(F94) Metodi statistici per l'apprendimento

Rischio nei predittori ad albero

Docente: *Nicolò Cesa-Bianchi* versione 20 marzo 2017

L'analisi di rischio ci dice che per un training set di taglia m, con probabilità almeno $1-\delta$ si ha

$$\mathrm{er}(\widehat{h}) \leq \min_{h\in\mathcal{H}} \mathrm{er}(h) + \sqrt{\frac{2}{m}\ln\frac{2|\mathcal{H}|}{\delta}}\ . \tag{1}$$

Se la classe $\mathcal{H}$ è molto grande, questo risultato è debole. Consideriamo per esempio un problema di classificazione dove $\mathcal{X} = \{0,1\}^d$ (i dati hanno d attributi binari).

Fatto 1 *Sia $\mathcal{H}$ l'insieme di tutti i classificatori calcolati da predittori ad albero su $\mathcal{X} = \{0,1\}^d$. Allora $\mathcal{H}$ contiene tutte le funzioni della forma $h : \{0,1\}^d \to \{-1,+1\}$.*

Questo è una conseguenza del fatto che ogni funzione booleana è rappresentabile con un albero di decisione.

Dato che ci sono 2^{2^d} funzioni della forma $h : \{0,1\}^d \to \{-1,+1\}$, $|\mathcal{H}| = 2^{2^d}$ e il maggiorante (1) diventa

$$\mathrm{er}(\widehat{h}) \leq \min_{h\in\mathcal{H}} \mathrm{er}(h) + \sqrt{\frac{2}{m}\left(2^d \ln 2 + \ln\frac{2}{\delta}\right)}\ .$$

Quindi, perché il rischio di $\widehat{h}$ sia piccolo, il training set deve contenere un numero m di esempi esponenziale in d, che risulta essere un numero troppo grande anche per valori moderati di d. Si tratta di un tipico esempio di overfitting.

Per controllare l'overfitting possiamo minimizzare il training error focalizzandoci sulla classe $\mathcal{H}_N$ dei predittori ad albero con N nodi su $\{0,1\}^d$.

Fatto 2 $|\mathcal{H}_N| \leq (2de)^N$.

Dimostrazione. $|\mathcal{H}_N|$ è esprimibile come il prodotto fra: il numero di alberi binari con N nodi, il numero di modi di assegnare test binari su attributi ai nodi interni, il numero di modi di assegnare etichette binarie alle foglie. Assegnando convenzionalmente il figlio sinistro al risultato negativo di un test e il figlio destro al risultato positivo, un test è definito solamente dall'indice i dell'attributo testato. Quindi, se l'albero ha M nodi interni, ci sono d^M modi per assegnare i test ai nodi interni. Inoltre, dato che le foglie sono $N-M$, ci sono 2^{N-M} modi per assegnare etichette binarie alle foglie. Quindi, ogni albero di N nodi può implementare fino a $d^M 2^{N-M} \leq d^N$ (dato che $d \geq 2$) classificatori. Infine, il numero di alberi binari con N nodi è dato dall'$(N-1)$-simo numero di Catalano $C_{N-1} = \frac{1}{N}\binom{2N-2}{N-1}$. Quindi, utilizzando la maggiorazione $\binom{n}{k} \leq \left(\frac{en}{k}\right)^k$ derivata dall'approssimazione di Stirling per i coefficienti binomiali, otteniamo

$$|\mathcal{H}_N| \leq \frac{1}{N}\left(\frac{2e(N-1)}{N-1}\right)^{N-1} d^N \leq (2ed)^N\ .$$

□

Quindi, se $\widehat{h} = \operatorname{argmin}_{\mathcal{H}_N} \widehat{\mathrm{er}}(h)$ per un N fissato, il maggiorante (1) diventa

$$\mathrm{er}(\widehat{h}) \leq \min_{h \in \mathcal{H}_N} \mathrm{er}(h) + \sqrt{\frac{2}{m}\left(N\big(1+\ln(2d)\big) + \ln\frac{2}{\delta}\right)} \ .$$

Da ciò si deduce che in questo caso un training set la cui taglia è dell'ordine di $N \ln d$ è sufficiente per controllare il rischio di $\widehat{h}$ in $\mathcal{H}_N$.

Il risultato appena dimostrato vale per un predittore specifico, quello che minimizza il training error in $\mathcal{H}_N$ per un dato N fissato. In pratica, non è chiaro come scegliere N, che dovrebbe dipendere dalle caratteristiche del training set. Per aggirare questo problema, invece di limitare l'errore di varianza del predittore che minimizza il training error, come abbiamo fatto finora, mostriamo un risultato diverso. Cioè, maggioriamo simultaneamente il rischio di tutti i predittori ad albero, dove il maggiorante del rischio di ogni albero dipende dal training error e dal numero di nodi dell'albero. A questo scopo introduciamo una funzione $w : \mathcal{H} \to [0,1]$ e chiamiamo $w(h)$ il peso del predittore h. Assumiamo che valga

$$\sum_{h \in \mathcal{H}} w(h) \leq 1 \ . \tag{2}$$

Possiamo allora scrivere la seguente catena di disuguaglianze, dove $\varepsilon_h > 0$ è scelto alla fine,

$$\mathbb{P}\left(\exists h \in \mathcal{H} : \ \big|\widehat{\mathrm{er}}(h) - \mathrm{er}(h)\big| > \varepsilon_h\right) \leq \sum_{h \in \mathcal{H}} \mathbb{P}\left(\big|\widehat{\mathrm{er}}(h) - \mathrm{er}(h)\big| > \varepsilon_h\right) \leq \sum_{h \in \mathcal{H}} 2e^{-2m\varepsilon_h^2} \ .$$

