(F94) Metodi statistici per l'apprendimento

Stability bounds e controllo del rischio in SVM

Docente: *Nicolò Cesa-Bianchi* versione 1 giugno 2016

Come sappiamo, un modo per controllare l'overfitting è quello di impedire all'algoritmo di apprendimento di produrre dei predittori che abbiamo un training error arbitrariamente basso su qualunque training set. Questo comportamento può essere ottenuto imponendo una condizione di "stabilità" sull'algoritmo stesso. Ovvero, richiedendo che il training error del predittore prodotto dall'algoritmo non cambi significativamente quando il training set viene leggermente perturbato.

Fissiamo un training set $S = \big((\boldsymbol{x}_1, y_1), \dots, (\boldsymbol{x}_m, y_m)\big)$ e usiamo $\boldsymbol{z}_t$ per denotare il t-esimo esempio $(\boldsymbol{x}_t, y_t) \in \mathbb{R}^d \times \mathbb{R}$. Data una funzione di perdita ℓ e un predittore h, indichiamo con $\ell(h, \boldsymbol{z}_t)$ la perdita di h sull'esempio $\boldsymbol{z}_t$ e usiamo

$$\ell_S(h) = \frac{1}{m} \sum_{t=1}^{m} \ell(h, \boldsymbol{z}_t)$$

per denotare il training error di h. Fissato inoltre un modello statistico $\mathcal{D}$, denotiamo con $\ell_{\mathcal{D}}(h)$ il rischio statistico di h. Nel seguito, assumiamo che S sia un campione statistico estratto da $\mathcal{D}$ e usiamo la notazione $S^{(t)}$ per indicare il training set S in cui in t-esimo esempio $(\boldsymbol{x}_t, y_t)$ è stato sostituito con un altro esempio $\boldsymbol{z}'_t = (\boldsymbol{x}'_t, y'_t)$ estratto da $\mathcal{D}$ in modo indipendente da S.

Fissato un algoritmo di apprendimento A, indichiamo con h_S e $h_{S^{(t)}}$ i predittori generati da A con input, rispettivamente, S e $S^{(t)}$. Diciamo che A è ε-stabile se, per ogni $t = 1, \dots, m$

$$\mathbb{E}\big[\ell(h_S, \boldsymbol{z}'_t) - \ell(h_{S^{(t)}}, \boldsymbol{z}'_t)\big] \le \varepsilon$$

dove il valore atteso è calcolato rispetto alle estrazioni di S e $\boldsymbol{z}'_t = (\boldsymbol{x}'_t, y'_t)$. In altri termini, un algoritmo è stabile se, in media rispetto all'estrazione del training set, l'algoritmo produce un predittore il cui errore su un qualunque elemento del training set è vicino a quello commesso sullo stesso elemento da un predittore addestrato su un training set dove quell'elemento è sostituito da un altro.

I due prossimi risultati mostrano che un algoritmo stabile non soffre di overfitting.

Teorema 1 *Se A è ε-stabile allora $\mathbb{E}\big[\ell_{\mathcal{D}}(h_S) - \ell_S(h_S)\big] \le \varepsilon$.*

Dimostrazione. Consideriamo $S = \big((\boldsymbol{x}_1, y_1), \dots, (\boldsymbol{x}_m, y_m)\big)$ e $S' = \big((\boldsymbol{x}'_1, y'_1), \dots, (\boldsymbol{x}'_m, y'_m)\big)$ estratti in modo indipendente da $\mathcal{D}$. Allora

$$\mathbb{E}\big[\ell_S(h_S)\big] = \mathbb{E}\left[\frac{1}{m}\sum_{t=1}^{m} \ell(h_S, \boldsymbol{z}_t)\right] = \frac{1}{m}\sum_{t=1}^{m} \mathbb{E}\big[\ell(h_S, \boldsymbol{z}_t)\big] = \frac{1}{m}\sum_{t=1}^{m} \mathbb{E}\big[\ell(h_{S^{(t)}}, \boldsymbol{z}'_t)\big] \ .$$

Inoltre,

$$\ell_{\mathcal{D}}(h_S) = \mathbb{E}\big[\ell(h_S, \boldsymbol{z}'_t) \mid S\big] = \frac{1}{m}\sum_{t=1}^{m} \mathbb{E}\big[\ell(h_S, \boldsymbol{z}'_t) \mid S\big]$$

che implica, facendo la media rispetto all'estrazione di S,

$$\mathbb{E}\big[\ell_{\mathcal{D}}(h_S)\big] = \frac{1}{m}\sum_{t=1}^{m}\mathbb{E}\big[\ell(h_S, \boldsymbol{z}'_t)\big] \ .$$

