Alma Mater Studiorum - Università di Bologna

Relazione per il corso di Data Science

Liam Cavini
Semestre Invernale 2024/2025

3° Foglio, Regressione Lineare e Modelli
30/10/2024

Risorse

Il codice utilizzato, insieme al file .tex di questo documento, possono essere trovati nella seguente repository github: https://github.com/LazyLagrangian/data_science.

Esercizio 1 - Regressione con funzioni di base Gaussiane

L'esercizio ha lo scopo di compiere un fit di un polinomio tramite regressione lineare, utilizzando un modello della forma:

$$f(x) = \theta_0 + \sum_{i=1}^{n} \theta_i \phi_i(x, \mu_i, \alpha)$$

dove:

$$\phi_i(x, \mu_i, \alpha) = \exp(-\frac{(x-\mu_i)^2}{\alpha})$$

Le variabili θ_i, con $i \in \{0, 1, \cdots\}$, sono i parametri da determinare tramite la regressione, mentre i valori di α[1], μ_i e n devono essere definiti prima di eseguire il fit.

Si è scelto come polinomio:

$$p(x) = -0.1x^5 - 0.4x^4 + 1.2x^3 + x^2 - 2.3x$$

I dati sono stati ottenuti campionando $p(x)$ nell'intervallo $[-1.5, 3.0]$, ed è stato aggiunto un termine casuale per simulare del rumore. Sono stati generati in totale 100 datapoints, di cui 60 sono stati usati nel dataset di train e 40 in quello di test.

Si sono compiute regressioni per vari valori di α e n, con lo scopo di valutare quali parametri risultassero ottimali, mentre i μ_i sono stati scelti equispaziati nell'intervallo $[-1.5, 3.0]$. Il coefficiente di determinazione in funzione di questi parametri è riportato in figura 1.

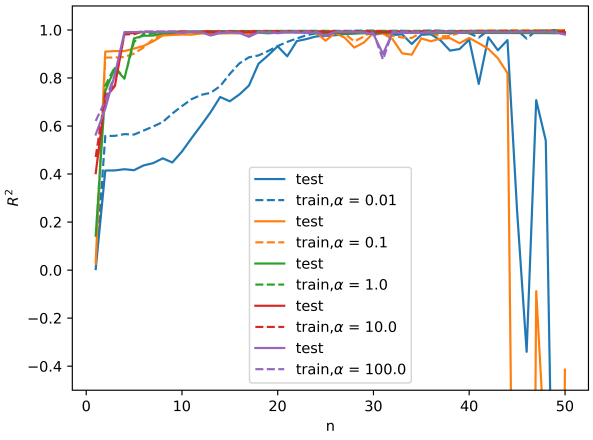

Figura 1: *sulle ascisse il numero di funzioni ϕ_i, in ordinata il coefficiente di determinazione R^2.*

[1] In questo esercizio, a ogni funzione ϕ_i è associato un parametro α_i distinto; tuttavia, si è scelto di fissare un valore comune per tutti i parametri α_i, riducendoli a un unico α.

Si osserva che per valori sufficientemente alti di α si ottiene un buon fit. Per valori bassi di α e alti di n invece si riscontrano problemi di overfitting. Questo comportamento è mostrato in figura 2.

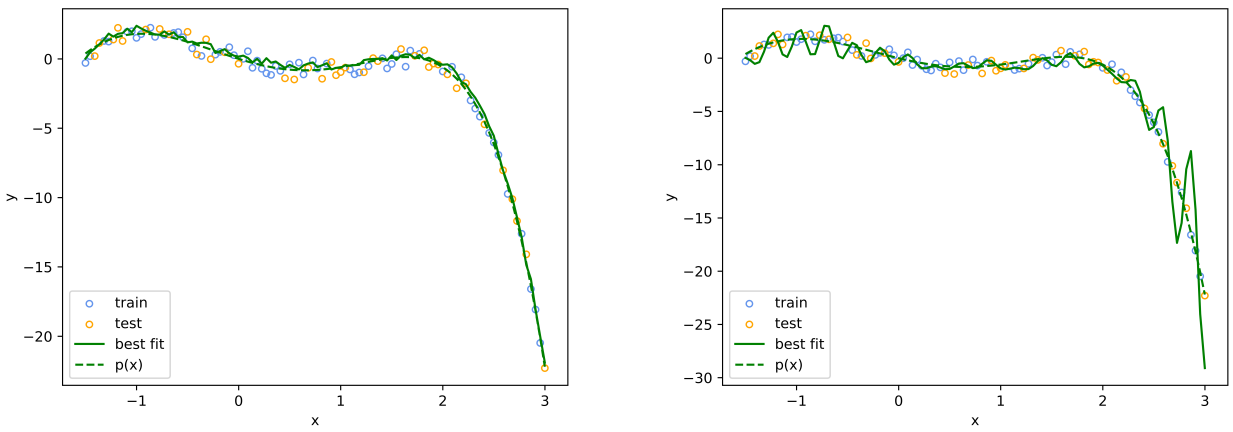

Figura 2: *Due fit dei dati generati, il grafico a sinistra mostra il fit con parametri $\alpha = 100$ e $n = 10$, il grafico di destra con parametri $n = 18$ e $\alpha = 0.01$. Il primo di questi risulta essere un buon fit, mentre il secondo non predice correttamente i dati di test.*

Esercizio 2 - Regressione lineare e modello di Ising

L'esercizio consiste nell'applicare la regressione lineare per trovare i corretti coefficienti della hamiltoniana di un modello 1D di Ising con interazioni a primi vicini. La regressione lineare è stata implementata in tre modi distinti: tramite il metodo non regolarizzato dei minimi quadrati, tramite la regolarizzazione L1, e tramite la regolarizzazione L2.

