(F94) Metodi statistici per l'apprendimento

Rischio in Nearest Neighbour

Docente: *Nicolò Cesa-Bianchi* versione 27 marzo 2018

In questa lezione vediamo come limitare il valore atteso del rischio statistico del classificatore prodotto da 1-NN rispetto all'estrazione del training set. Il risultato che vogliamo ottenere è del tipo

$$\mathbb{E}\big[\mathrm{er}(\widehat{h}_S)\big] \leq 2\,\mathrm{er}(f^*) + \varepsilon_m \tag{1}$$

dove $\widehat{h}_S$ indica il classificatore 1-NN prodotto sul training set S di taglia m, $\mathrm{er}(f^*)$ è il rischio del classificatore Bayesiano ottimo e ε_m è una quantità che dipende da m. Questo tipo di risultato è diverso da quelli ottenuti in precedenza per il classificatore ERM (minimizzatore del training error), qui denotato con $\widetilde{h}_S$, che erano del tipo

$$\mathbb{P}\left(\mathrm{er}(\widetilde{h}_S) > \min_{h\in\mathcal{H}} \mathrm{er}(h) + \varepsilon\right) \leq \delta_{m,\varepsilon}\ . \tag{2}$$

dove $\delta_{m,\varepsilon} > 0$ è una quantità che dipende sia da m che da ε. Le differenze sono di due tipi:

1. Il maggiorante (1) maggiora il rischio del classificatore 1-NN in termini assoluti (cioè si confronta direttamente col Bayes risk), mentre il maggiorante (2) lo maggiora relativamente al rischio del miglior classificatore in $\mathcal{H}$ (che potrebbe essere arbitrariamente peggiore del Bayes risk).
2. Il maggiorante (2) descrive una proprietà della distribuzione del rischio dei classificatori ERM al variare di $\varepsilon > 0$, mentre il maggiorante (1) limita semplicemente il rischio di un "tipico" classificatore 1-NN

Dato un training set $S = (\boldsymbol{X}_1, Y_1), \ldots, (\boldsymbol{X}_m, Y_m)$ estratto da un modello statistico (D, η) definiamo la mappa $\pi_S : \mathbb{R}^d \to \{1, \ldots, m\}$ come

$$\pi_S(\boldsymbol{x}) = \operatorname*{argmin}_{t=1,\ldots,m} \|\boldsymbol{x} - \boldsymbol{X}_t\|\ .$$

Il predittore 1-NN su input S è quindi definito da $\widehat{h}_S(\boldsymbol{x}) = y_{\pi_S(\boldsymbol{x})}$.

Nel seguito, assumiamo che i dati $\boldsymbol{x}$ generati dalla sorgente siano tali che $\max_i |x_i| \leq 1$ con probabilità uno. Ovvero, le componenti x_i delle istanze $\boldsymbol{x} = (x_1, \ldots, x_d)$ generate dalla sorgente hanno sempre valori compresi fra -1 e $+1$. Sia $\mathcal{X} \subset \mathbb{R}^d$ il sottoinsieme delle istanze $\boldsymbol{x}$ con questa proprietà. Esprimeremo il maggiorante sul valore atteso del rischio di 1-NN in termini di una quantità che caratterizza il modello statistico (D, η), ovvero il più piccolo $c > 0$ tale che

$$\big|\eta(\boldsymbol{x}) - \eta(\boldsymbol{x}')\big| \leq c\,\big\|\boldsymbol{x} - \boldsymbol{x}'\big\| \qquad \text{per ogni } \boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X}\ .$$

Si noti che $c < \infty$ implica che η è una funzione continua. Possiamo quindi scrivere

$$\eta(\boldsymbol{x}') \leq \eta(\boldsymbol{x}) + c\,\big\|\boldsymbol{x} - \boldsymbol{x}'\big\| \tag{3}$$

$$1 - \eta(\boldsymbol{x}') \leq 1 - \eta(\boldsymbol{x}) + c\,\big\|\boldsymbol{x} - \boldsymbol{x}'\big\| \tag{4}$$

