(F94) Metodi statistici per l'apprendimento

Regressione logistica, funzioni surrogate e consistenza

Docente: *Nicolò Cesa-Bianchi* versione 1 giugno 2017

Come abbiamo visto in precedenza, un approccio molto seguito per affrontare i problemi di classificazione binaria è quello di apprendere un modello predittivo $g : \mathcal{X} \to \mathbb{R}$ che rappresenta il classificatore $\operatorname{sgn}\big(g(\boldsymbol{x})\big) \in \{-1, +1\}$ e poi utilizzare una funzione di perdita convessa $\ell(z)$ che maggiori la funzione indicatrice dell'errore di classificazione, $\ell(z) \geq \mathbb{I}\{z \leq 0\}$ dove $z = y\, g(\boldsymbol{x})$ e $y \in \{-1, +1\}$ è l'etichetta di $\boldsymbol{x}$. Funzioni convesse che maggiorano $\mathbb{I}\{z \leq 0\}$ vengono dette surrogate. Un esempio di funzione surrogata è la hinge loss $h(z) = [1 - z]_+$.

In molte situazioni, oltre a utilizzare $g(\boldsymbol{x})$ per classificare, lo si vorrebbe utilizzare anche come stima della probabilità $\eta(\boldsymbol{x}) = \mathbb{P}(Y = +1 \mid \boldsymbol{x})$. Un modo elegante per utilizzare una qualunque $g : \mathcal{X} \to \mathbb{R}$ per stimare η è attraverso la funzione logistica

$$\sigma(z) = \frac{1}{1 + e^{-z}}$$

dove $0 < \sigma(z) < 1$ per ogni $z \in \mathbb{R}$. La stima di $\eta(\boldsymbol{x})$ tramite $1\big/\big(1 + e^{-g(\boldsymbol{x})}\big)$ prende quindi il nome di regressione logistica.

Chiaramente, se vogliamo usare g per fare regressione logistica dobbiamo apprendere usando una funzione di perdita appropriata. Dato che vogliamo usare g per stimare una probabilità, usiamo la funzione di perdita logaritmica,

$$\ell\big(\widehat{y}, y\big) = \mathbb{I}\{y = +1\} \ln \frac{1}{\widehat{y}} + \mathbb{I}\{y = -1\} \ln \frac{1}{1 - \widehat{y}}$$

dove $\widehat{y} = \sigma\big(g(\boldsymbol{x})\big)$. Osservando che $1 - \sigma(z) = \sigma(-z)$, possiamo scrivere la perdita logaritmica in funzione di $z = y\, g(\boldsymbol{x})$ ottenendo

$$\ell\big(\widehat{y}, y\big) = \ln\big(1 + e^{-z}\big)$$

che vale quando $\widehat{y} = \sigma\big(g(\boldsymbol{x})\big)$. La parte destra dell'identità prende il nome di funzione di perdita logistica. Dato che è convessa, per renderla una funzione surrogata è sufficiente dividerla per $\ln 2$. Infatti, è facile dimostrare che

$$\frac{\ln\big(1 + e^{-z}\big)}{\ln 2} = \log_2\big(1 + e^{-z}\big) \geq \mathbb{I}\{z \leq 0\} \ .$$

Da un esame dei grafici delle due funzioni surrogate in Figura 1 si nota come la funzione logistica $\ell_{\log}(z) = \log_2\big(1 + e^{-z}\big)$ penalizza (seppure di poco) anche le classificazioni corrette, ovvero $\ell_{\log}(z) > 0$ per ogni $z \in \mathbb{R}$, mentre la funzione hinge h penalizza le classificazioni corrette soltanto quando hanno margine piccolo, ovvero $h(z) > 0$ per $z < 1$. Inoltre, la logistica penalizza molto più della hinge le classificazioni scorrette con margine grande, ovvero $\ell_{\log}(z) \gg h(z)$ quando $z \ll 0$.

Chiaramente, hinge e logistica non sono le uniche due possibili funzioni surrogate. Ma allora ci chiediamo quale sia un criterio che ci permetta di dire che alcune funzioni surrogate sono migliori

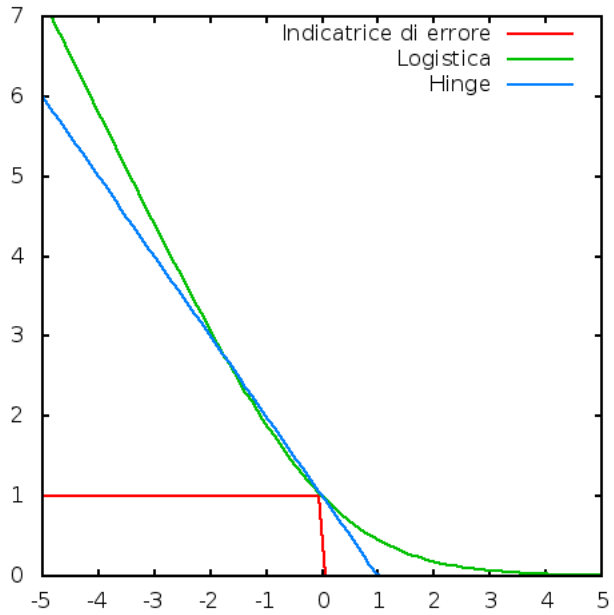

Figura 1: Confronto fra le due funzioni surrogate hinge $h(z) = [1 - z]_+$ e logistica $\ell_{\log}(z) = \log_2\big(1 + e^{-z}\big)$.

