \boldsymbol{x}\big] \Big) \qquad \text{(scartando il termine che non dipende da } \widehat{y}) \\
&= \mathbb{E}\big[Y \,\big|\, \boldsymbol{X} = \boldsymbol{x}\big] \qquad \text{(minimizzando la funzione } F(\widehat{y}) = \widehat{y}^2 - 2\widehat{y}\,\mathbb{E}\big[Y \,\big|\, \boldsymbol{X} = \boldsymbol{x}\big].)
\end{aligned}$$

Ovvero, il predittore Bayesiano ottimo per la funzione di perdita quadratica è il valore atteso dell'etichetta condizionato sull'istanza.

Sostituendo nella formula del rischio condizionato $\mathbb{E}\big[\big(Y - f^*(\boldsymbol{X})\big)^2 \,\big|\, \boldsymbol{X} = \boldsymbol{x}\big]$ il predittore ottimo $f^*(\boldsymbol{x}) = \mathbb{E}[Y \mid \boldsymbol{X} = \boldsymbol{x}]$ otteniamo

$$\mathbb{E}\Big[\big(Y - f^*(\boldsymbol{X})\big)^2 \,\Big|\, \boldsymbol{X} = \boldsymbol{x}\Big] = \mathbb{E}\Big[\big(Y - \mathbb{E}[Y \mid \boldsymbol{x}]\big)^2 \,\Big|\, \boldsymbol{X} = \boldsymbol{x}\Big] = \operatorname{Var}\big[Y \,\big|\, \boldsymbol{X} = \boldsymbol{x}\big] \ .$$

Ovvero il rischio condizionato del predittore Bayesiano ottimo per la perdita quadrata è la varianza dell'etichetta condizionata sull'istanza.

Focalizziamoci ora sul caso di classificazione binaria, dove $\mathcal{Y} = \{-1, +1\}$. In questo caso è conveniente specificare una distribuzione congiunta sugli esempi $(\boldsymbol{X}, Y)$ con la coppia (D, η), dove D è la marginale su $\mathcal{X}$ e η rappresenta la distribuzione su $\{-1, +1\}$ condizionata su $\mathcal{X}$. Dato che η è una distribuzione concentrata su due valori, possiamo rappresentarla con la funzione $\eta : \mathcal{X} \to [0, 1]$ dove $\eta(\boldsymbol{x}) = \mathbb{P}\big(Y = +1 \mid \boldsymbol{X} = \boldsymbol{x}\big)$ è la probabilità che $\boldsymbol{x}$ assuma l'etichetta $+1$.

Sia $\mathbb{I}\{A\} \in \{0, 1\}$ la funzione indicatrice di un evento A: $\mathbb{I}\{A\} = 1$ se e solo se A è vero. Il rischio statistico rispetto alla funzione di perdita zero-uno $\ell(y, \widehat{y}) = \mathbb{I}\{\widehat{y} \neq y\}$ risulta quindi essere

$$\operatorname{er}(h) = \mathbb{E}\big[\ell(Y, h(\boldsymbol{X}))\big] = \mathbb{E}\big[\mathbb{I}\{h(\boldsymbol{X}) \neq Y\}\big] = \mathbb{P}\big(h(\boldsymbol{X}) \neq Y\big) \ .$$

In altre parole, il rischio di h è la probabilità che h sbagli a classificare un $\boldsymbol{X}$ estratto dalla distribuzione D su $\mathcal{X}$ e avente etichetta Y estratta a caso dalla distribuzione $\{1 - \eta(\boldsymbol{X}), \eta(\boldsymbol{X})\}$ su $\{-1, +1\}$.

Ipotizzando di conoscere il modello statistico (D, η), possiamo costruire il classificatore Bayesiano ottimo $f^* : \mathcal{X} \to \{-1, +1\}$. Allora

