Homework 2

Giuliano Mirabella
Matricola 1000062280

Descrizione del Dataset

- Il CIFAR-10 è stato pensato per testare modelli come regressione logistica, SVM, k - NN, alberi decisionali, usato anche per le reti neurali
- Il dataset contiene un totale di 60.000 immagini a colori, ciascuna di 32x32 pixel, con 3 canali RGB, suddivise equamente in 10 classi semantiche
- Inoltre è predefinito in due set: 50.000 immagini per l'addestramento e 10.000 per il test.
- Nel progetto, per motivi di efficienza computazionale, è preferibile scegliere una porzione ridotta (10%) del dataset, mantenendo comunque una rappresentazione bilanciata delle classi tramite sotto-campionamento stratificato.

Il dataset è strutturato in sei blocchi principali:

- 5 batch di training, ciascuno composto da 10.000 immagini;
- 1 batch di test, anch'esso con 10.000 immagini.
- Il batch di test contiene 1.000 immagini per ciascuna classe, selezionate in modo casuale, così da avere un test set bilanciato.
- I batch di training contengono le immagini rimanenti, mescolate casualmente. Alcuni batch possono avere più immagini di una classe rispetto ad un'altra. Tuttavia, nel totale, il training set è bilanciato, con 5.000 immagini per ciascuna classe, cioè 50.000 immagini in totale

Introduzione al problema

- Dataset: CIFAR10 (60.000 immagini, 10 classi, 32x32 RGB)
- Obiettivo: classificazione delle immagini in 10 classi
- Task: confrontare vari modelli di ML classici

Strategia adottata

- Il dataset CIFAR10 è composto da immagini RGB 32x32 suddivise in 10 classi (oggetti e animali)
- • È stato utilizzato il sotto-campionamento al 10% per ridurre i tempi di training e ottimizzazione
- • Le immagini sono state convertite da 3D a 1D (flattening) per l'uso con modelli classici
- • I modelli sono stati selezionati tra quelli tradizionali di Machine Learning:
- - Regressione Logistica, SVM, k-NN e Decision Tree
- • Per ciascun modello è stata applicata la Grid Search per ottimizzare gli iperparametri
- • L'accuratezza è stata valutata con Cross-Validation (CV=3), report di classificazione e matrici di confusione
- • La strategia ha bilanciato precisione dei risultati e sostenibilità computazionale

Risultati

- Logistic Regression Accuracy = 0.27;

Parametri ottimali

C =1, max_iter = 100.

- SVM Accuracy = 0.286;

Parametri ottimali C =1,
gamma = 'scale',
kernel = 'linear'

- k-NN: Accuracy = 0.274;

Parametri ottimali
metric = 'euclidean',
n_neighbors = 5, weights = 'uniform'.

- Decision Tree Accuracy = 0.20;

Parametri ottimali

max_depth = None,
min_samples_leaf = 1,
min_samples_split = 2.

Conclusione

- Il modello con le migliori prestazioni è: SVM
- I modelli classici soffrono il flattening delle immagini
- Possibili miglioramenti: reti neurali (CNN), più dati, augmentazione
- Il sotto-campionamento ha ridotto i tempi mantenendo la rappresentatività