Si noti che abbiamo usato il maggiorante di Chernoff-Hoeffding all'ultimo passo. Scegliendo ora

$$\varepsilon_h = \sqrt{\frac{1}{2m}\left(\ln\frac{1}{w(h)} + \ln\frac{2}{\delta}\right)}$$

abbiamo che

$$\mathbb{P}\left(\exists h \in \mathcal{H} : \ \big|\widehat{\mathrm{er}}(h) - \mathrm{er}(h)\big| > \varepsilon_h\right) \leq \sum_{h \in \mathcal{H}} \delta w(h) \leq \delta$$

dove abbiamo usato la proprietà (2) della funzione w.

Una conseguenza di questa analisi è che, con probabilità almeno $1 - \delta$ rispetto all'estrazione del training set abbiamo che

$$\mathrm{er}(h) \leq \widehat{\mathrm{er}}(h) + \sqrt{\frac{1}{2m}\left(\ln\frac{1}{w(h)} + \ln\frac{2}{\delta}\right)} \tag{3}$$

simultaneamente per ogni $h \in \mathcal{H}$. Questo suggerisce un algoritmo alternativo alla minimizzazione del training error. Infatti, mentre ERM suggerisce di usare

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

per un dato N fissato, l'approccio suggerito dalla nuova analisi propone di usare

$$\widehat{h} = \operatorname*{argmin}_{h \in \mathcal{H}} \left(\widehat{\mathrm{er}}(h) + \sqrt{\frac{1}{2m} \left(\ln \frac{1}{w(h)} + \ln \frac{2}{\delta} \right)} \right) \ . \tag{4}$$

La funzione w può essere naturalmente vista come una misura di complessità del classificatore h. Si noti che questa analisi offre una prospettiva nuova sull'overfitting: $\widehat{\mathrm{er}}(h)$ diventa una buona stima di $\mathrm{er}(h)$ quando viene "penalizzato" dal termine

$$\sqrt{\frac{1}{2m} \left(\ln \frac{1}{w(h)} + \ln \frac{2}{\delta} \right)}$$

che rende conto del fatto che abbiamo usato gli m esempi del training set per scegliere un predittore h di complessità $w(h)$.

Vediamo un esempio concreto per i predittori ad albero su $\mathcal{X} = \{0,1\}^d$. Sia $\mathcal{H}$ l'insieme dei 2^{2^d} predittori ad albero che codificano tutti i classificatori $h : \{0,1\}^d \to \{-1,+1\}$. Usando tecniche di teoria dei codici, possiamo codificare ogni predittore ad albero h con N_h nodi usando una stringa binaria $\sigma(h)$ di lunghezza $|\sigma(h)| = (N_h + 1)\lceil \log_2(d+3) \rceil + 2\lfloor \log_2 N_h \rfloor + 1 = \mathcal{O}(N_h \log d)$ in modo che non ci siano due predittori h e h' tali che $\sigma(h)$ è prefisso di $\sigma(h')$. Codici di questo tipo si chiamano *istantanei* e soddisfano la disuguaglianza di Kraft

$$\sum_{h \in \mathcal{H}} 2^{-|\sigma(h)|} \leq 1 \ .$$

Grazie alla disuguaglianza di Kraft —che implica la proprietà (2)— possiamo quindi assegnare il peso $w(h) = 2^{-|\sigma(h)|}$ ad un classificatore h calcolato da un predittore ad albero con N_h nodi. Applicando il maggiorante (3) otteniamo che, con probabilità almeno $1 - \delta$ rispetto all'estrazione del training set,

$$\mathrm{er}(h) \leq \widehat{\mathrm{er}}(h) + \sqrt{\frac{1}{2m} \left(|\sigma(h)| + \ln \frac{2}{\delta} \right)} \qquad \text{con } |\sigma(h)| = \mathcal{O}(N_h \log d)$$

simultaneamente per ogni $h \in \mathcal{H}$. Quindi, un algoritmo di apprendimento per alberi può controllare l'overfitting generando predittori $\widehat{h}$ definiti come

$$\widehat{h} = \operatorname*{argmin}_{h \in \mathcal{H}} \left(\widehat{\mathrm{er}}(h) + \sqrt{\frac{1}{2m} \left(|\sigma(h)| + \ln \frac{2}{\delta} \right)} \right) \ .$$

Questo tipo di analisi giustifica l'osservazione empirica che, a parità di training error, risulta generalmente più affidabile il classificatore ad albero con minor numero di nodi. D'altra parte, esiste un'arbitrarietà nella scelta della funzione di complessità w. In particolare, non è detto che w debba necessariamente essere inversamente proporzionale al numero di nodi dell'albero: possiamo scegliere una qualsiasi altra w a patto che essa soddisfi (2). È quindi corretto interpretare w come un bias che orienta la nostra preferenza verso certi tipi di alberi rispetto ad altri. La scelta del bias che, a parità di training error, privilegia alberi piccoli è conforme al principio del Rasoio di Occam: una regola euristica secondo la quale fra due spiegazioni alternative di uno stesso fenomeno, la più corta è anche quella più affidabile.