$$\begin{aligned}
f^*(\boldsymbol{x}) &= \operatorname*{argmin}_{\widehat{y} \in \{-1,+1\}} \mathbb{E}\big[\ell(Y, \widehat{y}) \,\big|\, \boldsymbol{X} = \boldsymbol{x}\big] \\
&= \operatorname*{argmin}_{\widehat{y} \in \{-1,+1\}} \mathbb{E}\big[\mathbb{I}\{Y = +1\}\mathbb{I}\{\widehat{y} = -1\} + \mathbb{I}\{Y = -1\}\mathbb{I}\{\widehat{y} = +1\} \,\big|\, \boldsymbol{X} = \boldsymbol{x}\big] \\
&= \operatorname*{argmin}_{\widehat{y} \in \{-1,+1\}} \Big(\mathbb{P}(Y = +1 \mid \boldsymbol{X} = \boldsymbol{x})\mathbb{I}\{\widehat{y} = -1\} + \mathbb{P}(Y = -1 \mid \boldsymbol{X} = \boldsymbol{x})\mathbb{I}\{\widehat{y} = +1\} \Big) \\
&= \operatorname*{argmin}_{\widehat{y} \in \{-1,+1\}} \Big(\eta(\boldsymbol{x})\mathbb{I}\{\widehat{y} = -1\} + \big(1 - \eta(\boldsymbol{x})\big)\mathbb{I}\{\widehat{y} = +1\} \Big) \\
&= \begin{cases} -1 & \text{se } \eta(\boldsymbol{x}) < 1/2, \\ +1 & \text{se } \eta(\boldsymbol{x}) \geq 1/2. \end{cases}
\end{aligned}$$

Quindi il classificatore Bayesiano ottimo predice l'etichetta che massimizza la probabilità condizionata sull'istanza. Non è difficile verificare che il Bayes error è pari a $\mathrm{er}(f^*) = \mathbb{E}\big[\min\{\eta(\boldsymbol{X}),\, 1 - \eta(\boldsymbol{X})\}\big]$.

Stima del rischio per classificazione binaria. Passiamo ora a capire come valutare il rischio di un algoritmo di apprendimento. Per interpretare il modello statistico, possiamo pensare che ogni esempio $(\boldsymbol{x}, y)$ disponibile sia ottenuto attraverso un'estrazione indipendente di $\boldsymbol{x} \in \mathcal{X}$ secondo la distribuzione D seguito dall'attribuzione dell'etichetta $y \in \{-1, +1\}$ secondo la distribuzione $\big\{1 - q(\boldsymbol{x}), q(\boldsymbol{x})\big\}$.

Dovrebbe esssere chiaro che, dato un qualunque classificatore h, non possiamo calcolare direttamente il suo rischio $\mathrm{er}(h)$ rispetto ad un modello statistico (D, η). Infatti, il calcolo di $\mathrm{er}(h)$ richiede di conoscere esattamente D e η. Ma se conoscessimo η potremmo direttamente trovare il classificatore Bayesiano ottimo per il problema.

Consideriamo ora il problema di stimare il rischio di un dato classificatore h. Per calcolare questa stima si usa un cosiddetto **test set**, ovvero un insieme $(\boldsymbol{x}'_1, y'_1), \ldots, (\boldsymbol{x}'_n, y'_n)$ di dati etichettati. Quindi, si stima $\mathrm{er}_{D,\eta}(h)$ con il **test error**, ovvero la frazione degli esempi del test set scorrettamente classificati da h,

$$\widetilde{\mathrm{er}}(h) = \frac{1}{n} \sum_{t=1}^{n} \ell\big(y'_t, h(\boldsymbol{x}'_t)\big) \ .$$

Sotto l'ipotesi che il test set sia stato generato mediante estrazioni indipendenti dal modello statistico (D, η), il test error altro non è che la media campionaria del rischio in quanto, per ogni $t = 1, \ldots, n$ abbiamo che $(\boldsymbol{X}'_t, Y'_t)$ è un'estrazione indipendente da (D, η). Quindi

$$\mathbb{E}\Big[\ell\big(Y'_t, h(\boldsymbol{X}'_t)\big)\Big] = \mathbb{P}\big(h(\boldsymbol{X}'_t) \neq Y'_t\big) = \mathrm{er}(h) \ .$$

Per valutare la precisione con la quale possiamo vedere il test error come stima del rischio per una data taglia di test set utilizziamo il risultato seguente legato alla legge dei grandi numeri.

