Smoothing Splines

Vito Simone Lacatena 747810

3/7/2021

Introduzione

Prima di entrare nel cuore dell'algoritmo occorre introdurre dei concetti fondamentali.

Il problema della regressione

In generale un problema di regressione ha lo scopo di stimare un modello che descriva uan relazione tra una **variabile di risposta** Y dipendente è un insime di **variabili predittive** indipententi $X_1, ..., X_p$.

Il più semplice modello di regressione è il modello di **regressione lineare** in cui si assume una relazione lineare tra una singola variabile predittiva X e la variabile di risposta, tale funzione lineare viene approssimata come segue: Y

$$Y = \beta_0 + \beta_1 X + \epsilon$$

I parametri β_0 e β_1 sono sconosciuti e vanno stimati utilizzando i dati. Per determinare il valore ottimale di questi parametri si può usare il criterio dei **minimi quadtrati**, ovvero scegliere le stime dei parametri $\hat{\beta}_0$ e $\hat{\beta}_1$ in modo da minimizzare la **somma dei residui al quadrato**, definita come :

$$RSS = \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{N} (y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))^2$$

Regressione Polinomiale

Il modello di regressione polinomiale estende il modello di regressione lineare attraverso l'aggiunta di ulteriori variabili ottenuti dalla variabile origianli elevandoli ad una potenza. Ovvero si estendo il modello lineare classico

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

con la seguente funzione polinomiale:

$$y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_i^2 + ... + \beta_d x_i^d + \epsilon_i$$

Funzioni a scalini

Tali funzioni dividono l'intervallo dei valori in K bins distinti creando un insieme di valori di soglia $c_1, c_2, ..., c_k$ nell'intervallo e per ogni bin si va ad adattare una funzione costante, Questo equivale a trasformare una variabile continua in una variabile categorica ordinale, avremo quindi:

$$y_i = \beta_0 + \beta_1 C_1(x_i) + \beta_2 C_2(x_i) + ... + \beta_k C_k(x_i) + \epsilon_i$$

dove C_j con $j = 1...k$ è una funzione che assume valore 1 se $c_{j-1} < x_i < c_j$ altrimenti 0.

Funzioni base

I modelli di regressione polinomiali e funzoni a scalini sono casi particolari di un framework generico di funzioni base, in cui si va ad adattare il seguente modello:

$$y_i = \beta_0 + \beta_1 b_1(x_i) + \beta_2 b_2(x_i) + ... + \beta_k b_k(x_i) + \epsilon_i$$

dove $b_j(.)$ sono funzioni base:

- regressione polinomiale : $b_j(x_i) = x_i^j$;
- funzioni a scalini : $b_j(x_i) = C_j(x_i)$;

Splines

Invece di adattare un polinomio di grado elevato su l'intero intervallo , l'idea base dell'approccio Spline è quello di utilizzare la regressione polinomiali attratti che comporta l'inserimento di polinomi di grado inferiore su diversi sottointervalli, Esempio se si considerano polinomi di terzo grado e si divide l'intervallo originale X sul punto c, avremmo due cubiche con differenti coefficenti:

$$y_i = \begin{cases} \beta_{01} + \beta_{11} x_i + \beta_{21} x_i^2 + \beta_{31} x_i^3 + \varepsilon_i & \text{se } x_i < c \\ \beta_{02} + \beta_{12} x_i + \beta_{22} x_i^2 + \beta_{32} x_i^3 + \varepsilon_i & \text{se } x_i \geq c \end{cases}$$

In questo caso si adatterano due diversi polinomi con i propri coefficenti, il punto c dell'intervallo in cui si passa da un polinomio ad un altro è chiamatato **nodo** (o **knot**).

Osservazione: con K nodi si avranno $K + 1$ polinomi.

I **gradi di libertà** per questo esempio sono 8, per gradi di libertà si intende il numero di parametri indipendenti da stimare che vengono utilizzati dal modello, nell'esempio precedente il numero di gradi di libertà è quattro per ogni polinomio (poichè di grado 3), quindi in tutto otto.

Basi della Spline

Una regressione spline può essere rappresentata in termini di basi di funzioni.

Caso semplice: regressione spline lineare con grado $d = 1$ e $K = 1$ nodi

$$y_i = \beta_0 + \beta_1 b_1(x_i) + \beta_2 b_2(x_i) + \epsilon_i$$

dove $b_1(x) = x$ e successivamente si usano funzioni di base troncate

$$h(x, c) = (x - c)_+ = \begin{cases} (x - c) & \text{se } x > c \\ 0 & \text{se } x \leq c \end{cases}$$

Se $x_i \leq c$ allora $(x_i - c)_+ = 0$ e $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$

Se $x_i > c$ allora $(x_i - c)_+ = (x_i - c)$ e $y_i = \beta_0 + \beta_1 x_i + \beta_1 (x_i - c) + \epsilon_i$

In generale: regressione spline di grado d e K nodi

$y_i = \beta_0 + \beta_1 b_1(x_i) + \beta_2 b_2(x_i) + \beta_K b_K(x_i) + \beta_{K+1} b_{K+1}(x_i) + ... + \beta_{K+d} b_{K+d}(x_i) + \epsilon_i$

le funzioni di base sono

$$x, x^2, ..., x^d, h(x, c_1), ..., h(x, c_K)$$

dove in questo caso

$h(x, c) = (x - c)_+^d = (x - c)^d$ se $x > c$ altrimenti 0.

