Alma Mater Studiorum - Università di Bologna

Relazione per il corso di Data Science

Liam Cavini
Semestre Invernale 2024/2025

5° Foglio, Softmax e PCA
19/11/2024

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 - Funzione Esponenziale Normalizzata (Softmax) e banca dati MNIST

L'esercizio consiste nell'allenare un regressore lineare sul dataset MNIST, che consiste in cifre numeriche, da 0 a 9, scritte a mano su griglie di pixel 28×28. Il nostro classificatore, se il training ha successo, deve quindi riuscire a riconoscere con un accuratezza sufficiente le 10 cifre numeriche.

Come è solito si è diviso il dataset in training e test, affidando alla libreria sklearn il compito di randomizzare i due batch. Sempre usando sklearn, ed in particolare il metodo di regressione logistica 'sag', si è addestrato il modello. La sparsità (la percentuale dei pesi non nulli) è mostrata in tabella 1.

cifra	sparsità
0	0.3648
1	0.2997
2	0.4349
3	0.3980
4	0.2755
5	0.3520
6	0.3827
7	0.3673
8	0.3546
9	0.2857

Tabella 1: *La tabella mostra nella seconda colonna le sparsità arrotondate alla quarta cifra significativa, nella prima la cifra a cui sono associati i pesi.*

I pesi relativi a ciascuna cifra sono stati visualizzati su una matrice 28×28 in figura 1. Si osserva che in alcuni casi i pesi più rilevanti tracciano in maniera distinguibile la cifra corrispondente.

Esercizio 2 - Modello di Ising e analisi delle componenti principali

L'esericizo consiste nel compiere l'analisi dei componenti pricipali al modello di Ising bidimensionale discusso nel laboratorio precendente. Prima di discutere i risultati dell'analisi, consideriamo i risulati attesi in linea teorica.

Possiamo immaginare ogni sistema come un punto in uno spazio di dimensione $n = 1600$. Ogni coordinata può prendere il valore 1 o -1, dunque i punti saranno disposti sui vertici di un ipercubo $[-1, 1]^n$. I sistemi con temperatura bassa hanno tutti i valori uguali: di conseguenza si devono trovare sul vertice $(1, \cdots, 1)$ oppure $(-1, \cdots, -1)$. Osserviamo che la retta passante per questi due vertici comprende anche l'origine, ed è quindi un asse, che denotiamo $\hat{n}$. Dato che buona parte dei sistemi si trovano su questi due vertici, è ragionevole supporre che una notevole parte della varianza sia disposta lungo questo asse.

Ad alte temperature ciascuno dei 1600 valori è -1 oppure 1 con uguale probabilità. I vertici su cui si collocano i punti corrispondenti a questi sistemi sono quindi molteplici, e nessun singolo asse li comprende tutti.

Possiamo quindi fare le seguenti predizioni:

- $\hat{n}$ è un asse principale
- la varianza lungo $\hat{n}$ è superiore a quella lungo gli altri assi.