Lemma 1 (Chernoff-Hoeffding) *Siano $Z_1, \ldots, Z_n$ variabili casuali indipendenti, identicamente distribuite con media μ, e tali che $0 \leq Z_t \leq 1$ per ogni $t = 1, \ldots, n$. Allora, per ogni $\varepsilon > 0$ fissato,*

$$\mathbb{P}\left(\frac{1}{n}\sum_{t=1}^{n} Z_t > \mu + \varepsilon\right) \leq e^{-2\varepsilon^2 n} \qquad e \qquad \mathbb{P}\left(\frac{1}{n}\sum_{t=1}^{n} Z_t < \mu - \varepsilon\right) \leq e^{-2\varepsilon^2 n} \ .$$

Utilizzando il maggiorante di Chernoff-Hoeffding con $Z_t = \ell(y_t, h(x_t))$ dove ℓ è la funzione di perdita per classificazione binaria, possiamo quindi valutare la precisione della nostra stima come segue

$$\mathbb{P}\Big(\big|\mathrm{er}(h) - \widetilde{\mathrm{er}}(h)\big| > \varepsilon\Big) = \mathbb{P}\Big(\mathrm{er}(h) - \widetilde{\mathrm{er}}(h) > \varepsilon\Big) + \mathbb{P}\Big(\widetilde{\mathrm{er}}(h) - \mathrm{er}(h) > \varepsilon\Big) \leq 2\, e^{-2\varepsilon^2 n} \qquad (1)$$

dove la probabilità è calcolata rispetto all'estrazione del test set. Questa disuguaglianza ci dice che la misura (secondo D e η) dei test set che producono stime $\widetilde{\mathrm{er}}(h)$ che differiscono dal valor medio $\mathrm{er}(h)$ per più di ε descresce rapidamente con l'aumentare di n, cioè del numero di esempi nel test set.

In particolare, ponendo la parte destra di (1) pari a δ, per un qualunque $\delta \in (0,1)$ fissato, otteniamo che

$$\left|\mathrm{er}(h) - \widetilde{\mathrm{er}}(h)\right| \leq \sqrt{\frac{1}{2n} \ln \frac{2}{\delta}}$$

vale con probabilità almeno $1 - \delta$ rispetto all'estrazione del test set.

La disuguaglianza (1) ci indica come stimare il rischio di un'ipotesi prodotta da un qualche algoritmo di apprendimento mediante un test set. Viceversa, la stessa disuguaglianza ci mostra come il test set, che è il modo con cui in pratica misuriamo le prestazioni di un classificatore su dati ignoti, sia ben correlato col rischio nel modello di apprendimento statistico.

Studiamo ora un algoritmo di apprendimento nel modello statistico che abbiamo appena delineato e verifichiamo se compare l'overfitting. Come di consueto, assumiamo che l'algoritmo riceva in ingresso un training set $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_m, y_m)$, dove $(\boldsymbol{x}_t, y_t) \in \mathcal{X} \times \{-1, +1\}$, e generi un classificatore $h : \mathcal{X} \to \{-1, +1\}$ appartenente a un dato spazio di modelli $\mathcal{H}$.

Sotto queste ipotesi, la cosa migliore che può fare l'algoritmo è scegliere il miglior classificatore possibile $h \in \mathcal{H}$, ovvero il classificatore h^* tale che

$$\mathrm{er}(h^*) = \min_{h \in \mathcal{H}} \mathrm{er}(h) \ .$$

Grazie alla legge dei grandi numeri, sappiamo che il training error $\widehat{\mathrm{er}}(h^*)$ è vicino a $\mathrm{er}(h^*)$ con alta probabilità rispetto all'estrazione del training set su cui $\widehat{\mathrm{er}}(h^*)$ viene calcolato.

Consideriamo l'algoritmo che sceglie $\widehat{h} \in \mathcal{H}$ in modo da minimizzare il training error, ovvero

$$\widehat{h} = \operatorname*{argmin}_{h \in \mathcal{H}} \widehat{\mathrm{er}}(h) \ .$$

Putroppo non possiamo applicare direttamente a $\widehat{h}$ il maggiorante di Chernoff-Hoeffding al fine di dimostrare che $\mathrm{er}(\widehat{h})$ è vicino a $\widehat{\mathrm{er}}(\widehat{h})$, per poi concludere che $\mathrm{er}(\widehat{h})$ è vicino a $\mathrm{er}(h^*)$. Il motivo è che $\widehat{h}$ è una funzione del training set e quindi una variabile casuale. Chernoff-Hoeffding dice che $\widehat{\mathrm{er}}(h)$ è vicino a $\mathrm{er}(h)$ per ogni h fissato, mentre $\widehat{h}$ non è fissato, ma dipende dal campione rispetto al quale calcoliamo la probabilità.