Quindi,

$$\mathbb{E}\big[\ell_{\mathcal{D}}(h_S) - \ell_S(h_S)\big] = \frac{1}{m}\sum_{t=1}^{m}\mathbb{E}\big[\ell(h_S, \boldsymbol{z}'_t) - \ell(h_{S^{(t)}}, \boldsymbol{z}'_t)\big] \leq \varepsilon$$

per l'assunzione di stabilità. □

Dato che la stabilità e la minimizzazione del training error sono obiettivi antagonisti, l'algoritmo ERM (che minimizza il rischio empirico) non è necessariamente stabile. D'altra parte, se un algoritmo di apprendimento è stabile e simultaneamente in grado di produrre predittori con basso rischio empirico (non necessariamente minimo) allora il suo errore di varianza è basso.

Teorema 2 *Se A è ε-stabile e inoltre minimizza in modo approssimato il rischio empirico in una classe $\mathcal{H}$ di predittori, ovvero*

$$\ell_S(h_S) \leq \inf_{h\in\mathcal{H}} \ell_S(h) + \gamma$$

per un qualche $\gamma > 0$, allora

$$\mathbb{E}\big[\ell_{\mathcal{D}}(h_S)\big] \leq \inf_{h\in\mathcal{H}} \ell_{\mathcal{D}}(h) + \varepsilon + \gamma \ .$$

Dimostrazione. Sia h^* il predittore a rischio minimo in $\mathcal{H}$. Allora

$$\begin{aligned}\mathbb{E}\big[\ell_{\mathcal{D}}(h_S)\big] &= \mathbb{E}\Big[\underbrace{\ell_{\mathcal{D}}(h_S) - \ell_S(h_S)}_{\leq\varepsilon \text{ (stabilità)}}\Big] + \mathbb{E}\Big[\ \underbrace{\ell_S(h_S) - \ell_S(h^*)}_{\leq\gamma \text{ (ERM approssimato)}}\ \Big] + \mathbb{E}\big[\ell_S(h^*)\big] \\ &\leq \varepsilon + \gamma + \mathbb{E}\big[\ell_S(h^*)\big] \ .\end{aligned}$$

La dimostrazione è conclusa osservando che $\mathbb{E}\big[\ell_S(h^*)\big] = \ell_{\mathcal{D}}(h^*)$, dato che il valore atteso del rischio empirico è il rischio. □

Nel caso di predittori parametrizzati da un vettore $\boldsymbol{w} \in \mathbb{R}^d$ (come quelli lineari), il minimizzatore del rischio empirico, rispetto ad una funzione di perdita data ℓ, può essere reso stabile aggiungendo a ℓ un termine cosiddetto di regolarizzazione. In realtà abbiamo bisogno anche di un'altra condizione, ovvero che la funzione $\ell(\cdot, \boldsymbol{z})$, dove $\ell(\boldsymbol{w}, \boldsymbol{z})$ è l'errore di $\boldsymbol{w}$ sull'esempio $\boldsymbol{z}$, sia convessa e Lipschitz. Ricordiamo che Lipschitz significa che esiste una costante L tale che $\big|\ell(\boldsymbol{w}, \boldsymbol{z}) - \ell(\boldsymbol{w}', \boldsymbol{z})\big| \leq L\,\|\boldsymbol{w} - \boldsymbol{w}'\|$ per ogni $\boldsymbol{w}, \boldsymbol{w}' \in \mathbb{R}^d$ e per ogni $\boldsymbol{z} = (\boldsymbol{x}, y)$. Non sono richieste altre assunzioni su ℓ.

Teorema 3 *Sia $\ell(\boldsymbol{w}, \boldsymbol{z})$ una funzione di perdita tale che $\ell(\cdot, \boldsymbol{z})$ è convessa, differenziabile e Lipschitz con costante $L > 0$. Allora l'algoritmo di apprendimento che, su input S, produce*

$$\boldsymbol{w}_S = \operatorname*{argmin}_{\boldsymbol{w}\in\mathbb{R}^d} \left(\ell_S(\boldsymbol{w}) + \frac{\lambda}{2}\,\|\boldsymbol{w}\|^2\right)$$

è $(2L)^2/(\lambda m)$-stabile per ogni $\lambda > 0$.

