Università degli Studi di Verona

“Face Clustering”

Fondamenti di Machine Learning

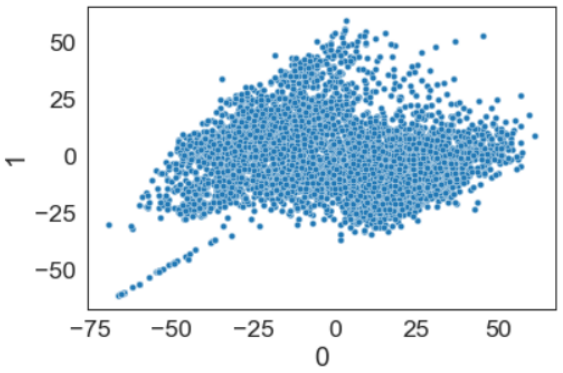
A.A. 2021/2022

Elisa Acciari - VR478828

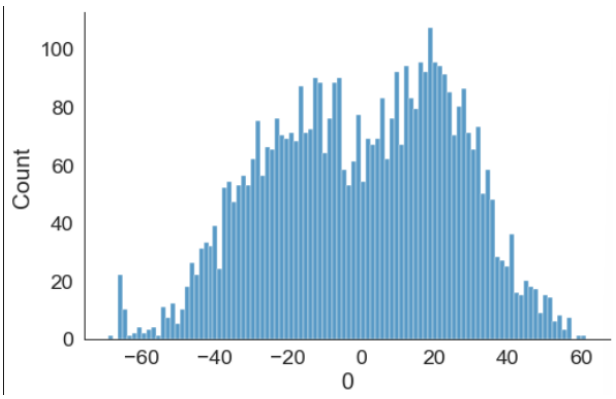
Dato un set di immagini di volti, clusterizzare e identificare le classi.
In quanto si andrà a fare Clustering, verranno trattati dati non etichettati procedendo con un Learning Non Supervisionato.

Pre-elaborazione dei dati

I dati si presentavano grezzi e poco lavorabili in quanto l'immagine comprendeva tutta la figura della persona ed erano presenti molte immagini con la figura rivolta di spalle o in cui il viso non era visibile.
Quindi, per quanto possibile, grazie ad un file di keypoints presente nel file da scaricare, si sono andate a fare delle operazioni di ritaglio andando ad intercettare i punti che contornavano il viso.
Mentre si sono andate ad eliminare le immagini che presentavano un dimensione uguale a zero dopo il ritaglio, in quanto non presentavano punti che delimitavano il viso.
Si è andato anche ad applicare il filtro sulle immagini per convertirle in scala di grigio.



Per comprendere meglio l'andamento dei dati, si è andato a plottare il seguente istogramma



Possiamo notare due picchi distinti alla coordinata x=-20 e x=20, che rappresenteranno le medie di due classi.

Feature Extraction

Una volta completata la prima fase, si è passato all'estrazione delle feature per ridurre la dimensionalità delle immagini e compattando l'informazione.
Tale procedimento è stato fatto attraverso PCA.

Andando a plottare i dati sul nuovo sistema di coordinate, osserviamo i vari addensamenti di punti che si creano per capire in generale quanti cluster avrà il nostro dataset.

Training

Per fare l'addestramento si è andato innanzitutto a splittare il dataset in dati di training e testing.
Successivamente si è passato alla fase di addestramento svolto con KMeans su un diverso numero di cluster.

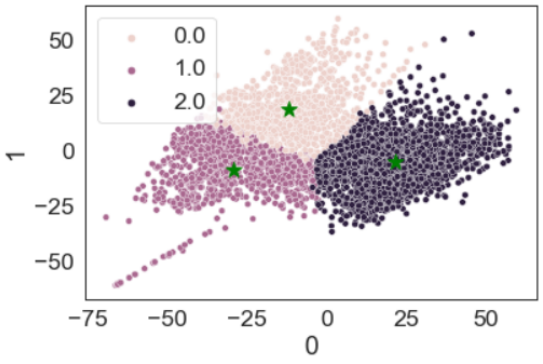
Testing

La fase successiva all'addestramento è il testing, in cui si vanno a testare nuovi dati.

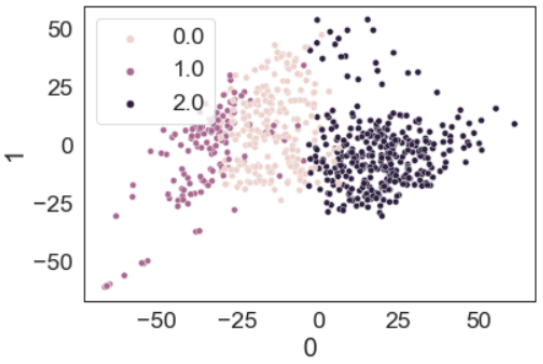
Si andranno a testare questi dati sia con la regola di Bayes che con Knn, andando poi a plottare i punti di testing colorati in base all'etichetta a loro assegnata e facendo un confronto visivo con i dati clusterizzati da KMeans. Andando a clusterizzare con K=2, K=3 e K=4 , si osserva che la procedura che meglio si avvicina alla clusterizzazione con KMeans è quella di Bayes. Mostriamo un esempio.

L'Immagine n. 1 mostra il train set clusterizzato e plottato con gli indici calcolati da KMeans.
L'immagine n. 2 rappresenta il test set indicizzato con le etichette calcolate da Knn.
L'immagine n. 3 rappresenta invece il test set indicizzato con le etichette calcolate con la regola di Bayes.
Si nota con evidenza che il terzo plot è quello che si avvicina di più al primo di Kmeans.

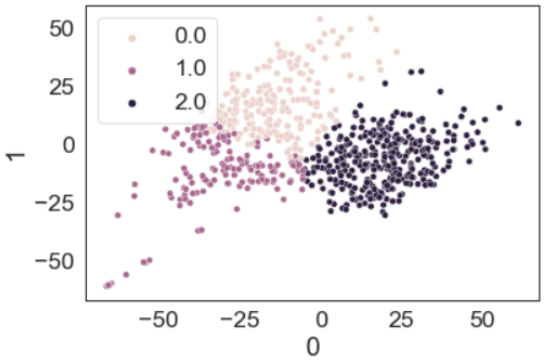
1.



2.



3.



Script aggiuntivo

Per fare un ulteriore confronto sia con le tecniche di estrazione delle feature che con Classificazione e Cluster, si è andato a creare un vettore di label rappresentanti le classi *Maschio* e *Femmina*.
Con questo andiamo ad applicare oltre a PCA anche ad LDA, andando ad osservare una maggiore distinzione delle classi, in quanto LDA non solo riduce la dimensionalità delle feature, ma va anche a considerare la distanza fra le classi e la varianza per evitare la sovrapposizione dei punti.

Inoltre si è andato a svolgere un processo di Cross Validation durante il Training per verificare l'accuratezza e l'errore commesso, utilizzando come classificatore Knn.
Il risultato è un'accuratezza di più del 90%, che è un buon risultato, in conferma del fatto che LDA è un ottima procedura se si hanno dati etichettati per fare classificazione, anche se nella pratica è difficile avere già in partenza dei vettori di label.