(F94) Metodi statistici per l'apprendimento

Compression bounds

Docente: *Nicolò Cesa-Bianchi* versione 30 aprile 2018

In questa lezione vediamo come limitare il rischio statistico del classificatore prodotto da un algoritmo di apprendimento in grado di rappresentare tale classificatore usando un piccolo sottoinsieme del training set.

Consideriamo una sequenza di esempi $S = \big((\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_m, y_m)\big)$ da utilizzare come training set per un problema di classificazione binaria e consideriamo un algoritmo di classificazione A che dato il training set S in input genera un classificatore $A(S)$. Per ogni $K \subseteq \{1, \ldots, m\}$ indichiamo con S_K la sottosequenza del training set che contiene soltanto gli elementi indicizzati da K e indichiamo con $S_{\overline{K}}$ la sottosequenza che contiene soltanto gli elementi indicizzati da $\{1, \ldots, m\} \setminus K$. Chiamiamo sketch per A una qualsiasi sottosequenza S_K di S tale che $A(S') = A(S)$ per ogni sottosequenza $S_K \subseteq S' \subset S$. Quindi A con input S_K genera lo stesso classificatore di A con input S (si noti che una tale sottosequenza esiste sempre dato che S_K può essere anche uguale ad S). Denotando

$$\widehat{h} = A(S) = A(S_K)$$

vogliamo limitare il rischio $\mathrm{er}(\widehat{h})$ in termini di $\widetilde{\mathrm{er}}(\widehat{h})$, dove

$$\widetilde{\mathrm{er}}(\widehat{h}) = \frac{1}{|S_{\overline{K}}|} \sum_{(\boldsymbol{x}_t, y_t) \in S_{\overline{K}}} \mathbb{I}\{\widehat{h}(\boldsymbol{x}_t) \neq y_t\}$$

è la frazione di errori commessi da $\widehat{h}$ sulla sottosequenza di training set $S_{\overline{K}}$ che non include gli esempi dello sketch.

Per comodità, nel resto dell'analisi consideriamo solo algoritmi che ammettano sempre uno sketch di cardinalità al più metà di quella del training set di partenza. Fissiamo quindi un algoritmo A, un training set S di cardinalità m e uno sketch S_K per A di cardinalità $|K| \leq \frac{m}{2}$. È importante notare a questo punto che, per A fissato, K è una funzione del campione casuale S. Introducendo le costanti $\varepsilon_j > 0$ da determinare in seguito, notiamo che

$$\mathrm{er}\big(A(S)\big) > \widetilde{\mathrm{er}}\big(A(S)\big) + \varepsilon_{|K|} \qquad \text{implica} \qquad \exists \text{ sketch } J,\ |J| \leq \tfrac{m}{2},\ \mathrm{er}\big(A(S_J)\big) > \widetilde{\mathrm{er}}\big(A(S_J)\big) + \varepsilon_{|J|}\ .$$

La precedente implicazione permette di spezzare la variabile casuale $|K|$ nell'unione dei suoi possibili valori $|J|$, ottenendo

$$\begin{aligned}
\mathbb{P}\Big(\mathrm{er}\big(A(S)\big) > \widetilde{\mathrm{er}}\big(A(S)\big) + \varepsilon_{|K|}\Big) &\leq \mathbb{P}\Big(\exists \text{ sketch } J,\ |J| \leq \tfrac{m}{2},\ \mathrm{er}\big(A(S_J)\big) > \widetilde{\mathrm{er}}\big(A(S_J)\big) + \varepsilon_{|J|}\Big) \\
&\leq \sum_{j=0}^{m/2} \sum_{\substack{J \text{ sketch}: \\ |J|=j}} \mathbb{P}\Big(\mathrm{er}\big(A(S_J)\big) > \widetilde{\mathrm{er}}\big(A(S_J)\big) + \varepsilon_j\Big)
\end{aligned}$$

dove abbiamo usato la regola della somma $\mathbb{P}(A \cup B) \leq \mathbb{P}(A) + \mathbb{P}(B)$ ed ora le ε_j sono quantità deterministiche. Si noti che $\widetilde{\mathrm{er}}\big(A(S_J)\big)$ denota la frazione di errori di $A(S_J)$ sugli esempi $(\boldsymbol{x}_t, y_t)$ del training set tali che $t \notin J$. Ora, in ciascuna probabilità

$$\mathbb{P}\Big(\mathrm{er}\big(A(S_J)\big) > \widetilde{\mathrm{er}}\big(A(S_J)\big) + \varepsilon_j\Big)$$

il classificatore $A(S_J)$ per definizione è indipendente da tutti gli $m - j$ esempi $(\boldsymbol{x}_t, y_t)$ di training tali che $t \notin J$. Quindi $\widetilde{\mathrm{er}}\big(A(S_J)\big)$, che è proprio determinato da questi $m - j$ esempi, è una media campionaria di un classificatore fissato ed ha valore atteso $\mathrm{er}\big(A(S_J)\big)$. Possiamo dunque applicare il maggiorante di Chernoff-Hoeffding ottenendo

$$\mathbb{P}\Big(\mathrm{er}\big(A(S)\big) > \widetilde{\mathrm{er}}\big(A(S)\big) + \varepsilon_{|K|}\Big) \leq \sum_{j=0}^{m/2} \sum_{J\,:\,|J|=j} e^{-2(m-j)\varepsilon_j^2} \ .$$

Ora, scegliendo

$$\varepsilon_j = \sqrt{\frac{1}{m}\left(\ln\frac{1}{w_j} + \ln\frac{1}{\delta}\right)} \tag{1}$$

dove i pesi $w_j \geq 0$, che determineremo in seguito, soddisfano

$$\sum_{j=0}^{m/2} \sum_{J\,:\,|J|=j} w_j \leq 1 \tag{2}$$

otteniamo

$$\begin{aligned}
\sum_{j=0}^{m/2} \sum_{J\,:\,|J|=j} e^{-2(m-j)\varepsilon_j^2} &\leq \sum_{j=0}^{m/2} \sum_{J\,:\,|J|=j} e^{-m\varepsilon_j^2} && \text{(dato che } j \leq m/2\text{)} \\
&\leq \sum_{j=0}^{m/2} \sum_{J\,:\,|J|=j} \delta\, w_j && \text{(usando (1))} \\
&\leq \delta \ . && \text{(usando (2))}
\end{aligned}$$

Quindi, con probabilità almeno $1 - \delta$ rispetto all'estrazione del training set abbiamo

$$\mathrm{er}\big(A(S)\big) \leq \widetilde{\mathrm{er}}\big(A(S)\big) + \sqrt{\frac{1}{m}\left(\ln\frac{1}{w_{|K|}} + \ln\frac{1}{\delta}\right)} \ .$$

Si noti ora che per soddisfare (2) è sufficiente definire

$$w_j = \frac{1}{m\binom{m}{j}} \ .$$

Utilizzando il maggiorante $\binom{m}{j} \leq \left(\frac{em}{j}\right)^j$ vediamo che

$$\ln\frac{1}{w_j} = \ln m + \ln\binom{m}{j} \leq \ln m + j\ln\frac{em}{j} \leq j + (j+1)\ln m \ . \tag{3}$$

Quindi, con probabilità almeno $1 - \delta$ rispetto all'estrazione del training set abbiamo infine

$$\mathrm{er}\big(A(S)\big) \leq \widetilde{\mathrm{er}}\big(A(S)\big) + \sqrt{\frac{1}{m}\left(|K| + \big(|K|+1\big)\ln m + \ln\frac{1}{\delta}\right)} \ .$$