Machine Learning per la fisica applicata e la fisica delle alte energie

Lezione 5: Modelli lineari - regressione

Come introdotto nelle lezioni precedenti, lo scopo del machine learning (supervised) consiste nel determinare il mapping:

$$y = f(\vec{\vartheta}/x)$$

Una classe di modelli per f molto importante è quella dei modelli lineari in cui f è una funzione lineare nei parametri $\vec{\vartheta}$:

$$f(\vec{\vartheta}/x) = \beta_0 + \sum_{j=1}^{p} x_j \beta_j \qquad \vec{\vartheta} = \{\beta_0, \beta_1, \dots, \beta_p\} \quad j = 1, \dots, p$$

N.B. Le x_i possono essere: input quantitativi; la loro trasformazione ($\log$, $\sqrt{\ }$, quadrato); un'espansione in una base p. es. di tipo polinomiale come $x_2 = x_1^2$; $x_3 = x_1^3$; "interazioni" tra variabili, p. es. $x_3 = x_1 \cdot x_2$.
Dato un data set di training $(x_1, y_1); \dots; (x_N, y_N)$ i parametri β vengono tipicamente stimati col metodo dei minimi quadrati

$$RSS(\beta) = \sum_{i=1}^{N} \left(y_i - f(x_i)\right)^2 \qquad \text{RSS: residual square sum}$$

$$= \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij} \beta_j\right)^2$$

La RSS altro non è che la NLL per variabili casuali INDIPENDENTI distribuite in modo gaussiano. Riscriviamo la $RSS(\beta)$ come

$$RSS(\beta) = (\vec{y} - X\beta)^T (y - X\beta)$$

e differenziamo

$$\begin{cases} \dfrac{\partial RSS}{\partial \beta} = -2X^T(\vec{y} - X\beta) \\[2ex] \dfrac{\partial^2 RSS}{\partial\beta\,\partial\beta^T} = 2X^T X \end{cases}$$

$$\frac{\partial RSS}{\partial \beta} = 0 \iff X^T(\vec{y} - X\beta) = 0$$

$$\iff \hat{\beta} = (X^T X)^{-1} X^T \vec{y}$$

Le predizioni per un generico input vector x_0 sono date da

$$\hat{f}(x_0) = (1, \dots, x_0)^T \hat{\beta} \implies \hat{\vec{y}} = X\hat{\beta} = X(X^TX)^{-1}X^T\vec{y} \quad \text{con } \hat{y}_i = \hat{f}(x_i)$$

<u>Osservazione</u>: può capitare che le colonne di X non siano linearmente indipendenti, quindi X non è full rank. Questo accade se due degli input sono perfettamente correlati ($x_2 = 3x_1$). In tal caso X^TX è singolare e i coefficienti $\hat{\beta}$ non sono definiti. Occorre allora pre-trattare X, per esempio rimuovendo le colonne linearmente dipendenti. Questo è fatto automaticamente negli algoritmi correnti.

Supponiamo ora che le y_i siano variabili non correlate, che abbiano varianza costante e che le x_i siano fissate. La matrice di varianze è $\mathrm{Var}(\hat{\beta}) = (X^TX)^{-1}\sigma^2$

Assumiamo ora che il modello lineare sia il modello corretto per il valor medio, cioè il valore di aspettazione condizionale per Y è lineare in $X_1, \dots, X_p$. Assumiamo inoltre che le deviazioni di Y intorno al suo valore di aspettazione siano gaussiane e additive. Allora

$$y_p = \beta_0 + \sum_{j=1}^{p} x_j \beta_j + \varepsilon \qquad \text{con } \varepsilon \sim N(0, \sigma^2)$$

Segue che

$$\hat{\beta} \sim N(\beta, (X^TX)^{-1}\sigma^2)$$

Supponiamo di testare l'ipotesi per cui un particolare coefficiente $\beta_j = 0$. Si definisce lo Z-score

$$z_j = \frac{\hat{\beta}_j}{\hat{\sigma}\sqrt{v_j}}$$ dove v_j è il j-esimo elemento diagonale di $(X^TX)^{-1}$

Sotto l'ipotesi nulla che $\beta_j = 0$, allora z_j è distribuito come t_{N-p-1} e quindi un grande valore di z_j conduce a rigettare l'ipotesi. Se $\hat{\sigma}$ è sostituito da un valore noto di σ, allora z_j è distribuito gaussiano=

mente. La differenza nei quantili nelle coda della distribuzione t e delle gaussiana diventa trascurabile per N grande. L'intervallo di credibilità per $\hat{\beta}$ è:

$$\left[\hat{\beta}_j - z^{(1-\alpha)} v_j^{1/2}\hat{\sigma}, \; \hat{\beta}_j + z^{(1-\alpha)} v_j^{1/2}\hat{\sigma}\right]$$

dove $z^{(1-\alpha)}$ è il percentile $(1-\alpha)$ della distribuzione gaussiana.

5.1. Il teorema di Gauss-Markov

La stima ottenuta con il metodo dei minimi quadrati per i parametri β ha la varianza minore tra tutti gli stimatori lineari unbiased. Definiamo $\vartheta = a^T\beta$. Il metodo dei minimi quadrati su $a^T\hat{\beta}$

$$\hat{\vartheta} = a^T\hat{\beta} = a^T(X^TX)^{-1}X^T\vec{y}$$

Se consideriamo X una matrice fissata, dev'essere una funzione lineare $c_0^T\vec{y}$. Se il modello lineare è corretto, $a^T\hat{\beta}$ è unbiased poiché

$$\mathbb{E}(a^T\hat{\beta}) = \mathbb{E}(a^T(X^TX)^{-1}X^T\vec{y}) = a^T(X^TX)^{-1}X^TX\beta = a^T\beta$$

Il teorema di Gauss-Markov afferma che, se abbiamo qualunque altro stimatore $\tilde{\vartheta} = c^T\vec{y}$ che è unbiased per $a^T\beta$, cioè $\mathbb{E}(c^T\vec{y}) = a^T\beta$, allora $\mathrm{Var}(a^T\hat{\beta}) \leq \mathrm{Var}(c^T\vec{y})$. La dimostrazione segue dalla disuguaglianza triangolare.

