**(F94) Metodi statistici per l'apprendimento**

# Da rischio sequenziale a rischio statistico

Docente: *Nicolò Cesa-Bianchi* versione 16 maggio 2018

Ci occupiamo ora mettere in relazione il rischio sequenziale di un algoritmo online (come OGD) con il rischio di un predittore da esso prodotto, assumendo che i dati sui quali l'algoritmo viene eseguito siano generati da una sorgente statistica. Questo ci permette di creare un ponte fra il modello di apprendimento sequenziale e quello statistico.

Fissiamo una funzione di perdita $\ell$. Dato un predittore lineare $\boldsymbol{w} \in \mathbb{R}^d$ e un esempio $(\boldsymbol{x}_t, y_t) \in \mathbb{R}^d \times \mathbb{R}$, sia $\ell_t(\boldsymbol{w}) = \ell\big(\boldsymbol{w}^\top \boldsymbol{x}, y_t\big)$ la perdita associata. Per esempio, $\ell\big(\boldsymbol{w}^\top \boldsymbol{x}, y_t\big) = \big(\boldsymbol{w}^\top \boldsymbol{x}_t - y_t\big)^2$ in regressione o $\ell\big(\boldsymbol{w}^\top \boldsymbol{x}, y_t\big) = \big[1 - y_t\, \boldsymbol{w}^\top \boldsymbol{x}_t\big]_+$ in classificazione con hinge loss. Nel seguito, assumeremo che la funzione $\ell_t$ sia convessa.

Consideriamo il caso dell'apprendimento statistico, in cui gli esempi $(\boldsymbol{x}_t, y_t)$ sono realizzazioni indipendenti di variabili casuali $(\boldsymbol{X}_t, Y_t)$ con distribuzione comune $\mathcal{D}$ su $\mathbb{R}^d \times \mathbb{R}$ fissata ma ignota. Il rischio statistico rispetto alla funzione di perdita $\ell$ di un predittore lineare $\boldsymbol{w}$ è definito da

$$\mathrm{er}_{\mathcal{D}}(\boldsymbol{w}) = \mathbb{E}\Big[\ell\big(\boldsymbol{w}^\top \boldsymbol{X}, Y\big)\Big]$$

dove l'esempio $(\boldsymbol{X}, Y)$ è estratto dalla distribuzione congiunta $\mathcal{D}$ su $\mathbb{R}^d \times \mathbb{R}$.

Consideriamo ora un training set $S$ composto da realizzazioni $(\boldsymbol{x}_1, y_1), \dots, (\boldsymbol{x}_m, y_m)$ di variabili casuali $(\boldsymbol{X}_1, Y_1), \dots, (\boldsymbol{X}_m, Y_m)$ estratte da $\mathcal{D}$. Eseguiamo un algoritmo online come OGD sul training set $S$ con la sequenza di funzioni di perdita $\ell_1, \dots, \ell_m$ definite da $\ell_t(\boldsymbol{w}) = \ell\big(\boldsymbol{w}^\top \boldsymbol{x}_t, y_t\big)$. Per definizione di algoritmo online, otteniamo una sequenza $\boldsymbol{w}_1, \dots, \boldsymbol{w}_m$ di predittori lineari (quelli prodotti da OGD durante il run). Vogliamo ora stabilire un maggiorante sul rischio statistico di un predittore lineare ottenuto in modo naturale dalla sequenza, ovvero il *predittore medio*

$$\overline{\boldsymbol{w}} = \frac{1}{m} \sum_{t=1}^{m} \boldsymbol{w}_t \ .$$

Anzitutto, dato che $\ell$ è convessa, la disuguaglianza di Jensen ci dà che

$$\mathrm{er}_{\mathcal{D}}(\overline{\boldsymbol{w}}) = \mathbb{E}\Big[\ell\big(\overline{\boldsymbol{w}}^\top \boldsymbol{X}, Y\big)\Big] \le \mathbb{E}\left[\frac{1}{m} \sum_{t=1}^{m} \ell\big(\boldsymbol{w}_t^\top \boldsymbol{X}, Y\big)\right] = \frac{1}{m} \sum_{t=1}^{m} \mathrm{er}_{\mathcal{D}}(\boldsymbol{w}_t)$$

dove l'ultima eguaglianza vale per linearità dell'aspettazione. Quindi il rischio del predittore medio è maggiorato dal rischio medio dei predittori della sequenza $\boldsymbol{w}_1, \dots, \boldsymbol{w}_m$.

