Relazione Progetto di Machine Learning
Classificazione di lettere dell'alfabeto ASL

Habasescu Alin Marian

December 11, 2024

Contents

1 Descrizione del Progetto

L'obiettivo di questo progetto è progettare un sistema di classificazione in grado di identificare lettere statiche dell'alfabeto americano dei segni (ASL) a partire da immagini di dimensione 28×28 pixel. Le lettere Y e Z sono escluse dall'analisi poichè richiedono movimento per essere rappresentate.

Sono stati considerati e confrontati diversi modelli di Machine Learning affontati durante il corso, al fine di identificare il modello che ottiene i migliori risultati in ternini di accuratezza e capacità di generalizzazione.

- **Naive Bayes**: Un classificatore probabilistico sul teorema di Bayes e sull'assunzione di indipendenza tra le features.
- **MLPClassifier**: Una rete neurale a più strati che utilizza funzioni di attivazione non lineari per apprendere pattern complessi dalle immagini di input.
- **Support Vector Machine (SVM)**: Un algoritmo di classificazione che trova un iperpiano ottimale che separa le classi.
- **Decision Tree**: Un modello basato su una struttura ad albero, che suddivide i dati in base a regole decisionali sequenziali.

Metodologia Il progetto segue i seguenti passi principali:

1. **Caricamento e visualizzazione dei dati**.
2. **Preprocessing dei dati**.
3. **Addestramento dei modelli** e **valutazione dei modelli**.
4. **Analisi dei Risultati ottenuti**.

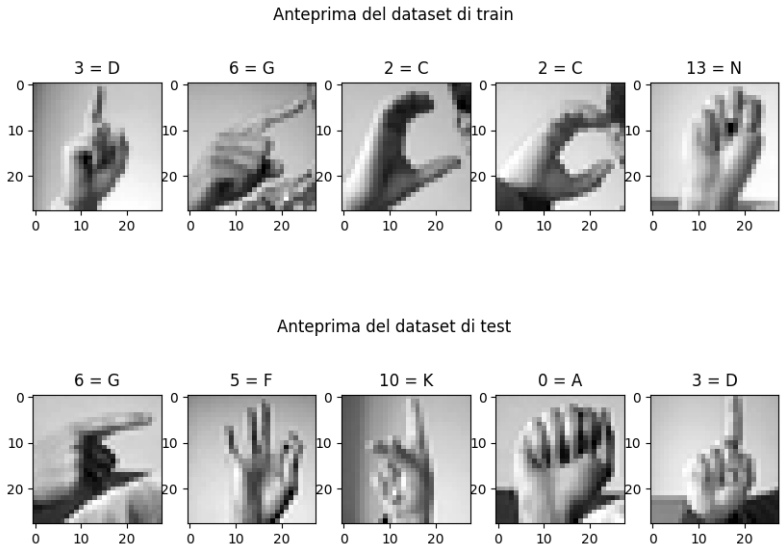
2 I Dataset

Nel progetto si utilizzano due dataset distinti per l'addestramento e la valutazione dei modelli.

Questi dataset vengono caricati dai file CSV forniti tramite la funzione `load_datasets`, che suddivide i dati in *features* (valori dei pixel delle immagini) e *target* (le lettere corrispondenti).

La funzione `visualize_dataset` consente invece di analizzare i dati caricati, mostrando un'anteprima grafica delle prime osservazioni sotto forma di immagini 28×28 pixel in scala di grigi e associandole alle lettere dell'alfabeto ASL (025, mappati su AZ).

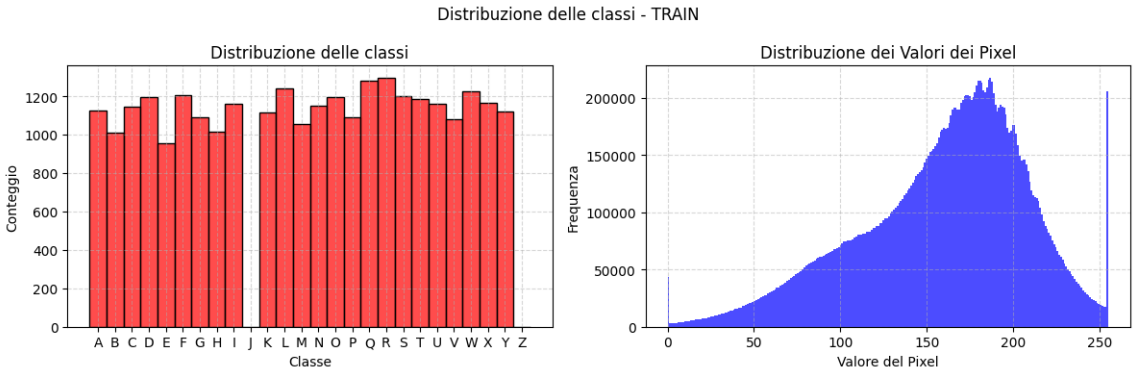
Il dataset di training è composto da 27455 campioni, mentre il dataset di test contiene 7172 campioni. Ogni campione rappresenta un'immagine di una lettera statica dell'alfabeto ASL, accompagnata dai valori di grigio relativi ai 784 pixel che costituiscono l'immagine.



3 Analisi Esplorativa dei Dati (EDA)

Dopo il caricamento dei dataset, viene eseguita una serie di operazioni per analizzare le caratteristiche dei dati.

- **Distribuzione delle classi**: Verifica di un bilanciamento uniforme tra le classi.
- **Distribuzione dei valori dei pixel**: Analisi della gamma di valori di pixel (0-255) per determinare la necessità di normalizzazione.



Durante l'analisi esplorativa dei dati, non sono stati osservati evidenti squilibri nella distribuzione delle classi, evinziando però valori prossimi a 255 per la distribuzione dei pixel.

4 Preprocessing dei dati

5 Addestramento dei modelli e metriche di valutazione

6 Risultati ottenuti