Per analizzare il rischio di $\widehat{h}$ procediamo allora come segue:

$$\begin{aligned} \mathrm{er}(\widehat{h}) = \mathrm{er}(\widehat{h}) - \mathrm{er}(h^*) \qquad & \text{errore di “varianza”} \\ + \mathrm{er}(h^*) - \mathrm{er}(f^*) \qquad & \text{errore di “bias”} \\ + \mathrm{er}(f^*) \qquad & \text{Bayes error} \end{aligned}$$

dove f^* è il classificatore Bayesiano ottimo per il modello statistico (D, η) soggiacente. Il Bayes error non è controllabile, dato che dipende unicamente dal modello (D, η). L'errore di bias dipende dal fatto che $\mathcal{H}$ può non contenere il classificatore Bayesiano ottimo. L'errore di varianza dipende dal fatto che $\mathrm{er}(\widehat{h})$ è generalmente diverso dall'errore di $\widehat{h}$ sul training set. Di conseguenza, scegliere $\widehat{h}$ comporta un errore dovuto al fatto che la frazione di esempi classificati scorrettamente nel training set da un classificatore h è soltanto una stima (possibilmente imprecisa) del rischio $\mathrm{er}(h)$.

Procediamo quindi a controllare l'errore di varianza. Per ogni training set fissato, abbiamo che

$$\begin{aligned}\mathrm{er}(\widehat{h}) - \mathrm{er}(h^*) &= \mathrm{er}(\widehat{h}) - \widehat{\mathrm{er}}(\widehat{h}) + \widehat{\mathrm{er}}(\widehat{h}) - \mathrm{er}(h^*) \\ &\leq \mathrm{er}(\widehat{h}) - \widehat{\mathrm{er}}(\widehat{h}) + \widehat{\mathrm{er}}(h^*) - \mathrm{er}(h^*) \\ &\leq \left|\mathrm{er}(\widehat{h}) - \widehat{\mathrm{er}}(\widehat{h})\right| + \left|\widehat{\mathrm{er}}(h^*) - \mathrm{er}(h^*)\right| \\ &\leq 2 \max_{h\in\mathcal{H}} \left|\widehat{\mathrm{er}}(h) - \mathrm{er}(h)\right|\end{aligned}$$

dove abbiamo usato l'ipotesi che $\widehat{h}$ minimizza $\widehat{\mathrm{er}}(h)$ in $\mathcal{H}$. Quindi, per ogni $\varepsilon > 0$,

$$\mathrm{er}(\widehat{h}) - \mathrm{er}(h^*) > \varepsilon \quad \Rightarrow \quad \max_{h\in\mathcal{H}} \left|\widehat{\mathrm{er}}(h) - \mathrm{er}(h)\right| > \frac{\varepsilon}{2} \quad \Rightarrow \quad \exists h \in \mathcal{H} : \ \left|\widehat{\mathrm{er}}(h) - \mathrm{er}(h)\right| > \frac{\varepsilon}{2} \ .$$

Dato che la catena di implicazioni qui sopra vale per qualsiasi realizzazione del training set, possiamo scrivere

$$\mathbb{P}\left(\mathrm{er}(\widehat{h}) - \mathrm{er}(h^*) > \varepsilon\right) \leq \mathbb{P}\left(\exists h \in \mathcal{H} : \ \left|\widehat{\mathrm{er}}(h) - \mathrm{er}(h)\right| > \frac{\varepsilon}{2}\right) \ .$$

Studiamo il caso $|\mathcal{H}| < \infty$, ovvero lo spazio dei modelli contiene un numero finito di classificatori. Dato che l'evento

$$\exists h \in \mathcal{H} : \ \left|\widehat{\mathrm{er}}(h) - \mathrm{er}(h)\right| > \frac{\varepsilon}{2}$$