Il passo cruciale sta ora nel legare il rischio statistico medio con il rischio sequenziale. Questo viene fatto osservando che, sotto l'ipotesi che $S$ sia un campione casuale estratto da $\mathcal{D}$, $\boldsymbol{w}_t$ è determinato dai primi $t-1$ esempi $(\boldsymbol{x}_1, y_1), \dots, (\boldsymbol{x}_{t-1}, y_{t-1})$ estratti. Quindi, applicando la definizione di rischio al valore atteso della perdita di $\boldsymbol{w}_t$ sul $t$-esimo esempio $(\boldsymbol{x}_t, y_t)$, possiamo scrivere

$$\mathbb{E}\Big[\mathrm{er}_{\mathcal{D}}(\boldsymbol{w}_t) - \ell\big(\boldsymbol{w}_t^\top \boldsymbol{X}_t, Y_t\big) \,\Big|\, (\boldsymbol{X}_1, Y_1), \dots, (\boldsymbol{X}_{t-1}, Y_{t-1})\Big] = 0 \ . \tag{1}$$

La relazione qui sopra dice la seguente cosa: se condizioniamo sui primi $t-1$ esempi, il valore atteso di $\ell_t(\boldsymbol{w}_t)$ rispetto all'estrazione del $t$-esimo esempio è semplicemente (per definizione) il rischio di $\boldsymbol{w}_t$.

Per semplicità, denotiamo con $\mathbb{E}_{t-1}$ il valore atteso condizionato come sopra. Se sommiamo entrambi i membri di (1) per $t = 1, \dots, m$ e dividiamo per $m$ otteniamo

$$\frac{1}{m} \sum_{t=1}^{m} \mathbb{E}_{t-1}\Big[\mathrm{er}_{\mathcal{D}}(\boldsymbol{w}_t) - \ell\big(\boldsymbol{w}_t^\top \boldsymbol{X}_t, Y_t\big)\Big] = 0 \ .$$

Per ogni $t = 1, \dots, m$ sia $Z_t$ la variabile casuale $\mathrm{er}_{\mathcal{D}}(\boldsymbol{w}_t) - \ell\big(\boldsymbol{w}_t^\top \boldsymbol{X}_t, Y_t\big)$. Le variabili casuali $Z_1, \dots, Z_m$ sono tutte funzioni del medesimo campione $S$ e sono tali che

$$\frac{1}{m} \sum_{t=1}^{m} \mathbb{E}_{t-1}[Z_t] = 0 \ .$$

Assumiamo che $\ell_t \in [0, M]$, allora $|Z_t| \leq M$. Questo tipo di variabili casuali, o più precisamente di processo, viene definito sequenza di differenze di martingale con incrementi limitati da $2M$. Si noti che le $Z_t$ non sono indipendenti. Da un punto di vista statistico, però, queste variabili si comportano come se lo fossero (almeno per certi aspetti). In particolare, vale una legge dei grandi numeri del tipo seguente

$$\frac{1}{m} \sum_{t=1}^{m} Z_t \leq 2M \sqrt{\frac{1}{2m} \ln \frac{1}{\delta}}$$

con probabilità almeno $1 - \delta$ rispetto all'estrazione di $S$. Questo significa che

$$\frac{1}{m} \sum_{t=1}^{m} \mathrm{er}_{\mathcal{D}}(\boldsymbol{w}_t) \leq \frac{1}{m} \sum_{t=1}^{m} \ell\big(\boldsymbol{w}_t^\top \boldsymbol{X}_t, Y_t\big) + M \sqrt{\frac{2}{m} \ln \frac{2}{\delta}} \tag{2}$$

con probabilità almeno $1 - \delta/2$ rispetto all'estrazione di $S$. Ritornando al predittore medio $\overline{\boldsymbol{w}}$, il risultato che otteniamo può essere sintetizzato come

$$\mathrm{er}_{\mathcal{D}}(\overline{\boldsymbol{w}}) \leq \frac{1}{m} \sum_{t=1}^{m} \ell\big(\boldsymbol{w}_t^\top \boldsymbol{x}_t, y_t\big) + \mathcal{O}\left(\frac{1}{\sqrt{m}}\right) \qquad \text{con alta probabilità.}$$

In altre parole, il rischio del predittore medio è limitato in probabilità dal rischio sequenziale sul training set.