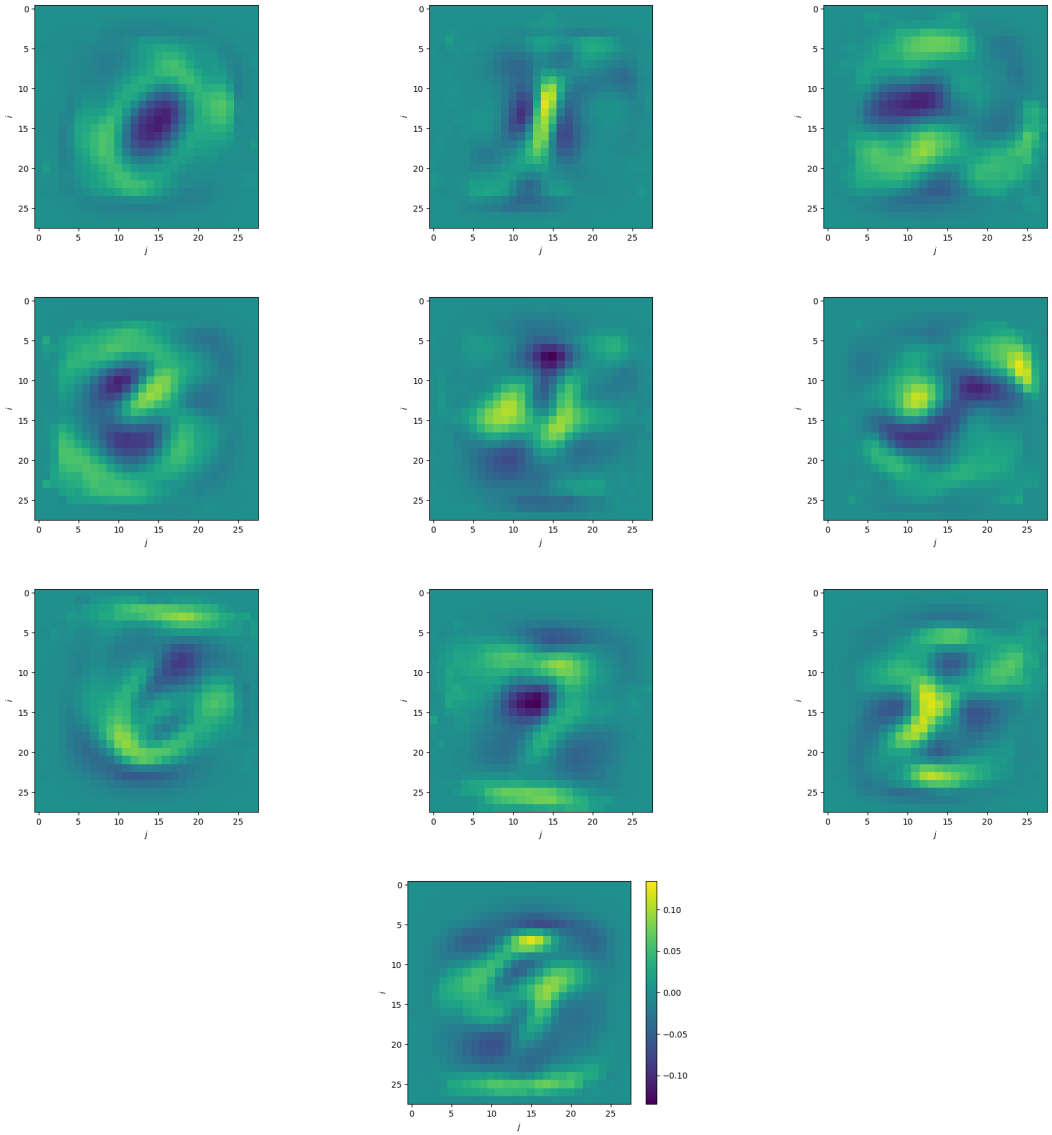

Figura 1: *La figura mostra i pesi di ciascuna classe, da 0 a 9 (partendo da in alto a sinistra), su una griglia* 28×28. *I valori numerici corrispondenti ai colori sono consultabili dalla legenda a fianco dell'ultima immagine.*

- I sistemi a basse temperature sono disposti lungo le estremità di questo asse, mentre quelli ad alte temperature vicino all'origine.

I risultati delle analisi confermano queste predizioni. Infatti $\hat{n}$ risulta essere il primo asse principale, ed è responsabile per circa il 50% della varianza, mentre il secondo asse principale è responsabile per circa il 0.6%. La terza predizione trova conferma in figura 2.

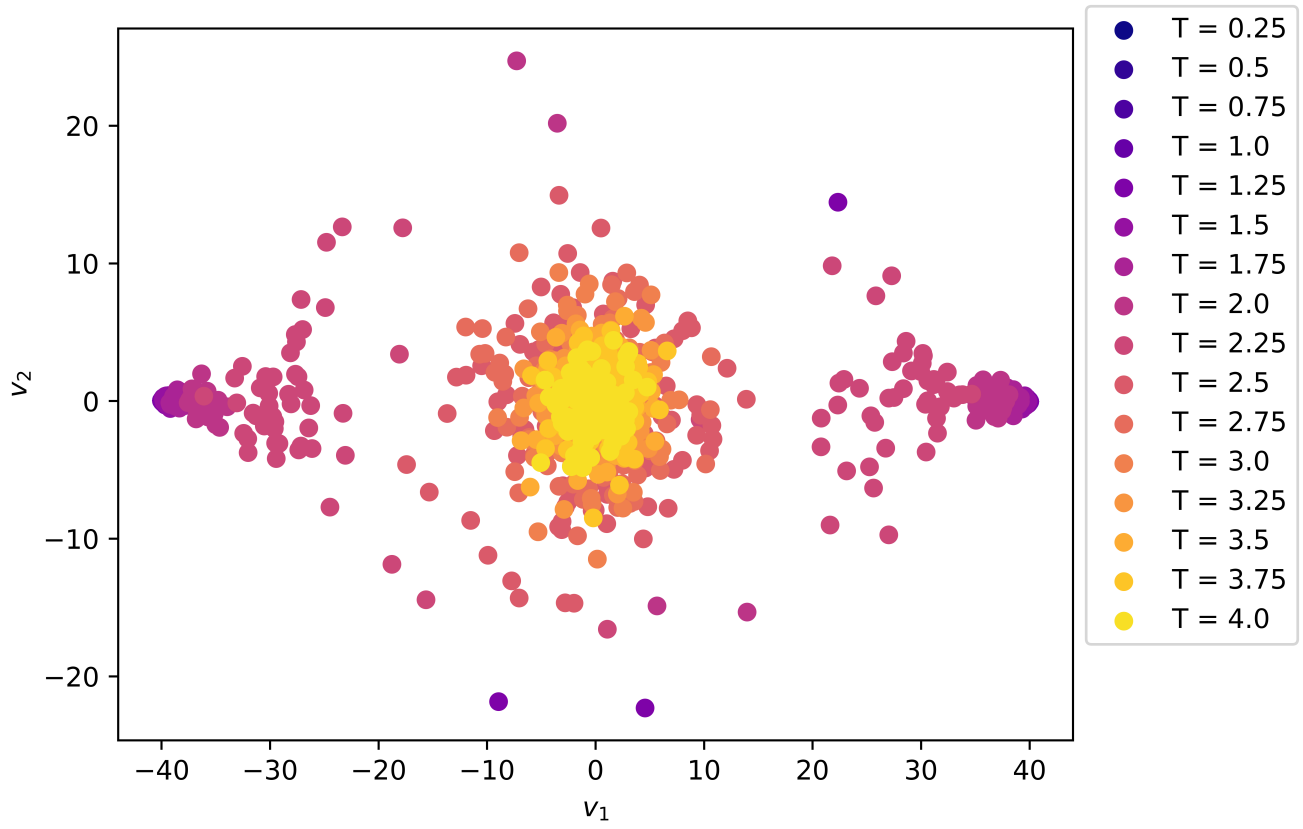

Figura 2: *La figura mostra la distribuzione dei sistemi lungo i primi due assi principali (con l'asse x come primo), e le rispettive temperature. I valori numerici corrispondenti ai colori sono consultabili dalla legenda a fianco.*