(F94) Metodi statistici per l'apprendimento

Compression bounds

Docente: *Nicolò Cesa-Bianchi* versione 4 giugno 2016

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 $A(S)$. Chiamiamo sketch una qualsiasi sottosequenza S_0 di S tale che $A(S_0) = A(S)$. Ovvero A con input S_0 genera lo stesso classificatore di A con input S (si noti che questo è vero per qualsiasi algoritmo A dato che S_0 può essere anche uguale ad S). Vogliamo limitare il rischio $\mathrm{er}(\widehat{h})$ in termini di $\widetilde{\mathrm{er}}(\widehat{h})$, dove

$$\widetilde{\mathrm{er}}(\widehat{h}) = \frac{1}{|S \setminus S_0|} \sum_{(\boldsymbol{x}_t, y_t) \in S \setminus S_0} \mathbb{I}\{\widehat{h}(\boldsymbol{x}_t) \neq y_t\}$$

è la frazione di errori commessi da h sulla sottosequenza di training set $S \setminus S_0$ che non include gli esempi dello sketch.

Indichiamo con $\sigma(A, S)$ lo sketch S_0 che l'algoritmo A produce con input S. Quindi, se $\sigma(A, S) = S_0$ allora $A(S_0) = A(S)$. Inoltre, sia $|S_0|$ la dimensione dello sketch S_0. Per ogni $J \subseteq \{1, \ldots, m\}$ sia S_J la sottosequenza del training set che contiene soltanto gli esempi indicizzati da J.

Procediamo ora a limitare il rischio di $A(S)$. Per comodità, nel resto dell'analisi consideriamo solo algoritmi A tali che $\big|\sigma(A, S)\big| \leq \frac{m}{2}$. Introducendo $\varepsilon_k > 0$ da determinare in seguito notiamo che, per ogni training set S fissato,

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

dove $S_J = \sigma(A, S)$. Quindi possiamo scrivere

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

dove abbiamo usato la regola della somma $\mathbb{P}(A \cup B) \leq \mathbb{P}(A) + \mathbb{P}(B)$. 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

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

il classificatore $A(S_J)$ per definizione è indipendente da tutti gli $m - k$ 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 $m - k$ esempi, è una media

campionaria di un classificatore fissato ed ha valore atteso $\mathrm{er}\big(A(S_J)\big)$. Possiamo quindi applicare il maggiorante di Chernoff-Hoeffding ottenendo

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

Ora, scegliendo

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

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

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

otteniamo

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

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

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

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

$$w_k = \frac{1}{m\binom{m}{k}} \ .$$

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

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

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

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