è l'unione su ogni $h \in \mathcal{H}$ degli eventi (non necessariamente disgiunti)

$$\left|\widehat{\mathrm{er}}(h) - \mathrm{er}(h)\right| > \frac{\varepsilon}{2}$$

usando la regola della somma

$$\mathbb{P}(A_1 \cup \cdots \cup A_n) \leq \sum_{i=1}^{n} \mathbb{P}(A_i)$$

che vale per qualsiasi collezione $A_1, \ldots, A_n$ di eventi, abbiamo che

$$\begin{aligned}\mathbb{P}\left(\exists h \in \mathcal{H} : \ \left|\widehat{\mathrm{er}}(h) - \mathrm{er}(h)\right| > \frac{\varepsilon}{2}\right) &= \mathbb{P}\left(\bigcup_{h\in\mathcal{H}} \left(\left|\widehat{\mathrm{er}}(h) - \mathrm{er}(h)\right| > \frac{\varepsilon}{2}\right)\right) \\ &\leq \sum_{h\in\mathcal{H}} \mathbb{P}\left(\left|\widehat{\mathrm{er}}(h) - \mathrm{er}(h)\right| > \frac{\varepsilon}{2}\right) \\ &\leq |\mathcal{H}| \max_{h\in\mathcal{H}} \mathbb{P}\left(\left|\widehat{\mathrm{er}}(h) - \mathrm{er}(h)\right| > \frac{\varepsilon}{2}\right) \\ &\leq |\mathcal{H}| 2e^{-m\varepsilon^2/2} \qquad (2)\end{aligned}$$

dove nell'ultimo passo abbiamo usato il maggiorante di Chernoff-Hoeffding.

Quindi, in conclusione, otteniamo che

$$\mathbb{P}\left(\mathrm{er}(\widehat{h}) - \mathrm{er}(h^*) > \varepsilon\right) \leq \mathbb{P}\left(\exists h \in \mathcal{H} : \ \left|\widehat{\mathrm{er}}(h) - \mathrm{er}(h)\right| > \frac{\varepsilon}{2}\right) \leq 2|\mathcal{H}|e^{-m\varepsilon^2/2} \ . \qquad (3)$$

Ponendo il membro destro di (3) uguale a δ e risolvendo rispetto a ε, otteniamo che

$$\mathrm{er}(\widehat{h}) \leq \mathrm{er}(h^*) + \sqrt{\frac{2}{m} \ln \frac{2|\mathcal{H}|}{\delta}}$$

vale con probabilità almeno $1-\delta$ rispetto all'estrazione casuale di un training set di cardinalità m.

Vediamo quindi che il rischio di $\widehat{h}$ si scompone in due contributi: $\mathrm{er}(h^*)$ e $\sqrt{\frac{2}{m}\ln\frac{2|\mathcal{H}|}{\delta}}$. In mancanza di informazioni su (D,η), e per una fissata cardinalità m del training set, per ridurre $\sqrt{\frac{2}{m}\ln\frac{2|\mathcal{H}|}{\delta}}$, ovvero il maggiorante sull'errore di varianza, dobbiamo ridurre $\mathcal{H}$. Ma questo fa potenzialmente aumentare $\mathrm{er}(h^*)$ e quindi l'errore di bias. La necessità di bilanciare l'errore di varianza e l'errore di bias per controllare il rischio di $\widehat{h}$ è il materializzarsi nella teoria del fenomeno dell'overfitting.

Nella dimostrazione del maggiorante sull'errore di varianza abbiamo anche dimostrato in (2) che

$$\forall h \in \mathcal{H} \quad \left|\widehat{\mathrm{er}}(h) - \mathrm{er}(h)\right| \leq \sqrt{\frac{1}{2m}\ln\frac{2|\mathcal{H}|}{\delta}}$$

con probabilità almeno $1-\delta$ rispetto all'estrazione del training set. Questo significa che quando la cardinalità m del training set è abbastanza grande rispetto a $\ln|\mathcal{H}|$, allora il training error $\widehat{\mathrm{er}}(h)$ diventa una buona stima del rischio $\mathrm{er}(h)$ *simultaneamente* per tutti i classificatori $h \in \mathcal{H}$. In queste condizioni, cioè quando la legge dei grandi numeri vale uniformemente rispetto alla scelta $h \in \mathcal{H}$, è chiaro che qualsiasi algoritmo che sceglie classificatori da $\mathcal{H}$ è protetto dall'overfitting.