Modelli Supervisionati in Dettaglio

Fabio Mardero

21 marzo 2019

Nel contesto di un modello ad apprendimento supervisionato, un dataset $\mathcal{D}$ può essere definito come le N determinazioni (i *datapoint*) di un campione di una coppia di variabili aleatorie (X, Y), le coppie del campione essendo indipendenti. La prima variabile è di tipo multidimensionale e rappresenta i dati di input, detti anche covariate o *features*; Y è invece la variabile target, chiamata anche variabile risposta o di output[1]. Si ha

$$\mathcal{D} = \left\{ (\boldsymbol{x}_i, y_i) \right\}_{i=1}^{N} \quad \text{con} \quad \boldsymbol{x}_i \in \mathcal{X}, y_i \in \mathcal{Y}, \tag{1}$$

dove il datapoint $(\boldsymbol{x}_i, y_i)$ in $\mathcal{D}$ è determinazione di $(\boldsymbol{X}_i, Y_i)$ con funzione di ripartizione congiunta $F_{\boldsymbol{X},Y}$. Si assume senza perdita di generalità, inoltre, che Y assuma valori in $\mathcal{Y}$, un intervallo reale. Si fa riferimento quindi ad un problema di regressione. Si ottengono i medesimi risultati nel caso di uno di classificazione, basta infatti interpretare $\mathcal{Y}$ come un insieme discreto.

L'obiettivo è quello di studiare la "relazione" che intercorre tra le due variabili aleatorie $\boldsymbol{X}$ e Y, ed in particolare di individuare una funzione in grado di approssimare $F_{Y|\boldsymbol{X}=\boldsymbol{x}}$ per ogni $\boldsymbol{x} \in \mathcal{X}$ o una sua qualche proprietà / parametro, dove $F_{Y|\boldsymbol{X}=\boldsymbol{x}}$ è la funzione di ripartizione di Y condizionata a $\boldsymbol{X} = \boldsymbol{x}$. Risulta infatti che, fissato $\boldsymbol{X} = \boldsymbol{x}$, Y non è funzione deterministica di $\boldsymbol{x}$, ma è una variabile aleatoria . In particolare il problema può essere affrontato secondo un punto di vista frequentista o bayesiano. Nel seguito sarà esposto il modello frequentista.

Si suppone di poter esprimere Y come somma di una componente deterministica, che coinvolge $\boldsymbol{x}$, e una stocastica ε, detta rumore,

$$Y = f(\boldsymbol{x}) + \varepsilon, \quad \boldsymbol{x} \in \mathcal{X}. \tag{2}$$

Fissata una determinazione $\boldsymbol{x}$, Y è stocastica in quanto frutto dell'aleatorietà di ε, coerentemente con quanto descritto sopra. Si sta infatti sommando un termine certo ad una variabile aleatoria. Da notare che ε può dipendere anch'essa dalle covariate: in questo caso si dice che il campione di variabili aleatorie è eteroschedastico; in caso contrario si parla di omoschedasticità. Con f si indica invece una funzione, detta *ipotesi induttiva*, tale che dato un input $\boldsymbol{x} \in \mathcal{X}$, restituisce un valore $\hat{y} \in \mathcal{Y}$.

$$\begin{aligned} f \colon \mathcal{X} &\longrightarrow \mathcal{Y} \\ \boldsymbol{x} &\longmapsto f(\boldsymbol{x}) = \hat{y} \end{aligned} \tag{3}$$

In particolare, ricordando (2), e ponendo ε a media nulla ($\mathbb{E}(\varepsilon) = 0$) è possibile scrivere

$$\mathbb{E}\left[Y|\boldsymbol{X} = \boldsymbol{x}\right] = f(\boldsymbol{x}), \quad \text{per ogni } \boldsymbol{x} \in \mathcal{X}. \tag{4}$$

[1]Nel testo si farà riferimento alle determinazioni delle variabili aleatorie indicandole con le rispettive lettere minuscole.

Se a priori non si ritiene di assegnare una distribuzione di probabilità all'errore, allora $F_{Y|\boldsymbol{X}=\boldsymbol{x}}$ rimane ignota, pur potendo in questo caso determinare la sua speranza matematica tramite (4). Analoghe considerazioni possono essere fatte per altre proprietà della distribuzione.

La funzione f può essere parametrica o non parametrica. Si dice che f è parametrica se essa è definita tramite un numero finito di coefficienti la cui numerosità non dipende dalla grandezza del dataset $\mathcal{D}$. Al contrario una funzione non parametrica "modifica" la sua struttura proporzionalmente al numero di datapoint disponibili (ad esempio $f(\boldsymbol{x}_k) = \frac{1}{N-1}\sum_{i=1}^{N} \alpha_i\, e^{-\boldsymbol{x}_i^2}\, \mathbb{I}_{i\neq k}$ possiede N parametri[2]). Si considereranno funzioni parametriche, le quali per maggiore chiarezza saranno indicate con $f(\boldsymbol{x};\boldsymbol{\theta})$ con $\boldsymbol{\theta}$ il vettore dei parametri. Al variare di $\boldsymbol{\theta}$ si ottiene la famiglia di funzioni

$$\mathcal{F} = \left\{ f\left(\boldsymbol{x};\boldsymbol{\theta}\right);\ \boldsymbol{\theta} \in \Theta \right\}, \quad \text{per ogni } \boldsymbol{x} \in \mathcal{X} \tag{5}$$

detta *spazio d'ipotesi*, con Θ *spazio parametrico*. Dato M il numero dei parametri, generalmente $\Theta \subseteq \mathbb{R}^M$.