Siccome i dati sono estratti in modo indipendente, per ogni $(\boldsymbol{x}, y)$ e $(\boldsymbol{x}', y')$ vale che

$$\mathbb{P}(Y = y, Y' = y' \mid \boldsymbol{X} = \boldsymbol{x}, \boldsymbol{X}' = \boldsymbol{x}') = \mathbb{P}(Y = y \mid \boldsymbol{X} = \boldsymbol{x})\mathbb{P}(Y' = y' \mid \boldsymbol{X}' = \boldsymbol{x}') \; . \tag{5}$$

Ricordando che $\widehat{h}_S(\boldsymbol{x}) = y_{\pi_S(\boldsymbol{x})}$, notiamo che, per ogni coppia di istanze $\boldsymbol{x}, \boldsymbol{x}'$,

$$\begin{aligned}\mathbb{P}(Y \neq Y' \mid \boldsymbol{X} = \boldsymbol{x}, \boldsymbol{X}' = \boldsymbol{x}') &= \mathbb{P}(Y = 1 \mid \boldsymbol{X} = \boldsymbol{x})\mathbb{P}(Y' = -1 \mid \boldsymbol{X}' = \boldsymbol{x}') \\ &\quad + \mathbb{P}(Y = -1 \mid \boldsymbol{X} = \boldsymbol{x})\mathbb{P}(Y' = 1 \mid \boldsymbol{X}' = \boldsymbol{x}') \\ &= \eta(\boldsymbol{x})\big(1 - \eta(\boldsymbol{x}')\big) + \big(1 - \eta(\boldsymbol{x})\big)\eta(\boldsymbol{x}')\end{aligned}$$

dove la probabilità è rispetto all'estrazione di S e $(\boldsymbol{X}, Y)$ e abbiamo usato (5). Applicando le due disuguaglianze (3) e (4), possiamo allora scrivere

$$\begin{aligned}&\mathbb{E}\big[\mathrm{er}(\widehat{h}_S)\big] = \mathbb{E}\big[\mathbb{I}\{\widehat{h}_S(\boldsymbol{X}) \neq Y\}\big] \qquad \text{il secondo valore atteso è rispetto all'estrazione di } S \text{ e } (\boldsymbol{X}, Y) \\ &= \mathbb{P}\Big(\widehat{h}_S(\boldsymbol{X}) \neq Y\Big) \qquad \text{ricordando che } \mathbb{P}(A) = \mathbb{E}\big[\mathbb{I}\{A\}\big] \text{ per ogni evento } A \\ &= \mathbb{P}\big(Y_{\pi_S(\boldsymbol{X})} \neq Y\big) \qquad \text{per definizione di 1-NN} \\ &= \mathbb{E}\big[\mathbb{I}\{Y_{\pi_S(\boldsymbol{X})} \neq Y\}\big] \\ &= \mathbb{E}\Big[\mathbb{E}\big[\mathbb{I}\{Y' \neq Y\}\big|\boldsymbol{X} = \boldsymbol{x}, \boldsymbol{X}_{\pi_S(\boldsymbol{X})} = \boldsymbol{x}'\big]\Big] \qquad \text{usando } \mathbb{E}[Y] = \mathbb{E}\big[\mathbb{E}[Y \mid X]\big] \text{ che vale per ogni } X, Y \\ &= \mathbb{E}\Big[\eta(\boldsymbol{X})\big(1 - \eta(\boldsymbol{X}')\big) + \big(1 - \eta(\boldsymbol{X})\big)\eta(\boldsymbol{X}')\Big] \quad \text{dove } \boldsymbol{X}' = \boldsymbol{X}_{\pi_S(\boldsymbol{X})} \\ &\leq \mathbb{E}\Big[\eta(\boldsymbol{X})\big(1 - \eta(\boldsymbol{X})\big) + \eta(\boldsymbol{X})c\,\big\|\boldsymbol{X} - \boldsymbol{X}'\big\| + \big(1 - \eta(\boldsymbol{X})\big)\eta(\boldsymbol{X}) + \big(1 - \eta(\boldsymbol{X})\big)c\,\big\|\boldsymbol{X} - \boldsymbol{X}'\big\|\Big] \\ &\leq 2\,\mathbb{E}\Big[\eta(\boldsymbol{X})\big(1 - \eta(\boldsymbol{X})\big)\Big] + c\,\mathbb{E}\Big[\big\|\boldsymbol{X} - \boldsymbol{X}_{\pi_S(\boldsymbol{X})}\big\|\Big]\end{aligned}$$

dove i valori attesi e le probabilità sono rispetto alle estrazioni indipendenti di S e di $(\boldsymbol{X}, Y)$.

