Decision tree nella classificazione di caratteri manoscritti

R.Murgia

Contents

1 Introduzione

In questo esercizio si è effettuata la classificazione di immagini di caratteri manoscritti sfruttando l'implementazione del Decision Tree fornita dalla libreria Python Scikit-learn. Prendendo in esame il data set EMNIST, si è analizzato l'errore di predizione commesso sul test set e sul training set al crescere delle dimensioni con cui quest'ultimo è stato considerato in fase di addestramento. Il data set EMNIST comprende un insieme di immagini di dimensione 28x28 pixel raffiguranti caratteri e cifre scritte a mano. Esistono sei differenti suddivisioni di questo data set, quella presa in esame é l'EMNIST letters che comprende 128000 immagini di caratteri maiuscoli e minuscoli raggruppati in 26 classi bilanciate.

2 Implementazione

Il codice da me realizzato si articola in due file, nel primo vengono implementate alcune funzioni per la gestione del data set mentre nel secondo vengono effettuati i test.
Le funzioni implementate sono le seguenti :

- manageDataSet: Il cui compito è quello di rendere il data set dimensionalmente compatibile con con gli input attesi dalle funzioni di addestramento e predizione proprie del Decision Tree. Ricevendo tra i parametri il numero di samples presenti nel data set, una matrice tridimensionale contenente le immagini e le relative labels, la funzione si occupa di effettuare il reshape della matrice e di restituire quindi una matrice bidimensionale che per righe avrà il numero degli samples e per colonne il numero di pixel che costituiscono l' immagine . Tale funzione permette quindi di avere una rappresentazione delle immagini sotto forma di Vettore monodimensionale anziché come matrice.

- balanceDataSet: Il cui compito é quello di garantire che, per ciascuna dimensione fissata, il data set venga opportunamente bilanciato e che quindi sia presente uno stesso numero di samples per ciascuna classe. Riceve tra i parametri il numero di samples, la matrice contenente le immagini e le relative leabel. Queste ultime vengono aggiunte come ultima colonna della matrice la quale é poi sottoposta a un processo di randomizzazione delle righe.

- saveDataSet: All'interno di questa funzione avviene l'estrazione del data set garantendo che questo venga accuratamente separato in Training set e Test set. Successivamente questi sono opportunamente trattati

dalla funzione mageDataset per poi essere salvati in alcuni file esterni di estensione .npy. Questo ha lo scopo di permettere che la procedura di reshape delle immagini possa essere applicata un'unica volta in maniera da minimizzare i tempi necessari per futuri test.

- lettersTest: All'interno di questa funzione vengono svolti i test che mirano ad evidenziare l'accuratezza raggiunta del Decision tree nella predizione del data set e del training set.

  Al fine di effettuare le operazioni di test é risultato utile l'utilizzo di alcune liste ausiliarie:

  - dimension: Questa lista contiene il numero di samples appartenenti al Training set utilizzatiti per l'apprendimento.
    Si é considerato dimensioni del Training set di 2^k, per k crescente da 10 fino al massimo valore compatibile con il numero di samples messo a disposizione dal data set EMNIST.

  - testAccuracy: Questa lista, viene utilizzata per tenere traccia dell'accuratezza raggiunta nella predizione sul Test set al crescere del numero di samples utilizzati in fase di addestramento.

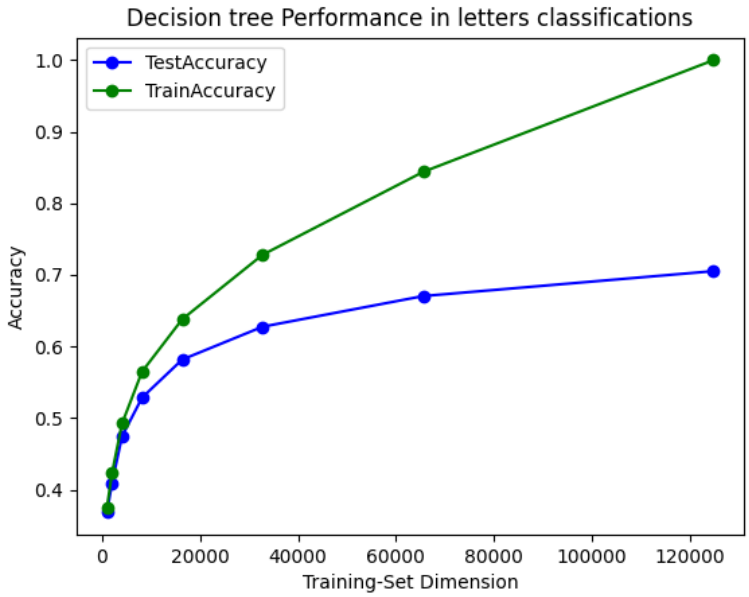
  - trainAccuracy: Questa lista, viene utilizzata per tenere traccia dell'accuratezza raggiunta nella predizione sul Training set al crescere del numero di samples utilizzati in fase di addestramento.

3 Test Effettuati

Per ciascuna cardinalitá del training set vengono effettuate operazioni di addestramento del modello, il quale successivamente, é utilizzato per svolgere delle predizioni sul Test set e sul Training set. Il risultato delle predizioni é quindi usato per il calcolo delle accuratezze raggiunte, le quali vengono poi salvate all'interno delle liste precedentemente descritte con l'intento di realizzare un grafico che dettagli le prestazioni ottenute al crescere del numero dei samples utilizzati.

4 Risultati

4.1 Grafico



4.2 Riflessioni

Il grafico sopra riportato rappresenta l'accuratezza del Decision tree ottenuta nella classificazione di caratteri manoscritti. Viene allegato con lo scopo di evidenziare come questo modello, al crescere della cardinalità dell'insieme di dati utilizzato in fase di addestramento, possa ottenere una maggiore accuratezza nella predizione. Come é possibile osservare i test effettuati hanno portato alla realizzazione di curve di apprendimento caratterizzate da un andamento logaritmico. Inoltre, come era ragionevole pensare, nei test che hanno preso in esame il Training set é stata raggiunta un'accuratezza maggiore rispetto a quella ottenuta negli esperimenti effettuati sul Test set. Si riscontrato ciò poiché si é effettuato delle predizione sullo stesso insieme di dati che ha permesso la realizzazione del modello di classificazione.

5 Riferimenti

- `https://pypi.org/project/emnist/`
- `https://scikit-learn.org/stable/`