Sia invece L (la *loss function*) una funzione in grado di misurare lo "scostamento" (l'errore / *loss*) tra due valori y_1, y_2 in $\mathcal{Y}$.

$$\begin{aligned} L\colon \mathcal{Y}\times\mathcal{Y} &\longrightarrow \mathbb{R} \\ (y_1, y_2) &\longmapsto L(y_1, y_2) \end{aligned} \tag{6}$$

Fissato un datapoint $(\boldsymbol{x}, y) \in \mathcal{D}$, applicando L è quindi possibile misurare l'errore puntuale tra la previsione $\hat{y} = f(\boldsymbol{x};\boldsymbol{\theta})$ e y:

$$L(\hat{y}, y) = L\left(f(\boldsymbol{x};\boldsymbol{\theta}), y\right), \quad \text{per ogni } (\boldsymbol{x}, y) \in \mathcal{D}.$$

Tramite L, si può anche valutare l'errore atteso ("*expected loss*", detto anche funzione di rischio o *risk function*) di una funzione $f(\ ;\boldsymbol{\theta})$[3] su tutte le possibili determinazioni di $(\boldsymbol{X}, Y)$.

$$\begin{aligned} R(\boldsymbol{\theta}) &= \mathbb{E}\left[L\left(f(\boldsymbol{X};\boldsymbol{\theta}), Y\right)\right] \\ &= \int L\left(f(\boldsymbol{x};\boldsymbol{\theta}), y\right)\, dF(\boldsymbol{x}, y) \end{aligned} \tag{7}$$

Date f e L per calcolare la (7) si deve applicare la $F_{\boldsymbol{X},Y}$, la legge di generazione dei dati. Se tale legge non è nota, risulta possibile approssimarla con

$$F_{emp}(\boldsymbol{x}, y) = \frac{1}{N}\sum_{i=1}^{N} I_{(\boldsymbol{x}\leq\boldsymbol{x}_i, y\leq y_i)},$$

con I_A funzione indicatrice dell'evento $A = (\boldsymbol{x} \leq \boldsymbol{x}_i, y \leq y_i)$. Si ottiene quindi che

$$R(\boldsymbol{\theta}) \approx R_{emp}(\mathcal{D}, \boldsymbol{\theta}) = \frac{1}{N}\sum_{i=1}^{N} L\left(f(\boldsymbol{x}_i;\boldsymbol{\theta}), y_i\right). \tag{8}$$

[2] Se $\alpha_i = 1$, $i = 1, \ldots, N$, la funzione si dice comunque non parametrica. Si osservi, infatti, che la "struttura" di f dipende da N: la numerosità degli addendi è pari a quella dei datapoint.

[3] Nell'ambito della famiglia di funzioni $\mathcal{F}$, il parametro $\boldsymbol{\theta}$ definisce univocamente la funzione. Per questo, per brevità, dove il contesto lo rende chiaro, specificando $\boldsymbol{\theta}$ si intenderà la funzione ad esso associata.

detta funzione di rischio empirica. Esempi di funzioni di rischio empiriche (dipendenti dalla scelta della loss function) sono lo scarto quadratico medio (MSE: *mean squared error*)

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2,$$

lo scarto medio assoluto (MAE: *mean absolute error*)

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|\hat{y}_i - y_i|$$

e lo scarto medio assoluto percentuale (MAPE: *mean absolute percentile error*)

$$MAPE = \frac{1}{N}\sum_{i=1}^{N}\frac{|\hat{y}_i - y_i|}{y_i}.$$

I principi utilizzati per ricavare (8) rientrano nella branca della statistica legata al machine learning chiamata *empirical risk minimization* (ERM).

Dato un dataset $\mathcal{D}$, si definisce modello parametrico ad apprendimento supervisionato la famiglia di funzioni $\mathcal{F}$, con le f definite in (3) e con proprietà (4), utilizzata per calcolare (8) sui datapoint al fine di ricercare i parametri $\hat{\boldsymbol{\theta}}$ tali per cui

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}\in\Theta} R_{emp}(\mathcal{D}, \boldsymbol{\theta}). \tag{9}$$

Proprio per (9) si parla di apprendimento supervisionato: il modello impara dai dati (che comprendono una variabile target) allo scopo di selezionare il migliore set di parametri in termini di funzione di rischio empirica. Si ricorda inoltre che la teoria della empirical risk minimization qui esposta si basa principalmente sull'approssimazione (8). Le capacità del modello di spiegare il fenomeno dipendono molto dalla quantità di dati disponibili: maggiore il numero dei datapoint collezionati, meglio potrà essere descritta la componente deterministica del fenomeno. Non a caso il machine learning, specie per i modelli più complicati, si dice essere *data hungry*, cioè "affamato di dati".

Dopo aver scelto un particolare modello di machine learning, cioè una famiglia di funzioni $\mathcal{F}$, il processo volto alla ricerca del minimo della risk function si dice allenamento e comprende le fasi di stima e validazione del modello.