Alma Mater Studiorum - Università di Bologna

**Esercizi per il corso di Data Science - Laurea in Scienza dei Materiali**

Prof. D. Di Sante, Dr. A. Consiglio **3° Foglio, Regressione Lineare e Modelli**
Semestre Invernale 2024/2025 23/10/2024

**Esercizio 1 - Regressione con funzioni di base Gaussiane**

La regressione lineare richiede che i pesi si presentino esclusivamente come coefficienti lineari. Tuttavia, i termini di cui i pesi sono coefficienti possono assumere forme arbitrarie. La regressione col metodo del nucleo (kernel) è un metodo per stimare la relazione tra una variabile dipendente e una o più variabili indipendenti utilizzando una funzione kernel. Questa è una funzione che misura la somiglianza tra due punti in un dato spazio. Per eseguire questa regressione, è necessario scegliere una funzione kernel e un parametro di larghezza di banda. Il metodo utilizza la funzione kernel per trasformare i dati in uno spazio diverso in cui è possibile applicare un metodo lineare.
Una tipica scelta è data dalla funzione Gaussiana:

$$K(x, \mu) = \exp\left(-\frac{(x-\mu)^2}{\alpha}\right)$$

$\mu$ è la posizione del centro e $\alpha$ determina la deviazione standard.
Per brevità scriviamo:

$$\phi_1(x_i) = K(x_i; \mu_1, \alpha_1),\ \ldots\ ,\ \phi_n(x_i) = K(x_i; \mu_n, \alpha_n)$$

L'ipotesi per i pesi $\mathbf{w}$ è dunque:

$$h(x_i; \mathbf{w}) = [1 \quad \phi_1(x_i)\ \ldots\ \phi_n(x_i)] \cdot [w_0 \quad w_1 \ldots w_n]^T$$

Questa formulazione può essere usata nel contesto della regressione lineare.

**(a)** Si generi un insieme di dati da un polinomio e si vada a sovrapporre del rumore random. Ad esempio:

$$x \in [-1.5, 3.0], \quad y_{\text{vera}} = -0.1x^5 - 0.4x^4 + 1.2x^3 + x^2 - 2.3x$$

e

$$y = y_{\text{vera}} + 1.5 * (\text{np.random.rand(X.size)} - 0.5)$$

**(b)** Si definisca il numero di funzioni kernel che volete utilizzare, la loro posizione a la loro ampiezza. In seguito, si chiami la funzione che popola la "matrice delle features" con i dati valutati dalle funzioni kernel e si calcolino i pesi utilizzando la soluzione analitica in forma chiusa.

**(c)** Si visualizzino i dati di training e le predizioni, analizzando i risultati rispetto al numero di funzioni di base e rispetto al valore della deviazione standard delle Gaussiane. A tal proposito, si discuta il sottoadattamento e il sovradattamento (underfitting e overfitting) dei dati.

**Esercizio 2 - Regressione lineare e modello di Ising**

Vogliamo applicare ora la regressione lineare a un esempio familiare della meccanica statistica: il modello di Ising.
Consideriamo dunque il modello 1D di Ising con interazioni a primi vicini:

$$H[S] = -J \sum_{j=1}^{L} S_j S_{j+1}$$

Stiamo considerando una catena di lunghezza $L$ (ad esempio $L = 40$), ove sono presenti delle condizioni periodiche al contorno (il termine con indice $L + 1$ coincide con il termine avente indice 0). $S_j = \pm 1$ sono le variabili di spin. Essendo in una dimensione, questo modello non presenta una transizione di fase a temperatura finita.
$H$ sta ad indicare la Hamiltoniana del problema, che potete pensare come l'energia totale del sistema; in effetti, per ogni sito della catena, essa dipende contemporaneamente dalla configurazione dello spin presente sul sito $j$ e dalla sua interazione con il sito primo vicino alla sua destra avente indice $j + 1$.

**(a)** Si metta $J = 1$ e si calcoli un gran numero $n \sim 10000$ configurazioni di spin e le loro corrispondenti energie di Ising. Da qui si ottiene un insieme di dati di $i = 1, ..., n$ punti della forma $(H[\mathbf{S}^i], \mathbf{S}^i)$ per ciascuna delle $i$ configurazioni di spin $\mathbf{S}^i$.

**(b)** Vogliamo riformulare il problema di Ising come una regressione lineare. Prima di tutto, dobbiamo decidere quale classe di modello utilizzeremo per adattare i dati. Supponendo di essere in assenza di qualsiasi conoscenza preliminare, una scelta sensata è data dal modello "completo" ove le interazioni vanno oltre i primi vicini:

$$H[\mathbf{S}^i] = -\sum_{j=1}^{L} \sum_{k=1}^{L} J_{j,k} S_j^i S_k^i$$

Si noti che questo modello è definito univocamente dalle forze di accoppiamento non locali $J_{j,k}$ che vogliamo imparare. È importante sottolineare che questo modello è lineare in $\mathbf{J}$, il che rende possibile utilizzare la regressione lineare.
Per applicare la regressione lineare, si inizi con il riformulare questo modello nella forma

$$H^i_{\text{model}} = \mathbf{X}^i \mathbf{J} = X^i_p J_p, \qquad\qquad p = \{j, k\}$$

dove i vettori $\mathbf{X}^i$ rappresentano tutte le interazioni a due corpi $\{S_j^i S_k^i\}_{j,k=1}^{L}$, e l'indice $i$ corre sui campioni della banca dati.

**(c)** Si applichino al problema il metodo dei minimi quadrati (soluzione analitica non regolarizzata), la regolarizzazione L1 (Lasso) e la regolarizzazione L2 (ridge regression). Si discutano criticamente i risultati ottenuti.