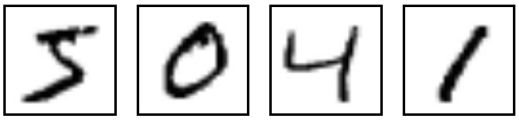
ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING
BIG DATA ANALYTICS
POLITECNICO DI BARI

A.A. 2019/2020
29/11/2019

Si vogliono classificare immagini in scala di grigi della dimensione di 28x28 pixel relative a caratteri numerici da 0 a 9 scritti a mano, come ad esempio quelli mostrati nella figura in basso.



Ciascun campione è presente nel dataset mnist.csv in forma di un vettore di 785 elementi, di cui il primo rappresenta la cifra rappresentata e i successivi 784 rappresentano ciascuno un pixel dell'immagine in un intervallo di interi da 0 (nero) a 255 (bianco).

1) Si riduca la dimensionalità del dataset iniziale utilizzando l'algoritmo PCA, scegliendo il numero di componenti in modo che la varianza che si ottiene sia almeno il 95% di quella originale. *Suggerimento: valutare un numero di componenti maggiore di 100*.

Clustering

2) Si suddivida il dataset ottenuto in 10 cluster utilizzando i seguenti algoritmi:
- K-means (implementato da zero);
- K-medoids (implementato da zero);
- clustering gerarchico (implementato da zero);
- misture di gaussiane (con libreria, oppure opzionalmente da zero);
- DBSCAN (con libreria, oppure opzionalmente da zero);
- HDBSCAN (opzionale, con libreria).

3) Si verifichi, utilizzando il metodo Elbow con uno degli algoritmi di clustering del punto 2, se il numero ideale di cluster sia proprio pari a 10.

4) Per ciascuno dei clustering realizzati al punto 2, si calcoli il Silhouette Coefficient al fine di determinare l'approccio che ha mostrato i risultati migliori.

Classificazione

5) Si suddivida il dataset ottenuto al punto 1 utilizzando uno splitting hold-out 80/20. Dunque, utilizzando un'apposita libreria, si addestri un classificatore SVM con costante C=5 per distinguere la cifra "5" da tutte le altre.

6) Si utilizzi il test set generato al punto 5 per misurare l'accuratezza del modello realizzato.

ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING
BIG DATA ANALYTICS
POLITECNICO DI BARI

A.A. 2019/2020
29/11/2019

The task is the classification of 28x28 grayscale images displaying handwritten digits from 0 to 9, like the ones shown in the below figure.

Each sample is contained in the dataset mnist.csv as an array of 785 elements, where the first element represents the specific digit and the remaining 784 represent every pixel from the image whose values range between 0 (black) and 255 (white).

1) Reduce the dimensionality of the starting dataset by means of PCA algorithm, choosing the number of components such that the obtained variance is at least 95% of the original one.
Tip: consider a minimum component number of 100.

Clustering

2) Split the obtained dataset into 10 clusters by means of one the following algorithms:
- K-means (implemented from scratch);
- K-medoids (implemented from scratch);
- Hierarchical clustering (implemented from scratch);
- Gaussian mixture (with the libraries, or optional from scratch);
- DBSCAN (with the libraries, or optional from scratch);
- HDBSCAN (optional, with the libraries).

3) Check whether 10 is the ideal number of clusters by using Elbow method with one of the clustering algorithms at point 2.

4) For each of the clusterings obtained at point 2, compute the Silhouette Coefficient in order to find out the approach with best results.

Classification

5) Split the dataset obtained at point 1 by using an hold-out splitting 80/20. The, using a specific library, train a SVM classifier with constant C=5 in order to distinguish digit “5” from the other ones.

6) Use the test set generated at point 5 to measure accuracy of the trained model.