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
5° Foglio, Softmax e PCA
Semestre Invernale 2024/2025
13/11/2024

Esercizio 1 - Funzione Esponenziale Normalizzata (Softmax) e banca dati MNIST

Il problema di classificazione MNIST è uno dei classici problemi di apprendimento automatico per l'apprendimento della classificazione su dati ad alta dimensione con un numero elevato di esempi.
In questa banca dati, ogni cifra scritta a mano è disponibile in un'immagine quadrata in scala di grigi sotto forma di una griglia di pixel da 28×28. Ogni pixel assume un valore compreso nell'intervallo $[0, 255]$, che rappresenta una delle 256 sfumature del colore grigio.

(a) Si scarichi il dataset MNIST, che contiene immagini di cifre (0-9) scritte a mano in formato 28×28 pixel, e lo si trasformi in un vettore di 784 (28×28) caratteristiche per ogni immagine.
Per scaricare i dati MNIST da https://www.openml.org/d/554 è possibile utilizzare il comando fetch_openml di sklearn.datasets:
X, y = fetch_openml('mnist_784', version=1, return_X_y=True)

(b) Si mescolino i dati in ordine casuale per garantire che il training e il test set siano bilanciati e si selezionino 50.000 campioni come insieme di training e 10.000 come insieme di test.

(c) Si Applichi 'StandardScaler' per normalizzare le caratteristiche, in modo che abbiano media zero e varianza unitaria, condizione importante per la convergenza dell'algoritmo di regressione.

(d) Si crei e si addestri un modello di regressione logistica utilizzando il metodo del gradiente stocastico (sag) per l'ottimizzazione.

(e) Per valutare il modello si calcoli la sparsità, ossia la percentuale di pesi non nulli. Si calcoli inoltre l'accuratezza del modello sull'insieme di test.

(f) Si visualizzino ora i pesi appresi dal modello, per comprendere come il modello distingue le diverse cifre.
Per fare ciò si copino i pesi dal modello per poterli modificare senza alterare il modello originale. Successivamente si calcoli la scala per visualizzare i pesi come immagini, determinando il valore massimo da usare per normalizzare la visualizzazione dei pesi (scale = np.abs(coef).max()).
Per ciascuna delle 10 classi di cifre (0-9), crea un sottografico visualizzando i pesi come un'immagine 28×28, che rappresenta i pesi del modello come intensità di pixel.

Esercizio 2 - Modello di Ising e analisi delle componenti principali

Facendo riferimento all'esercizio sul modello di Ising 2D della precedente sessione di laboratorio, vogliamo vedere ora come raggruppare il set di dati, per poi visualizzare i risultati utilizzando l'analisi delle componenti principali.

(a) Il primo passo nella PCA consiste nello stabilire la matrice di covarianza del DataSet. La matrice di covarianza mostrerà se le caratteristiche sono correlate positivamente o negativamente. E, poiché le caratteristiche sono regolarizzate, mostra anche quanto esse siano fortemente correlate.

(b) Dopo aver ottenuto la matrice di covarianza, vogliamo ottenere i suoi autovettori. Per fare ciò, si utilizzi la decomposizione ai valori singolari della matrice DataSet, o la diagonalizzazione diretta della matrice covariante. Pertanto, per ridurre un sistema da n-dimensioni a k-dimensioni, basta prendere i primi k-vettori dalla matrice V (prime k colonne).

(c) Per ottenere il sistema ridotto si utilizzi una trasformazione del tipo: $X_{\text{ridotto}} = X.V_{\text{ridotto}}$
Si proiettino i dati nelle prime due componenti, il che significa $D = 2$. Successivamente si raffigurino in un grafico i dati dimensionalmente ridotti. Si colorino i dati in base alla temperatura dello stato di Ising associato.
Quanti ammassi di punti riuscite a visualizzare, e a che cosa corrispondono fisicamente?

(d) Si ripeta l'analisi utilizzando la funzione PCA dal pacchetto sklearn.

(e) Come sono distribuite le componenti del primo vettore PCA?