Ora ricordiamo che il rischio del classificatore Bayesiano ottimo f^* soddisfa

$$\mathrm{er}(f^*) = \mathbb{E}\Big[\min\big\{\eta(\boldsymbol{X}), 1 - \eta(\boldsymbol{X})\big\}\Big] \geq \mathbb{E}\Big[\eta(\boldsymbol{X})\big(1 - \eta(\boldsymbol{X})\big)\Big] \; .$$

Quindi abbiamo che

$$\mathbb{E}\big[\mathrm{er}(\widehat{h}_S)\big] \leq 2\,\mathrm{er}(f^*) + c\,\mathbb{E}\Big[\big\|\boldsymbol{X} - \boldsymbol{X}_{\pi_S(\boldsymbol{X})}\big\|\Big] \; .$$

Per gestire il termine contenente il valore atteso di $\big\|\boldsymbol{X} - \boldsymbol{X}_{\pi_S(\boldsymbol{X})}\big\|$ suddividiamo lo spazio delle istanze $\mathcal{X}$ in ipercubetti d-dimensionali di lato $\varepsilon > 0$ —si veda la Figura 1 per un esempio bidimensionale. Siano $C_1, \ldots, C_r$ gli ipercubetti. Ora possiamo limitare $\big\|\boldsymbol{X} - \boldsymbol{X}_{\pi_S(\boldsymbol{X})}\big\|$ distinguendo il caso in cui $\boldsymbol{X}$ appartiene un C_i in cui c'è almeno un punto $\boldsymbol{X}_t$ di S e il caso in cui $\boldsymbol{X}$ appartiene ad un C_i dove non c'è neanche un punto di S. Nel primo caso $\big\|\boldsymbol{X} - \boldsymbol{X}_{\pi_S(\boldsymbol{X})}\big\|$ è al più la lunghezza della diagonale dell'ipercubetto, cioè $\varepsilon\sqrt{d}$ —si veda la parte sinistra della Figura 1. Nel secondo caso $\big\|\boldsymbol{X} - \boldsymbol{X}_{\pi_S(\boldsymbol{X})}\big\|$ è maggiorata dalla lunghezza della diagonale di $\mathcal{X}$, cioè $2\sqrt{d}$ —si veda la parte destra della Figura 1. Quindi possiamo scrivere

$$\begin{aligned}\mathbb{E}\Big[\big\|\boldsymbol{X} - \boldsymbol{X}_{\pi_S(\boldsymbol{X})}\big\|\Big] &\leq \mathbb{E}\left[\varepsilon\sqrt{d}\sum_{i=1}^{r}\mathbb{I}\{C_i \cap S \neq \emptyset\}\mathbb{I}\{\boldsymbol{X} \in C_i\} + 2\sqrt{d}\sum_{i=1}^{r}\mathbb{I}\{C_i \cap S = \emptyset\}\mathbb{I}\{\boldsymbol{X} \in C_i\}\right] \\ &= \varepsilon\sqrt{d}\,\mathbb{E}\left[\sum_{i=1}^{r}\mathbb{I}\{C_i \cap S \neq \emptyset\}\mathbb{I}\{\boldsymbol{X} \in C_i\}\right] + 2\sqrt{d}\sum_{i=1}^{r}\mathbb{E}\left[\mathbb{I}\{C_i \cap S = \emptyset\}\mathbb{I}\{\boldsymbol{X} \in C_i\}\right]\end{aligned}$$