Un caso interessante è quello delle **spline cubiche**, ritenute popolari poichè difficile individuare la discontinuità ai nodi, una spline cubica con K nodi utilizza $K + 4$ gradi di libertà. Le Spline di grado superiore ad secondo mostrano un elevata variabilità agli estremi dell'intervallo, una **spline naturale** è una spline con vincoli di linearità agli estremi.

Struttura dell'algoritmo

Smoothing Splines:Panoramica

L'obiettivo della regressione è adattare una funzione $f(x)$ ad un set di dati tale che minimizzi la **somma dei quadtrati del residui**

$$RSS = \sum_{i=1}^{n} (y_i - f(x_i))^2$$

In questo modo si potrebbe semplicemente ottenere un RSS pari a zero scegliendo la funzione f che interpoli perfettamente tutti i puntii y_i, ma tale risultato sarebbe troppo soggetto ad overfitting poichè la funzione sarebbe troppo flessibile. Occorre quindi trovare il modo di smussare la funzione, ridefinendo la RSS nel seguente modo:

$$RSS(f, \lambda) = \sum_{i=1}^{n} (y_i - f(x_i))^2 + \lambda \int f''(t)^2 dt$$

La funzione f è definita **smoothing spline**.

La prima parte di RSS definisce una misura di distanza dai dati mentre la seconda parte definisce una penalità che penalizza la variabilità della funzione f , la quantità $\lambda \int f''(t)^2 dt$ è quindi una misura del cambiamento globale nella funzione $f'(t)$

- se f è smussata $\int f'(t)$ sarà costante e $\int f''(t)$ avrà un valore molto piccolo;
- se f è irregolare $\int f'(t)$ sarà molto variabile e $\int f''(t)$ avrà un valore molto grande;

Il parametro λ è un **parametro di smoothing** che porterà f ad essere smussata, il parametro assume valori compresi in $(0, \infty)$:

- Con valore $\lambda = 0$ i suoi effetti sono nulli e quindi la funzione assumerà un comportamento molto irregolare;
- con $\lambda = \infty$ la funzione sarà una linea retta che passa il più vicino possibile ai dati.

Quidi grandi valori di λ producono curve smussata, mentre piccoli valori di questo parametro producono curve più irregolari.

Si consideri il vettore n-dimensionale $\hat{f}_\lambda$ contenenti i valori adattati ai dati di training $x_1, ..., x_n$, nonchè soluzione di $RSS(f, \lambda)$ per un determinato valore di λ. Questo vettore può essere scritto come:

$$\hat{f}_\lambda = S_\lambda y$$

Dove la matrice di dimensione $n \times n$ S_λ è nota come **smoother matrix**, e il vettore y e il **vettore di risposta**, il numero dei gradi di libertà corrisponde alla traccia della matrice S_λ.

$$df_\lambda = trace(S_\lambda)$$

Selezione Automatica dei parametri di Smoothing

In generale i parametri di smoothing da stimare per la regressione spline sono:

- il grado delle spline;
- il numero e la posizione dei nodi.

In pratica si può però dimostrare che la funzione $f(x)$ che minimizza la RSS è una **spline cubica naturale** avente nodi ad ogni punto $x_1, ...x_n$, questo significa che per quanto riguarda lo smoothing splines si ha un solo parametro λ da stimare poichè i nodi corrispondono ai punti di training e si utilizzano polinomi di grado 3.

Occorre notare che all'aumentare del valore di λ da 0 a ∞ il numero di gradi di libertà descresce da n a 2.

Uno dei metodi per valutare la scelta del parametro λ è la cross-validation, in particolare l'errore di cross-validation Leave One Out (LOOCV) viene calcolato con la seguente formula:

$$LOOCV(\hat{f}_\lambda) = \frac{1}{N}\sum_{i=1}^{N}(\frac{y_i - \hat{f}_\lambda(x_i)}{1 - S_\lambda ii})^2$$

In questa formula $\hat{f}_\lambda$ indica la funzione smoothing spline adattata su tutti i dati di training tranne x_i e $S_\lambda ii$ l'elemento diagonale i-esimo della matrice S_λ.

Caso Multidimensionale

Sino ad ora si è considerato il caso unidimensionale dell'aproccio smoothing splines,ma tale metodo può essere generalizzato anche per un numero maggiore di dimensioni. Data la coppia y_i, x_i , con $y_i \in \mathbb{R}$ e $x_i \in \mathbb{R}^d$, e una funzione di regressione d-dimensionale $f(x)$, si consideri il seguente problema così impostato:

$$\min_f \sum_{i=1}^{N}\{y_i - f(x_i)\}^2 + \lambda J[f]$$

dove J è una funzione di penalità appropriata per stabilizzare una funzione f in $\mathbb{R}^d$. Esempio in $\mathbb{R}^2$:

$$J[f] = \int\int_{\mathbb{R}^2}\left[\left(\frac{\partial^2 f(x)}{\partial x_1^2}\right)^2 + 2\left(\frac{\partial^2 f(x)}{\partial x_1 \partial x_2}\right)^2 + \left(\frac{\partial^2 f(x)}{\partial x_2^2}\right)^2\right] dx_1 dx_2.$$

Utilizzando per l'otimizzazione questa penalità si ottiene una superficie bidimensionale smussata, nota come spline a superficie sottile. Condivide molte proprietà con la smoothing spline unidimensionale.

- Per $\lambda \to 0$ si ottiene come soluzione una funzione di interpolazione;
- Per $\lambda \to \infty$ la soluzione converge al piano dei minimi quadtrati;

Implementazioni

Illustrare le principali implementazioni disponibili in R;

Esempio con dati reali

presentare un esempio realistico in R.