di altre. Una proprietà importante è quella di *consistenza*, che definiamo rispetto ad un modello statistico (D, η) —in realtà, come vedremo, la consistenza dipende solo da η.

Diciamo che una funzione di perdita surrogata ℓ è consistente quando, per ogni $\boldsymbol{x}$,

$$\operatorname{sgn}\big(y^*_{\boldsymbol{x}}\big) = f^*(\boldsymbol{x}) \qquad \text{per} \qquad y^*_{\boldsymbol{x}} = \operatorname*{argmin}_{\widehat{y}\in\mathbb{R}} \mathbb{E}\big[\ell(Y\widehat{y}) \mid \boldsymbol{X} = \boldsymbol{x}\big] \ . \tag{1}$$

In altre parole, il segno della predizione che minimizza il valore atteso della perdita su $\boldsymbol{x}$ dev'essere uguale alla classificazione del Bayesiano ottimo su $\boldsymbol{x}$.

Dato che $\mathbb{P}(Y = +1 \mid \boldsymbol{x}) = \eta(\boldsymbol{x})$, possiamo scrivere $\mathbb{E}\big[\ell(Y\widehat{y}) \mid \boldsymbol{X} = \boldsymbol{x}\big] = \eta(\boldsymbol{x})\ell(\widehat{y}) + \big(1-\eta(\boldsymbol{x})\big)\ell(-\widehat{y})$.

Verifichiamo ora la consistenza della funzione di perdita logistica. Non è difficile verificare che

$$y^*_{\boldsymbol{x}} = \operatorname*{argmin}_{\widehat{y}\in\mathbb{R}} \Big(\eta(\boldsymbol{x}) \log_2\big(1 + e^{-\widehat{y}}\big) + \big(1 - \eta(\boldsymbol{x})\big) \log_2\big(1 + e^{\widehat{y}}\big)\Big) = \ln \frac{\eta(\boldsymbol{x})}{1 - \eta(\boldsymbol{x})}$$

e quindi

$$\operatorname{sgn}\big(y^*_{\boldsymbol{x}}\big) = \operatorname{sgn}\left(\ln \frac{\eta(\boldsymbol{x})}{1 - \eta(\boldsymbol{x})}\right) = \operatorname{sgn}\big(\eta(\boldsymbol{x}) - \tfrac{1}{2}\big) = f^*(\boldsymbol{x}) \ .$$

La quantità $y^*_{\boldsymbol{x}} = \ln \frac{\eta(\boldsymbol{x})}{1-\eta(\boldsymbol{x})}$ si chiama *log-odds ratio*. Oltre a fornire il classificatore Bayesiano ottimo $\operatorname{sgn}\big(y^*_{\boldsymbol{x}}\big)$, $y^*_{\boldsymbol{x}}$ è anche la migliore stima di η rispetto alla funzione di perdita logistica. Se calcoliamo il rischio condizionato di $y^*_{\boldsymbol{x}}$ rispetto alla funzione logistica otteniamo

$$\mathbb{E}\Big[\log_2\Big(1 + e^{-Y\, y^*_{\boldsymbol{x}}}\Big) \,\Big|\, \boldsymbol{X} = \boldsymbol{x}\Big] = -\eta(\boldsymbol{x}) \log_2 \eta(\boldsymbol{x}) + \big(1 - \eta(\boldsymbol{x})\big) \log_2\big(1 - \eta(\boldsymbol{x})\big) \ .$$

La quantità a destra è l'entropia $H\big(Y \mid \boldsymbol{X} = \boldsymbol{x}\big)$ di Y dato $\boldsymbol{x}$ e rappresenta il numero medio di bit che ricaviamo osservando Y quando $\boldsymbol{x}$ è noto. Ciò implica che il rischio Bayesiano (cioè il rischio del Bayesiano ottimo $g^*(\boldsymbol{x}) = y^*_{\boldsymbol{x}}$) per la perdita logistica è

$$\mathbb{E}\Big[\log_2\Big(1 + e^{-Y\, g^*(\boldsymbol{X})}\Big)\Big] = H(Y)$$

dove $H(Y)$ è l'entropia della distribuzione Bernoulliana $\big\{\mathbb{P}(Y=+1), \mathbb{P}(Y=-1)\big\}$.