DIMOSTRAZIONE. Definiamo

$$F_S(\boldsymbol{w}) = \ell_S(\boldsymbol{w}) + \frac{\lambda}{2}\|\boldsymbol{w}\|^2$$

e osserviamo che

$$\boldsymbol{w}_S = \operatorname*{argmin}_{\boldsymbol{w}\in\mathbb{R}^d} F_S(\boldsymbol{w}) \qquad \text{e} \qquad \boldsymbol{w}_{S^{(t)}} = \operatorname*{argmin}_{\boldsymbol{w}\in\mathbb{R}^d} F_{S^{(t)}}(\boldsymbol{w}) \ .$$

Per dimostrare la stabilità dobbiamo maggiorare $\mathbb{E}\big[\ell(\boldsymbol{w}_S, \boldsymbol{z}'_t) - \ell(\boldsymbol{w}_{S^{(t)}}, \boldsymbol{z}'_t)\big]$. Nella dimostrazione otterremo un risultato più forte, andando a maggiorare la quantità $\ell(\boldsymbol{w}_S, \boldsymbol{z}'_t) - \ell(\boldsymbol{w}_{S^{(t)}}, \boldsymbol{z}'_t)$ per ogni S e $\boldsymbol{z}'_t$. Come primo passo, usiamo Lipschitz per scrivere

$$\ell(\boldsymbol{w}_S, \boldsymbol{z}'_t) - \ell(\boldsymbol{w}_{S^{(t)}}, \boldsymbol{z}'_t) \le L\,\|\boldsymbol{w}_S - \boldsymbol{w}_{S^{(t)}}\| \ . \tag{1}$$

Quindi, procediamo a maggiorare $\|\boldsymbol{w}_S - \boldsymbol{w}_{S^{(t)}}\|$. Introduciamo le abbreviazioni $\boldsymbol{w} = \boldsymbol{w}_S$ e $\boldsymbol{w}' = \boldsymbol{w}_{S^{(t)}}$. Allora

$$\begin{aligned}
F_S(\boldsymbol{w}') - F_S(\boldsymbol{w}) &= \ell_S(\boldsymbol{w}') - \ell_S(\boldsymbol{w}) + \frac{\lambda}{2}\big(\|\boldsymbol{w}'\|^2 - \|\boldsymbol{w}\|^2\big) \\
&= \ell_{S^{(t)}}(\boldsymbol{w}') - \ell_{S^{(t)}}(\boldsymbol{w}) + \frac{\ell(\boldsymbol{w}', \boldsymbol{z}_t) - \ell(\boldsymbol{w}, \boldsymbol{z}_t)}{m} - \frac{\ell(\boldsymbol{w}', \boldsymbol{z}'_t) - \ell(\boldsymbol{w}, \boldsymbol{z}'_t)}{m} + \frac{\lambda}{2}\big(\|\boldsymbol{w}'\|^2 - \|\boldsymbol{w}\|^2\big) \\
&= F_{S^{(t)}}(\boldsymbol{w}') - F_{S^{(t)}}(\boldsymbol{w}) + \frac{\ell(\boldsymbol{w}', \boldsymbol{z}_t) - \ell(\boldsymbol{w}, \boldsymbol{z}_t)}{m} - \frac{\ell(\boldsymbol{w}', \boldsymbol{z}'_t) - \ell(\boldsymbol{w}, \boldsymbol{z}'_t)}{m} \\
&\le \frac{\big|\ell(\boldsymbol{w}', \boldsymbol{z}_t) - \ell(\boldsymbol{w}, \boldsymbol{z}_t)\big|}{m} + \frac{\big|\ell(\boldsymbol{w}', \boldsymbol{z}'_t) - \ell(\boldsymbol{w}, \boldsymbol{z}'_t)\big|}{m} \\
&\le \frac{2L}{m}\|\boldsymbol{w} - \boldsymbol{w}'\|
\end{aligned}$$

dove la prima disuguaglianza vale perché $\boldsymbol{w}' = \boldsymbol{w}_{S^{(t)}}$ minimizza $F_{S^{(t)}}$ e la seconda disuguaglianza vale per il fatto che $\ell(\cdot, \boldsymbol{z})$ è Lipschitz.

Proseguiamo notando che la funzione F_S è λ-fortemente convessa: infatti $\ell(\cdot, \boldsymbol{z})$ è convessa, $\frac{\lambda}{2}\|\boldsymbol{w}\|^2$ è λ-fortemente convessa e quindi la loro somma è λ-fortemente convessa. Allora, per definizione di funzione fortemente convessa,

$$F_S(\boldsymbol{w}') \ge F_S(\boldsymbol{w}) + \nabla F_S(\boldsymbol{w})^\top(\boldsymbol{w}' - \boldsymbol{w}) + \frac{\lambda}{2}\|\boldsymbol{w} - \boldsymbol{w}'\|^2 \ .$$

