Machine Learning (M-Z)

Prova pratica

28/11/2025

Analisi chimica e classificazione della qualità vinicola

Dataset: Red Wine Quality

URL dataset Kaggle:

https://www.kaggle.com/datasets/uciml/red-wine-quality-cortez-et-al-2009

Obiettivo: Lo studente dovrà preparare i dati, costruire un modello di regressione, esplorare la struttura dei dati con tecniche non supervisionate e infine addestrare e confrontare modelli di classificazione binaria.

Consegna: un notebook su Kaggle con il codice Python, **senza alcun commento**.

Introduzione al problema

Il dataset contiene 11 feature fisico-chimiche (come acidità, zuccheri, pH) e una variabile target (`quality`) che indica la qualità del vino con un punteggio da 0 a 10.

Il compito è duplice:

1. Regressione: predire il grado alcolico (`alcohol`) a partire dalle altre feature chimiche.
2. Classificazione: classificare un vino come "Standard" o "Ottimo".

Fase 1: Preprocessing

1. **Caricamento e preparazione dati**:
    - Caricare il dataset. Verificare l'assenza di valori mancanti e, se presenti, gestirli opportunamente.
    - La feature `alcohol` è la variabile target per la regressione.

    - Codificare la variabile target di classificazione in formato binario seguendo questa logica: i vini con punteggio `quality` tra 3 e 5 saranno etichettati come Standard (0), mentre quelli con punteggio tra 6 e 8 saranno etichettati come Ottimo (1).
    - Rimuovere le feature `alcohol` e `quality` dalla matrice dei dati risultante.
    - Dividere i dati in un training set (80%) e un test set (20%), assicurandosi che le classi siano equamente rappresentate negli split risultanti.
    - Estrarre un validation set dal training set, assicurandosi che le classi siano equamente rappresentate negli split risultanti.
2. **Standardizzazione**:
    - Standardizzare opportunamente le feature.

Fase 2: Riduzione della dimensionalità con PCA

1. Applicare la PCA al training set standardizzato.
2. Scegliere un numero di componenti `k` che spieghi almeno il 90% della varianza totale.
3. Trasformare gli split del dataset utilizzando le `k` componenti principali selezionate.

La PCA permette anche di visualizzare i dati, proiettando le feature sulle prime due componenti principali.

1. Applicare nuovamente la PCA al training set standardizzato, mostrando le prime due componenti principali.
2. Mostrare il dataset proiettato, colorando ciascun punto in base alla classe di appartenenza.

Questa seconda applicazione della PCA ha uno scopo puramente esplorativo e non finalizzato all'addestramento di modelli.

Fase 3: Regressione

Utilizzare i dati trasformati dalla PCA (con il 90% di varianza) per predire il grado alcolico del vino.

1. Effettuare la model selection tra un insieme di modelli e iperparametri a scelta, sui dati ridotti tramite la PCA. Utilizzare MSE come metrica prestazionale.
2. Riportare MSE e MAE del modello selezionato sul test set trasformato. A fine puramente speculativo, riportare i risultati sul test set anche degli altri modelli provati.

Fase 4: Classificazione

Utilizzare i dati trasformati (output della PCA con il 90% di varianza spiegata) per addestrare e valutare diversi modelli di classificazione sulla qualità del vino. In un contesto commerciale, l'obiettivo è identificare correttamente i vini di alta qualità per il posizionamento sul mercato, mantenendo un equilibrio tra precisione e recupero. Pertanto, la metrica di

riferimento per la model selection e la valutazione delle prestazioni è l'F1-score sulla classe positiva (1=Ottimo).

1. **Modello 1: Rete neurale**:
    - Progettare e implementare una rete neurale per il problema di classificazione binario.
    - La rete deve prendere in input i dati ridotti con la PCA.
    - Effettuare la model selection al variare degli iperparametri del modello e dell'ottimizzatore.
2. **Modello 2: Classificatore a scelta (SVM o Ensemble)**:
    - Scegliere uno tra **Support Vector Machine (SVM)** e un metodo di **Ensemble**.
    - Addestrare il modello scelto utilizzando gli stessi dati di training della rete neurale.
    - Effettuare la model selection al variare degli iperparametri del modello.
3. **Confronto e valutazione**:
    - Riportare i risultati dei modelli testati in una tabella di confronto, includendo **accuracy**, **precision**, **recall** e **F1-score** (calcolati sul validation set).
    - Riportare le prestazioni sul test set del modello selezionato in base alla tabella precedente. A fine puramente speculativo, riportare una tabella analoga alla precedente, mostrando i risultati sul test set.