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
6° Foglio, Dati ad alta dimensione
Semestre Invernale 2024/2025
20/11/2024

Esercizio 1 - Auto-facce e algoritmo PCA

Si applichi il metodo delle autofacce (eigenfaces) per rappresentare una foto in primo piano di un volto come combinazione lineare delle autofacce estratte da un set di dati.
Si utilizzino le tecniche di riduzione della dimensionalità come l'Analisi delle Componenti Principali (PCA) per svolgere l'esercizio.

(a) Per prima cosa si recuperi il dataset di volti fornito su Virtuale, comprendendo il formato delle immagini e la loro organizzazione, aiutandovi con il notebook jupyter.

(b) Caricate il dataset e appiattite le immagini da matrici a vettori.

(c) Si utilizzi la PCA per decomporre il dataset in autofacce. Successivamente, visualizzate le prime k autofacce per interpretarne il significato.

(d) Si carichi una nuova foto di un volto in primo piano. Per prima cosa, la si renda compatibile con la base della banca dati.
Successivamente, proiettate l'immagine nello spazio delle autofacce per calcolare i coefficienti di espansione lineare. Ricostruite l'immagine a partire da un numero k di autofacce e confrontatela con l'immagine originale. Variate k per analizzare l'accuratezza della ricostruzione.

(e) Valutate la perdita di informazione in base al numero di autofacce utilizzate. Create un grafico che mostri la differenza tra immagine originale e ricostruzione al variare di k.

(f) Si espanda l'immagine di un altro (s)oggetto, ad esempio una pianta, nella base delle autofacce. Quante autofacce sono necessarie per una rappresentazione accurata?

Esercizio 2 - Metodi PCA e t-SNE per la banca dati MNIST

Facendo riferimento al precedente foglio di esercizi, vogliamo utilizzare il set di dati MNIST. Ogni immagine è composta da 784 pixel (disposti in matrici di 28×28 pixel) con ciascun pixel corrispondente a un valore intero in scala di grigi compreso tra 0 e 255, dove 0 è bianco e 255 è nero.

(a) Si applichi l'Analisi delle Componenti Principali (PCA) per riuscire a ottenere un'immagine, utilizzando una frazione del numero di dimensioni (ovvero significativamente inferiore delle 784 variabili pixel).

(b) Si mostri quanta varianza nel set di dati originale è spiegata dalle prime k componenti principali.

(c) Si utilizzi la PCA per visualizzare i dati originali ad alta dimensionalità in due dimen-

sioni, per vedere se eventuali agglomerati di dati sono chiaramente visibili.

(d) Infine, utilizzando l'analisi t-SNE sia sul dataset originale e sia sui risultati dimensionalmente ridotti derivanti dalla PCA (considerare 40 componenti PCA), si proiettino in due dimensioni i dati del MNIST. Come si confrontano questi risultati, con quelli ottenuti esclusivamente attraverso la PCA?

Esercizio 3 - Metodo t-SNE e modello di Ising

Si ripeta l'analisi del modello di Ising 2D del precedente foglio di esercizi, questa volta utilizzando il metodo t-SNE. Come si confrontano i risultati con quelli ottenuti tramite PCA?