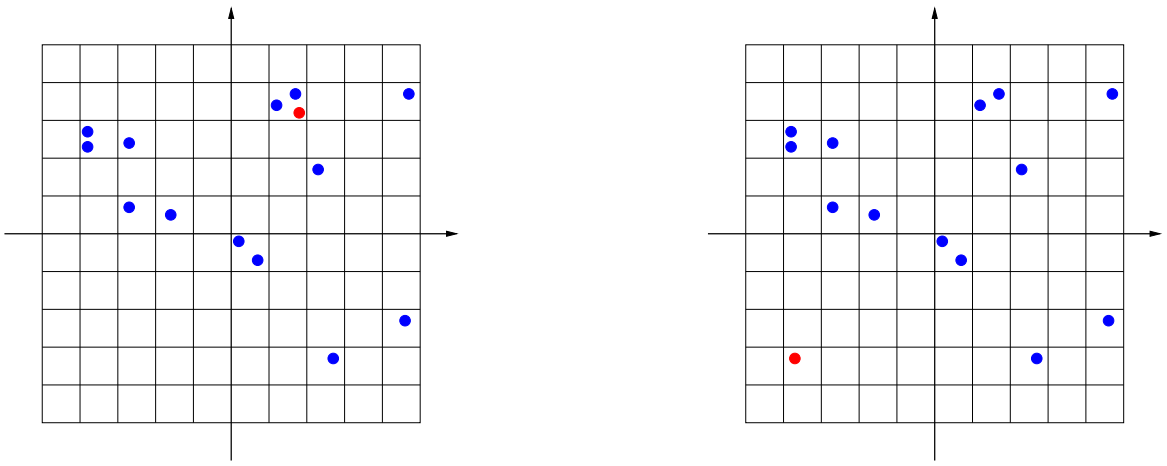

Figura 1: Esempio bidimensionale della costruzione utilizzata nell'analisi di 1-NN. Nella parte sinistra $\boldsymbol{X}$ (il punto rosso) si trova nello stesso quadratino C_i rispetto al punto $\boldsymbol{X}_{\pi_S(\boldsymbol{X})}$ del training set S più vicino ad esso e quindi $\left\|\boldsymbol{X} - \boldsymbol{X}_{\pi_S(\boldsymbol{X})}\right\|$ è limitato dalla lunghezza della diagonale del quadratino. Nella parte destra, invece, non ci sono punti di training nel quadratino di $\boldsymbol{X}$ e quindi $\left\|\boldsymbol{X} - \boldsymbol{X}_{\pi_S(\boldsymbol{X})}\right\|$ è limitato dalla lunghezza della diagonale del quadrato grande.

dove abbiamo usato la linearità del valore atteso nell'ultimo passo. Ora osserviamo che per ogni S e $\boldsymbol{X}$,

$$\sum_{i=1}^{r} \mathbb{I}\{C_i \cap S \neq \emptyset\}\mathbb{I}\{\boldsymbol{X} \in C_i\} \in \{0,1\}$$

dato che $\boldsymbol{X} \in C_i$ per un solo $i = 1, \ldots, d$. Quindi,

$$\mathbb{E}\left[\sum_{i=1}^{r} \mathbb{I}\{C_i \cap S \neq \emptyset\}\mathbb{I}\{\boldsymbol{X} \in C_i\}\right] \leq 1 \ .$$

Per l'altro termine, usiamo il fatto che il training set S e $\boldsymbol{X}$ sono indipendenti

$$\sum_{i=1}^{r} \mathbb{E}\left[\mathbb{I}\{C_i \cap S = \emptyset\}\mathbb{I}\{\boldsymbol{X} \in C_i\}\right] = \sum_{i=1}^{r} \mathbb{E}\left[\mathbb{I}\{C_i \cap S = \emptyset\}\right]\mathbb{E}\left[\mathbb{I}\{\boldsymbol{X} \in C_i\}\right] = \sum_{i=1}^{r} \mathbb{P}\left(C_i \cap S = \emptyset\right)\mathbb{P}\left(\boldsymbol{X} \in C_i\right) \ .$$