Possiamo spingerci ulteriormente per ottenere un maggiorante di rischio più esplicito che sfrutti direttamente i risultati dell'analisi online. Per esempio, per regressione con perdita quadratica, se facciamo girare OGD con proiezione nell'insieme $\big\{\boldsymbol{u} \in \mathbb{R}^d \,:\, \|\boldsymbol{u}\| \leq U\big\}$, e assumiamo che $\|\boldsymbol{x}_t\| \leq X$ e $|y_t| \leq UX$ per ogni $t$, otteniamo che, per *ogni* realizzazione $(\boldsymbol{x}_1, y_1), \dots, (\boldsymbol{x}_m, y_m)$ di $S$,

$$\frac{1}{m} \sum_{t=1}^{m} \ell\big(\boldsymbol{w}_t^\top \boldsymbol{x}_t, y_t\big) \leq \min_{\boldsymbol{u} \in \mathbb{R}^d \,:\, \|\boldsymbol{u}\| \leq U} \frac{1}{m} \sum_{t=1}^{m} \ell\big(\boldsymbol{u}^\top \boldsymbol{x}_t, y_t\big) + 8(UX)^2 \sqrt{\frac{2}{m}} \ .$$

Sostituendo il membro destro in (2) e notando che $M = 4(UX)^2$ per la funzione di perdita quadratica, possiamo quindi scrivere che

$$\mathrm{er}_{\mathcal{D}}(\overline{\boldsymbol{w}}) \leq \min_{\boldsymbol{u} \in \mathbb{R}^d \,:\, \|\boldsymbol{u}\| \leq U} \frac{1}{m} \sum_{t=1}^{m} \ell\big(\boldsymbol{u}^\top \boldsymbol{X}_t, Y_t\big) + 12(UX)^2 \sqrt{\frac{2}{m} \ln \frac{2}{\delta}}$$

con probabilità almeno $1 - \delta/2$ rispetto all'estrazione di $S$.

Infine, possiamo notare che, detto

$$\boldsymbol{u}^* = \operatorname*{argmin}_{\boldsymbol{u} \in \mathbb{R}^d \,:\, \|\boldsymbol{u}\| \leq U} \operatorname{er}_{\mathcal{D}}(\boldsymbol{u})$$

abbiamo, ovviamente,

$$\min_{\boldsymbol{u} \in \mathbb{R}^d \,:\, \|\boldsymbol{u}\| \leq U} \frac{1}{m} \sum_{t=1}^{m} \ell\big(\boldsymbol{u}^\top \boldsymbol{x}_t, y_t\big) \leq \frac{1}{m} \sum_{t=1}^{m} \ell\big(\boldsymbol{x}_t^\top \boldsymbol{u}^*, y_t\big) \ .$$

Dato che, per ogni $t = 1, \ldots, m$ si ha $\mathbb{E}\Big[\ell\big(\boldsymbol{X}_t^\top \boldsymbol{u}^*, Y_t\big)\Big] = \operatorname{er}_{\mathcal{D}}(\boldsymbol{u}^*)$, possiamo applicare il maggiorante di Chernoff-Hoeffding e dedurre che

$$\frac{1}{m} \sum_{t=1}^{m} \ell\big(\boldsymbol{X}_t^\top \boldsymbol{u}^*, Y_t\big) \leq \operatorname{er}_{\mathcal{D}}(\boldsymbol{u}^*) + 4(UX)^2 \sqrt{\frac{1}{2m} \ln \frac{2}{\delta}} \qquad \text{con probabilità almeno } 1 - \delta/2.$$

Abbiamo così ottenuto il seguente maggiorante esplicito sul rischio del predittore medio

$$\operatorname{er}_{\mathcal{D}}(\overline{\boldsymbol{w}}) \leq \operatorname{er}_{\mathcal{D}}(\boldsymbol{u}^*) + 14(UX)^2 \sqrt{\frac{2}{m} \ln \frac{2}{\delta}}$$

con probabilità almeno $1 - \delta$ rispetto all'estrazione di $S$.

È possibile dimostrare che l'algoritmo che minimizza il rischio empirico,

$$\widehat{\boldsymbol{w}} = \operatorname*{argmin}_{\boldsymbol{w} \in \mathbb{R}^d \,:\, \|\boldsymbol{w}\| \leq U} \frac{1}{m} \sum_{t=1}^{m} \ell\big(\boldsymbol{w}^\top \boldsymbol{x}_t, y_t\big)$$

ha un errore di varianza dello stesso ordine di grandezza,

$$\operatorname{er}_{\mathcal{D}}(\widehat{\boldsymbol{w}}) \leq \operatorname{er}_{\mathcal{D}}(\boldsymbol{u}^*) + \mathcal{O}\left(\frac{(UX)^2}{\sqrt{m}}\right)$$

con alta probabilità rispetto all'estrazione del training set.