Consideriamo l'errore quadratico medio di uno stimatore $\tilde{\vartheta}$ che stima ϑ

$$\mathrm{MSE}(\tilde{\vartheta}) = \mathbb{E}(\tilde{\vartheta} - \vartheta)^2 = \mathrm{Var}(\tilde{\vartheta}) + \left[\mathbb{E}(\tilde{\vartheta}) - \vartheta\right]^2$$

Il teorema di Gauss-Markov implica che lo stimatore ottenuto con il metodo dei minimi quadrati ha il più piccolo valore $\mathrm{Var}(\hat{\vartheta})$ e bias nullo. Tuttavia potrebbe esistere uno stimatore con $\mathrm{Var}(\hat{\vartheta})$ più piccola e bias tale che $\mathrm{MSE}(\tilde{\vartheta})$ sia complessivamente minore.

Consideriamo ora di voler fare una predizione

$$y_0 = f(x_0) + \varepsilon_0$$

Allora l'errore nella predizione di $\tilde{f}(x_0) = x_0^T \tilde{\beta}$ è:

$$\mathbb{E}(y_0 - \tilde{f}(x_0))^2 = \sigma^2 + \mathbb{E}(x_0^T \tilde{\beta} - f(x_0))^2 = \sigma^2 + MSE(\tilde{f}(x_0))$$

La differenza è la sola σ^2, che è la varianza della nuova osservabile

5.2 SUBSET SELECTION

Ci sono due ragioni per cui spesso non siamo soddisfatti del metodo dei minimi quadrati.

1) Il primo è l'accuratezza delle predizioni. Si può sacrificare il bias in favore di una varianza più piccola.

2) Il secondo è l'interpretabilità. Con un gran numero di predittori, vogliamo spesso determinare un sottoset piccolo che esibisca gli effetti più rilevanti. In altri termini, siamo disposti a sacrificare alcuni dei dataset più piccoli.

5.2.1 BEST SUBSET SELECTION

Si tratta di trovare $\forall\, k \in \{0, 1, \dots, p\}$ il sottoset di dimensione k che minimizza $RSS(\vec{\vartheta})$. La domanda è come scegliere k in modo da ottimizzare il trade-off tra bias e variance. L'idea è di usare il cross-validation per stimare l'errore nella predizione e selezionare k.

- forward stepwise selection: si parte con l'intercetta e poi si aggiunge sistematicamente il predittore che migliora il fit. Si produce perciò una sequenza di modelli, con indice k, che va determinato. Vantaggi: 1) COMPUTAZIONALE: per grandi p non si può calcolare la sequenza migliore ($p \geq 40$). 2) STATISTICO: varianza minore, bias maggiore.

- backward stepwise selection: si parte con il modello completo e si rimuove in maniera sequenziale il predittore che ha l'impatto minore nel fit (valutato sulla base del suo Z-score)

5.3 SHRINKAGE METHODS

L'idea è quella di rendere il processo (di selezione dei parametri) continuo.

5.3.1 RIDGE REGRESSION

Il metodo consiste nel ridurre i coefficienti di regressione imponendo una penalità sulla loro dimensione. I "ridge coefficients" minimizzano una RSS "penalizzata"

$$\hat{\beta}^{ridge} = \underset{\beta}{\operatorname{argmin}} \left\{ \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j\right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\}$$

$\lambda \geq 0$ è un parametro che controlla quanto "shrinkage" si applica ai parametri; maggiore è λ, più i parametri sono "schiacciati" a zero.
In forma matriciale, ottengo:

$$RSS(\lambda) = (\vec{y} - X\beta)^T(\vec{y} - X\beta) + \lambda\beta^T\beta \Longleftrightarrow \hat{\beta}^{ridge} = (X^TX + \lambda\mathbb{1})^{-1}X^T\vec{y}$$

5.3.2. LASSO

Il metodo è un metodo di shrinkage simile al precedente definito come

$$\hat{\beta}^{lasso} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j\right)^2 \quad \text{con} \quad \sum_{j=1}^{p} |\beta_j| \leq t$$

$$\hat{\beta}^{lasso} = \underset{\beta}{\operatorname{argmin}} \left\{ \frac{1}{2} \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j\right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}$$

La differenza sta nel fatto che la "ridge penalty" L_2 è stata sostituita con la lasso penalty L_1. Questo implica che le soluzioni siano non-lineari in y. Ponendo t sufficientemente piccolo, il metodo realizza una sorta di selezione continua sui parametri del modello.

- RIDGE REGRESSION: shrinkage proporzionale
- LASSO: ogni coefficiente è traslato di λ

Ridge regression e lasso possono essere generalizzati come

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \left\{ \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j\right)^2 + \lambda \sum_{j=1}^{p} |\beta_j|^q \right\} \quad q \geq 0$$

Si può pensare a $|\beta_j|^q$ come la log-prior density per β_j o come i contorni della distribuzione a priori dei parametri.

$q=0$ subset selection; $q=1$ lasso; $q=2$ ridge regression sono tutti Bayes estimates con un prior diverso.