

Progetto di Data Mining 1.3

Regressione

Introduzione:

Questo progetto mira a condurre un'analisi approfondita del dataset “meta”, un insieme di dati in cui ogni oggetto rappresenta delle caratteristiche di un dataset di letteratura su cui è applicato un algoritmo di learning. L'obiettivo primario è applicare tecniche di regressione.

Note sul dataset: il dataset meta si può scaricare dal repository openml (<https://openml.org/>), mentre le informazioni relative si trovano al link: <https://openml.org/search?type=data&id=566>

Fasi del Progetto:

1. Analisi del Dataset

Caricare il dataset tramite la libreria *scikit-learn*. Il database può essere scaricato mediante la funzione `fetch_openml`:

```
fetch_openml(name = 'meta')
```

N.B. la funzione restituisce un oggetto contenente i dati e le informazioni sul dataset. In particolare l'oggetto avrà i seguenti attributi:

- `data`: array contenente i dati;
- `target`: array contenente le label del dataset. Nel caso specifico, è un valore numerico continuo indicante l'errore normalizzato;
- `feature_names`: i nomi di ciascuna feature.

Inoltre, la descrizione completa del dataset è memorizzata nell'attributo `DESCR` dell'oggetto in questione.

Organizzare i dati scaricati in un DataFrame di **pandas**, e visualizzare le prime righe per acquisire una panoramica delle variabili disponibili.

2. Preprocessing

Eliminare le feature nominali dal set degli attributi. Esplorare il dataset per individuare valori mancanti e trattarli nei seguenti modi:

- Eliminare i dati aventi valori mancanti. (**N.B.** il dataset ottenuto è indicato da qui in poi con **D1**)

- Una volta creato il DataFrame utilizzare la funzione di pandas `interpolate`, che automaticamente effettua una interpolazione dei dati mancanti (**N.B.** il dataset ottenuto è indicato da qui in poi con **D2**). Utilizzare questa configurazione (sia `df` il DataFrame su cui applicare tale metodo):

```
df = df.interpolate(method='cubicspline', limit_direction='both', axis=0)
```

Per D1 e D2, eseguire successivamente la normalizzazione o standardizzazione delle variabili per garantire risultati più accurati durante la fase di predizione.

3. Regressione

Dividere entrambi i dataset in training e test set in maniera opportuna. Utilizzare per entrambi i seguenti modelli di regressione, lasciando i parametri di default:

- Linear Regression;
- Logistic Regression;
- Support Vector;
- Decision Trees;
- Random Forest;
- Gradient Boosting.

Addestrare e testare i modelli precedenti sia su D1 che su D2. Riportare i risultati in termini di MAE, MAPE e SMAPE (In questo ultimo caso implementare una funzione apposita, che prenda due array, uno per i valori reali e uno per quelli predetti, che restituisca lo SMAPE secondo la formula riportata nelle slide del corso). Plottare inoltre grafici opportuni in cui riportare i risultati, sia per D1 che per D2, e analizzare e discutere tali risultati.

Consegna del Progetto:

Gli studenti dovranno produrre il seguente materiale:

1. Il codice sorgente in linguaggio Python.
2. Una breve relazione che spiega le scelte di preprocessing, le metodologie e l'interpretazione dei risultati ottenuti.
3. Una presentazione (10-15 slide max), da presentare ai docenti in una sessione dedicata, che riassume l'intero progetto.

In dettaglio, il materiale dovrà rispettare le seguenti caratteristiche.

1. Codice Sorgente

Il codice sorgente dovrà possibilmente essere ben commentato o essere incluso in un notebook Jupyter. Fare in modo che lo script visualizzi le informazioni relative a ciascuna fase descritta in precedenza, comprendendo i grafici generati.

2. Report Finale

Il report dovrà includere i seguenti paragrafi:

1. *Introduzione*. Comprende descrizione del problema e obiettivi del task assegnato.
2. *Dataset*. Descrizione del dataset e principali caratteristiche.
3. *Preprocessing*. Descrizione dei processi di elaborazione e trasformazione del dataset effettuati, riportando risultati ed eventuali grafici associati.
4. *Regressione*. Descrivere metodologie e modelli utilizzati, motivando eventualmente le scelte effettuate, e riportare i risultati ottenuti (compresi di eventuali grafici).
5. *Discussione e conclusioni*. Analisi e discussione dei risultati, includendo un paragrafo riassuntivo che descriva le considerazioni finali sui task e sui risultati ottenuti.

3. Presentazione

L'obiettivo è preparare una serie di slide (max 10-15) che riassumano e mostrino tutti i punti chiave del report. Includere una slide introduttiva e una conclusiva.