Verifichiamo ora che anche la hinge loss è consistente,

$$\begin{aligned}
y^*_{\boldsymbol{x}} &= \operatorname*{argmin}_{\widehat{y}\in\mathbb{R}} \Big(\eta(\boldsymbol{x})\big[1-\widehat{y}\big]_+ + \big(1-\eta(\boldsymbol{x})\big)\big[1+\widehat{y}\big]_+\Big) \\
&= \operatorname*{argmin}_{\widehat{y}\in[-1,+1]} \Big(\eta(\boldsymbol{x})\big[1-\widehat{y}\big]_+ + \big(1-\eta(\boldsymbol{x})\big)\big[1+\widehat{y}\big]_+\Big) \\
&= \operatorname*{argmin}_{\widehat{y}\in[-1,+1]} \Big(1 + \big(1-2\eta(\boldsymbol{x})\big)\widehat{y}\Big) \\
&= \begin{cases} -1 & \text{se } \eta(\boldsymbol{x}) < 1/2, \\ +1 & \text{se } \eta(\boldsymbol{x}) \geq 1/2 \end{cases} \\
&= f^*(\boldsymbol{x}) \ .
\end{aligned}$$

Nella seconda uguaglianza, abbiamo potuto sostituire $\widehat{y} \in \mathbb{R}$ con $\widehat{y} \in [-1,+1]$ in quanto entrambe le funzioni $[1-\widehat{y}]_+$ e $[1+\widehat{y}]_+$ non aumentano se tronchiamo l'argomento $\widehat{y}$ nell'intervallo $[-1,+1]$.

In generale, è possibile dimostrare il seguente risultato.

Teorema 1 *Se una funzione di perdita surrogata $\ell : \mathbb{R} \to \mathbb{R}$ è convessa, differenziabile sullo zero e tale che $\ell'(0) < 0$ allora è anche consistente nel senso di (1).*

Questo implica immediatamente che anche le seguenti funzioni di perdita surrogate sono consistenti.

- Boosting loss: $\ell(z) = e^{-z}$.
- Square loss: $\ell(z) = (1-z)^2$.
- Hinge loss quadratica: $\ell(z) = \big([1-z]_+\big)^2$.

Passiamo ora a descrivere l'algoritmo di Online Gradient Descent per regressione logistica dove $g(\boldsymbol{x})$ è un modello lineare $\boldsymbol{w}^\top \boldsymbol{x}$. Data una sequenza di esempi $(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots \in \mathbb{R}^d \times \{-1,+1\}$ definiamo $\ell_t(\boldsymbol{w}) = \log_2\big(1 + e^{-y_t \boldsymbol{w}^\top \boldsymbol{x}_t}\big)$ e calcoliamo $\nabla \ell_t(\boldsymbol{w})$. Prima di tutto, osserviamo che

$$\frac{d}{dz}\ell_{\log}(yz) = \frac{d}{dz}\log_2\big(1+e^{-yz}\big) = \frac{1}{\ln 2} \times \frac{-ye^{-yz}}{1+e^{-yz}} = \frac{1}{\ln 2} \times \frac{-y}{1+e^{yz}} = \frac{-y\sigma(-yz)}{\ln 2} \ .$$

Quindi,

$$\nabla \ell_t(\boldsymbol{w}) = \frac{d}{dz}\ell_{\log}(y_t z)\Big|_{z=\boldsymbol{w}^\top \boldsymbol{x}_t} \times \boldsymbol{x}_t = \frac{-\sigma\big(-y_t \boldsymbol{w}^\top \boldsymbol{x}_t\big)}{\ln 2} y_t \boldsymbol{x}_t \ .$$

Il passo del gradiente si può scrivere allora come

$$\boldsymbol{w}_{t+1} = \boldsymbol{w}_t + \eta_t \sigma\big(-y_t \boldsymbol{w}^\top \boldsymbol{x}_t\big) y_t \boldsymbol{x}_t$$

dove il fattore $\ln 2$ è stato incorporato nel coefficente η_t. Dato che la funzione logistica non è fortemente convessa, dovremo usare $\eta_t = \eta/\big((\ln 2)\sqrt{t}\big)$ con $\eta > 0$ e far seguire ogni passo di discesa del gradiente da una proiezione sulla sfera di raggio $U > 0$.

Se vogliamo fare regressione logistica per minimizzare il rischio empirico in un dato training set, possiamo aggiungere un termine di regolarizzazione per rendere la funzione risultante fortemente convessa e quindi convergere più velocemente tramite discesa del gradiente stocastico (e senza dover più utilizzare la proiezione sulla sfera),

$$\ell_{\mathrm{reg}}(\boldsymbol{w}) = \log_2\left(1 + e^{-y\boldsymbol{w}^\top \boldsymbol{x}}\right) + \frac{\lambda}{2}\|\boldsymbol{w}\|^2 \ .$$

L'algoritmo risultante è simile a Pegasos dove la hinge loss è sostituita dalla perdita logistica.