Un singolo sistema consiste di L diversi elementi, e ciascuno di questi ha associato un valore S_i che può essere 1 o -1. La hamiltoniana del sistema è calcolata tramite la formula:

$$H = -\sum_{i=1}^{L} S_i S_{i+1}$$

dove con S_{L+1} si indica S_0 (si considera il sistema come periodico).

Con questa formula si sono generati i dati utilizzati nella regressione, mentre il modello utilizzato nella regressione associa a ciascun sistema la hamiltoniana:

$$H = \sum_{i=1}^{L}\sum_{j=1}^{L} J_{ij} S_i S_j$$

Dove i J_{ij} sono i parametri da determinare tramite la regressione.

Ci aspettiamo dalla formula della hamiltoniana usata per generare i dati, che soltanto i termini della forma $S_i S_{i+1}$ e $S_{i+1} S_i$ abbiano coefficiente J_{ij} ottenuto dalla regressione non nullo, e che $J_{i,i+1} + J_{i+1,i} = -1$.

Si possono disporre i parametri J_{ij} in una matrice $L \times L$:

$$\mathbf{J} := \begin{bmatrix} J_{1,1} & \cdots & J_{1,L} \\ & \vdots & \\ J_{L,1} & \cdots & J_{L,L} \end{bmatrix}$$

L'introduzione di questa matrice è volta soltanto a migliorare la presentazione dei dati ottenuti dalla regressione. Questi sono riportati in figura 3. Come si può osservare dalla figura, il metodo non regolarizzato fallisce nel trovare i coefficienti corretti, la matrice di Gram infatti risulta essere singolare. I metodi regolarizzati invece trovano dei

coefficienti corretti per certi valori del parametro della regressione α (come quelli scelti per la realizzazione della figura 3).

La figura 4 mostra l'evoluzione dei parametri e il coefficiente di determinazione R^2 al variare di α.

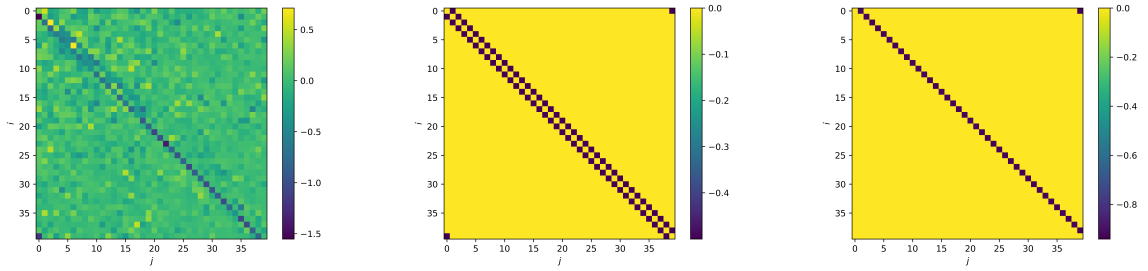

Figura 3: *Le tre immagini mostrano la matrice* $\boldsymbol{J}$ *rappresentata sul piano x-y. Ogni cella indica un elemento della matrice, mentre il colore indica il valore numerico, che si ricava consultando la legenda a destra di ciascuna immagine. Il grafico a sinistra mostra la* $\boldsymbol{J}$ *ottenuta tramite il metodo dei minimi quadrati non regolarizzato, il grafico centrale quella ottenuta dalla regolarizzazione L2, e il grafico a destra quella ottenuta dalla regolarizzazione L1.*

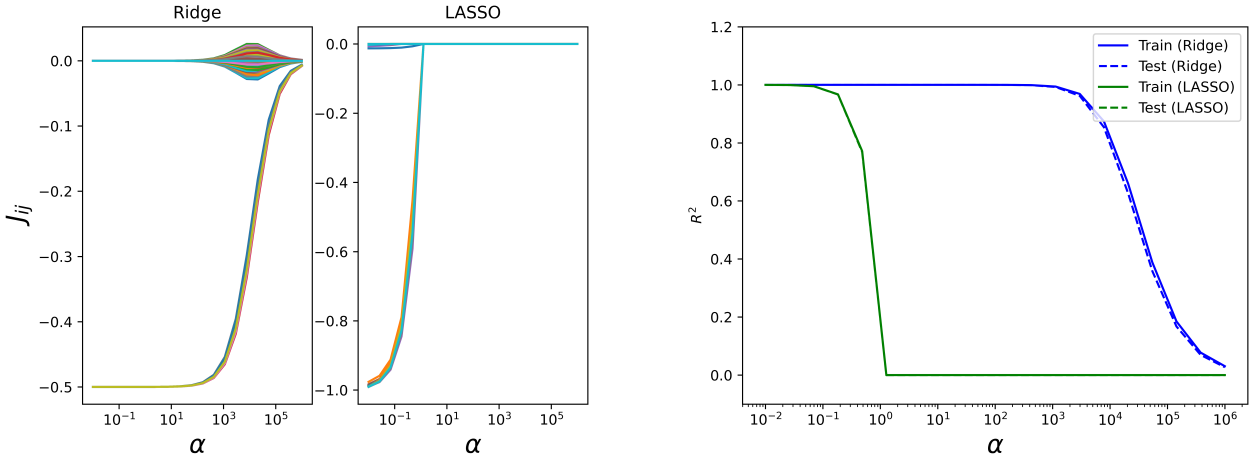

Figura 4: *A sinistra il valore dei parametri in funzione di* α*, a destra la performance(*R^2*) in funzione di* α*. La performance del dataset di test nel caso della regolarizzazione Lasso(L1) non è visibile in quanto coincide quasi con quella del training dataset.*