Dato che S è un campione statistico composto da m esempi indipendenti, possiamo scrivere

$$\mathbb{P}(C_i \cap S = \emptyset) = \big(1 - \mathbb{P}(\boldsymbol{X} \in C_i)\big)^m \leq \exp\big(-m\mathbb{P}(\boldsymbol{X} \in C_i)\big)$$

dove nell'ultimo passo abbiamo usato la maggiorazione elementare $(1-p)^m \leq e^{-pm}$. Ricapitolando,

$$\begin{aligned}
\mathbb{E}\Big[\left\|\boldsymbol{X} - \boldsymbol{X}_{\pi_S(\boldsymbol{X})}\right\|\Big] &\leq \varepsilon\sqrt{d} + \Big(2\sqrt{d}\Big)\sum_{i=1}^{r} \exp\big(-m\mathbb{P}(\boldsymbol{X} \in C_i)\big)\mathbb{P}\left(\boldsymbol{X} \in C_i\right) \\
&\leq \varepsilon\sqrt{d} + \Big(2\sqrt{d}\Big)\, r \max_{i=1,\ldots,r} \exp\big(-m\mathbb{P}(\boldsymbol{X} \in C_i)\big)\mathbb{P}\left(\boldsymbol{X} \in C_i\right) \\
&\leq \varepsilon\sqrt{d} + \Big(2\sqrt{d}\Big)\, r \max_{0\leq p\leq 1} e^{-pm} p \ .
\end{aligned}$$

Studiando la funzione $g(p) = e^{-pm}p$ troviamo il massimo per $p = \frac{1}{m}$. Quindi, sostituendo,

$$\mathbb{E}\Big[\big\|\boldsymbol{X} - \boldsymbol{X}_{\pi_S(\boldsymbol{X})}\big\|\Big] \leq \varepsilon\sqrt{d} + \Big(2\sqrt{d}\Big)\,\frac{r}{em} = \sqrt{d}\left(\varepsilon + \frac{2}{em}\left(\frac{2}{\varepsilon}\right)^d\right)$$

dove abbiamo usato il fatto che il numero r di ipercubetti è pari a $\left(\frac{2}{\varepsilon}\right)^d$. Mettendo tutto assieme troviamo che

$$\mathbb{E}\big[\mathrm{er}(\widehat{h}_S)\big] \leq 2\,\mathrm{er}(f^*) + c\,\sqrt{d}\left(\varepsilon + \frac{2}{em}\left(\frac{2}{\varepsilon}\right)^d\right)$$

Dato che la scelta di $0 < \varepsilon < 1$ è libera nell'analisi, possiamo porre $\varepsilon = 2\,m^{-1/(d+1)}$. Questo ci dà

$$\varepsilon + \frac{2}{em}\left(\frac{2}{\varepsilon}\right)^d = 2m^{-1/(d+1)} + \frac{2^{d+1}2^{-d}m^{d/(d+1)}}{em} = 2m^{-1/(d+1)}\left(1 + \frac{1}{e}\right) \leq 4m^{-1/(d+1)}\ .$$

Sostituendo, otteniamo infine

$$\mathbb{E}\big[\mathrm{er}(\widehat{h}_S)\big] \leq 2\,\mathrm{er}(f^*) + c\,4m^{-1/(d+1)}\sqrt{d}\ .$$

Da questa analisi possiamo trarre due conclusioni.

1. Per $m \to \infty$, $\mathrm{er}(f^*) \leq \mathbb{E}\big[\mathrm{er}(\widehat{h}_S)\big] \leq 2\,\mathrm{er}(f^*)$. Ovvero, il rischio di 1-NN è compreso fra il Bayes error e due volte il Bayes error.
2. Perché $\mathbb{E}\big[\mathrm{er}(\widehat{h}_S)\big]$ sia al più $2\,\mathrm{er}(f^*) + \varepsilon$ il training set dev'essere almeno di taglia $m \geq \left(\frac{4c}{\varepsilon}\sqrt{d}\right)^{d+1}$, cioè esponenziale nel numero d di attributi. Questo mostra che 1-NN paga un prezzo alto in termini di overfitting per potersi confrontare direttamente con il Bayes risk.

La dimostrazione può essere generalizzata per studiare, sotto le medesime assunzioni, il rischio del classificatore $\widehat{h}_S$ prodotto da k-NN, che risulta essere maggiorato come segue

$$\mathbb{E}\big[\mathrm{er}(\widehat{h}_S)\big] \leq \left(1 + \sqrt{\frac{8}{k}}\right)\mathrm{er}(f^*) + \mathcal{O}\big(k\,m^{-1/(d+1)}\big)\ .$$