(F94) Metodi statistici per l'apprendimento

Regressione logistica, funzioni surrogate e consistenza

Docente: *Nicolò Cesa-Bianchi* versione 31 maggio 2016

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 $z = yg(\boldsymbol{x})$ 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 classificazione lineare, $g(\boldsymbol{x}) = \boldsymbol{w}^\top \boldsymbol{x}$ e la hinge loss prende la forma $h(\boldsymbol{w}) = \big[1 - y\boldsymbol{w}^\top \boldsymbol{x}\big]_+$. In molte situazioni, però, si vuole utilizzare $g(\boldsymbol{x})$ per modellare $\eta(\boldsymbol{x}) = \mathbb{P}(Y = +1 \mid \boldsymbol{x})$. In altre parole, oltre a utilizzare $g(\boldsymbol{x})$ come classificatore $\operatorname{sgn}\big(g(\boldsymbol{x})\big)$, lo si vuole utilizzare anche come stima della probabilità $\eta(\boldsymbol{x})$ che l'etichetta su un dato $\boldsymbol{x}$ abbia valore $+1$. Questo evidentemente richiede che g abbia codominio $[0, 1]$ e di fatto rende innaturale l'utilizzo di modelli lineari $g(\boldsymbol{x}) = \boldsymbol{w}^\top \boldsymbol{x}$ che di norma possono assume valori positivi e negativi. Una maniera elegante per aggirare questo ostacolo è porre

$$\ln \frac{\mathbb{P}(Y = +1 \mid \boldsymbol{x})}{\mathbb{P}(Y = -1 \mid \boldsymbol{x})} = \boldsymbol{w}^\top \boldsymbol{x}$$

dove la quantità a sinistra dell'uguale prende il nome di *log-odds ratio*. Risolvendo l'equazione per $\mathbb{P}(Y = +1 \mid \boldsymbol{x})$, ricordando che $\mathbb{P}(Y = -1 \mid \boldsymbol{x}) = 1 - \mathbb{P}(Y = +1 \mid \boldsymbol{x})$, otteniamo

$$\mathbb{P}(Y = +1 \mid \boldsymbol{x}) = \frac{1}{1 + e^{-\boldsymbol{w}^\top \boldsymbol{x}}} \ .$$

La quantità a destra dell'uguale si chiama funzione logistica, ha codominio $[0, 1]$ e viene comunemente indicata con

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

La regressione logistica apprende quindi un classificatore utilizzando un modello logistico $g(\boldsymbol{x}) = \sigma\big(\boldsymbol{w}^\top \boldsymbol{x}\big) \in [0, 1]$ che rappresenta la probabilità $\mathbb{P}(Y = +1 \mid \boldsymbol{x})$. Dato che la predizione è probabilistica, una funzione di perdita appropriata è quella logaritmica,

$$\mathbb{I}\{y = +1\} \ln \frac{1}{\widehat{y}} + \mathbb{I}\{y = -1\} \ln \frac{1}{1 - \widehat{y}}$$

dove $\widehat{y} = \sigma\big(\boldsymbol{w}^\top \boldsymbol{x}\big)$. Osservando che $1 - \sigma(z) = \sigma(-z)$, possiamo infine scrivere la formula della funzione di perdita logistica $\ell(z) = \ln\big(1 + e^{-z}\big)$. La funzione di perdita logistica è convessa. Per renderla una funzione surrogata, è sufficiente dividerla per $\ln 2$ in quanto 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 si nota come la funzione logistica $\ell_{\log}(z) =$

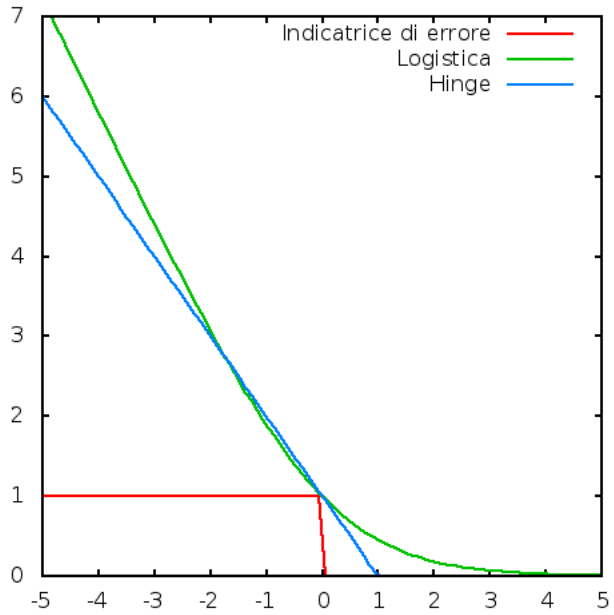

Figura 1: Confronto fra le due funzioni surrogate hinge $h(z) = [1 - z]_+$ e logistica $\ell_{\log}(z) = \log_2\left(1 + e^{-z}\right)$.

$\log_2\left(1 + e^{-z}\right)$ 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 di altre. Una prima proprietà 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 $\boldsymbol{x}$,

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

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 $\mathbb{E}\left[\ell(Y\widehat{y}) \mid \boldsymbol{X} = \boldsymbol{x}\right] = \eta(\boldsymbol{x})\ell(\widehat{y}) + \left(1 - \eta(\boldsymbol{x})\right)\ell(-\widehat{y})$.

Verifichiamo ora che 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 $[-1,+1]$.

Nel caso della funzione logistica è possibile dimostrare 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

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

Ovvero, anche la funzione logistica è consistente.

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 esplicitare l'algoritmo di Online Gradient Descent con la funzione di perdita logistica. Ovvero, 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,

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

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 $\eta_t = \eta/\sqrt{t}$ con $\eta > 0$ e far seguire ogni passo di discesa del gradiente da una proiezione sulla sfera di raggio $U > 0$.