Dato che $\boldsymbol{w}$ è il minimo di F_S, $\nabla F_S(\boldsymbol{w}) = \boldsymbol{0}$ e quindi

$$F_S(\boldsymbol{w}') - F_S(\boldsymbol{w}) = \Big(\ell_S(\boldsymbol{w}') + \frac{\lambda}{2}\|\boldsymbol{w}'\|^2\Big) - \Big(\ell_S(\boldsymbol{w}) + \frac{\lambda}{2}\|\boldsymbol{w}\|^2\Big) \ge \frac{\lambda}{2}\|\boldsymbol{w} - \boldsymbol{w}'\|^2$$

Combinando assieme le due disuguaglianze otteniamo

$$\frac{\lambda}{2}\|\boldsymbol{w} - \boldsymbol{w}'\|^2 \le \frac{2L}{m}\|\boldsymbol{w} - \boldsymbol{w}'\| \qquad \text{ovvero} \qquad \|\boldsymbol{w} - \boldsymbol{w}'\| \le \frac{4L}{\lambda m}$$

che, combinato con (1) dimostra la stabilità di $\boldsymbol{w} = \boldsymbol{w}_S$. □

Mostriamo ora come la nozione di stabilità può essere usata per ottenere un maggiorante sul rischio del predittore SVM. Prima di tutto, ricordiamo che la hinge loss $\ell\big(\boldsymbol{w}, (\boldsymbol{x}, y)\big) = \big[1 - y\,\boldsymbol{w}^\top\boldsymbol{x}\big]_+$ per $(\boldsymbol{x}, y) \in \mathbb{R}^d \times \{-1, +1\}$ è convessa e Lipschitz con costante $L = 1$.

Teorema 4 *Dato un training set S, la soluzione SVM*

$$\boldsymbol{w}_S = \operatorname*{argmin}_{\boldsymbol{w}\in\mathbb{R}^d} \left(\ell_S(\boldsymbol{w}) + \frac{\lambda}{2} \|\boldsymbol{w}\|^2 \right)$$

soddisfa

$$\mathbb{E}\big[\ell_{\mathcal{D}}(\boldsymbol{w}_S)\big] \leq \min_{\boldsymbol{u}\in\mathbb{R}^d} \Big(\ell_{\mathcal{D}}(\boldsymbol{u}) + \frac{\lambda}{2} \|\boldsymbol{u}\|^2 \Big) + \frac{4}{\lambda m} \ .$$

Dimostrazione. Chiaramente, per ogni $\boldsymbol{u} \in \mathbb{R}^d$ vale

$$\ell_S(\boldsymbol{w}_S) \leq \ell_S(\boldsymbol{w}_S) + \frac{\lambda}{2} \|\boldsymbol{w}_S\|^2 \leq \ell_S(\boldsymbol{u}) + \frac{\lambda}{2} \|\boldsymbol{u}\|^2 \ . \tag{2}$$

Quindi, dato che per il Teorema 3 la soluzione $\boldsymbol{w}_S$ è $(4L^2)/(\lambda m)$-stabile per $L = 1$,

$$\begin{aligned}
\mathbb{E}\big[\ell_{\mathcal{D}}(\boldsymbol{w}_S)\big] &\leq \mathbb{E}\big[\ell_S(\boldsymbol{w}_S)\big] + \frac{2}{\lambda m} \qquad \text{per il Teorema 1} \\
&\leq \mathbb{E}\left[\ell_S(\boldsymbol{u}) + \frac{\lambda}{2} \|\boldsymbol{u}\|^2\right] + \frac{4}{\lambda m} \qquad \text{usando (2)} \\
&= \ell_{\mathcal{D}}(\boldsymbol{u}) + \frac{\lambda}{2} \|\boldsymbol{u}\|^2 + \frac{4}{\lambda m}
\end{aligned}$$

come volevamo dimostrare. □

Si noti che scegliendo $\lambda = \sqrt{8/m}$ e usando il fatto che il rischio di classificazione $\mathrm{er}_{\mathcal{D}}(\boldsymbol{w}_S) = \mathbb{P}\big(Y\, \boldsymbol{w}_S^\top \boldsymbol{X} \leq 0\big)$ è maggiorato dal rischio della hinge loss $\mathbb{E}\big[\ell_{\mathcal{D}}(\boldsymbol{w}_S)\big]$, otteniamo

$$\mathbb{E}\big[\mathrm{er}_{\mathcal{D}}(\boldsymbol{w}_S)\big] \leq \min_{\boldsymbol{u}\in\mathbb{R}^d} \left(\ell_{\mathcal{D}}(\boldsymbol{u}) + \sqrt{\frac{2}{m}} \|\boldsymbol{u}\|^2 \right) + \sqrt{\frac{2}{m}} \ .$$