Traccia progetto

Corso di Statistica Numerica
Laurea Triennale in Informatica per il Management

a.a. 2022-23

Contents

1 Selezionare il Dataset

Su `https://www.kaggle.com` o `http://archive.ics.uci.edu/ml` scegliere il dataset di riferimento e scaricarlo.

Ricordarsi di utilizzare sempre, in maniera approfondita, la descrizione del dataset, che ci darà le informazioni necessarie per comprendere il significato delle variabili presenti, che spesso sono codificate da nomi non troppo intuitivi.

2 Caricare il Dataset

Il primo passaggio in un qualunque progetto di Machine Learning è quello di caricare il Dataset. Ricordando di impostare come working directory quella in cui è presente il dataset, in `.csv` che si vuole importare.

3 Pre-Processing

In questa seconda fase, bisogna:

1. Ripulire il dataset da eventuali NaN.
2. Controllare che le variabili di tipo numerico non presentino dei valori fuori soglia (numeri troppo bassi da essere realistici, o troppo alti).
3. Controllare, in generale, che gli elementi del dataset siano corretti ed eliminare eventuali dati corrotti.

4 Splitting

Nella fase di Splitting, il dataset deve essere diviso in training set e test set. Le dimensioni dei due sottoinsiemi così ottenute sono arbitrarie. E' consigliato far sì che il training set sia più grande del test set. Fare alcune prove fino a trovare una buona dimensione.

5 Exploratory Data Analysis (EDA)

In questa fase, ci si può sbizzarrire. L'idea è quella di sfruttare gli strumenti grafici messi a disposizione da Python per indagare alcune proprietà statistiche del dataset. Il numero e la tipologia di grafici dipende dal dataset a disposizione, l'importante è concludere questa fase avendo coscienza di come interagiscono tra loro (a livello statistico) le variabili di input del dataset. Ci sono indagini *univariate*, che coinvolgono una solo caratteristica , oppure *bivariate* che ne coinvolgono due contemporaneamente, oppure *multivariate* che ne coinvolgono più di due contemporaneamente. La stampa della matrice di correlazione è uno strumento che può essere particolarmente utile nell'indagine della maggior parte dei dataset. Ricordo che per stamparla, è necessario utilizzare solo colonne di tipo numerico.

6 Addestriamo il Modello

Siamo quindi pronti ad addestrare il nostro modello. Confrontare i due algoritmi:

1. Regressione Logistica
2. Support Vector Machine (SVM)

7 Hyperparameter Tuning

Abbiamo visto come le performance del modello dipendono drasticamente dalla scelta degli iperparametri (ovvero tutti quei parametri che vanno passati in input alla funzione del modello,

come per esempio, nel caso di `svm`, il `kernel`, il `cost` e il `degree` / `gamma`). Per esempio, se usiamo la funzione `svm` con kernel polinomiale, cerchiamo il parametro `degree` ottimale per la nostra situazione. Fissato quello, bisognerà poi ripetere per il parametro di costo e cercare una combinazione ottima di questi parametri. E' buona norma, seguendo quello fatto a lezione, visualizzare su uno stesso grafico anche la curva dell'errore valutata sul training set, per rendersi conto di come differiscono le due curve. Utilizzare per questa analisi il parametro `Accuracy`.

N.B. Nel caso di SVM considerate i tre tipi di Kernel e identificate il parametro x ognuno di essi.

8 Valutazione della Performance

Una volta identificata la combinazione ottimale di iperparametri per un determinato algoritmo, si può passare alla fase di valutazione, in cui il modello viene valutato (possibilmente utilizzando un numero variabile di metriche) sul test set, e se ne delineano punti di forza e di debolezza, utilizzando anche la matrice di confusione.

9 Studio statistico sui risultati della valutazione

L'esecuzione del modello una sola volta non è sufficiente per dare una valutazione corretta del modello, data la aleatorietà dei dati utilizzati. Per questo si suggerisce di ripetere le fasi di addestramento e testing un numero k di volte con $k \geq 10$. Di ogni metrica di errore abbiamo quindi un SRS(k).

- Usare strumenti di statistica descrittiva (calcolo centro dei dati, diffusione...) e grafici (istogramma, boxplot) per descrivere statisticamente il campione.
- Usare strumenti di statistica inferenziale per fare inferenza riguardo alla distribuzione cui appartiene il campione. In particolare, stimare la media e calcolare l' intervallo di confidenza con livello di confidenza $\alpha = 0.05$ (quindi con probabilità del 95%).

10 Feature selection (opzionale)

Un ulteriore studio sul modello viene effettuato realizzando la cosiddetta operazione di *feature selection*. Consiste nell'analizzare l' influenza che hanno alcune delle features o caratteristiche dei dati sulla performance del modello. In modo molto euristico, si possono selezionare alcune delle caratteristiche ed eseguire il modello SOLO sulle features selzionate. Questa operazione, ripetuta piu' volte scegliendo caratteristiche differenti, permette appunto di individuare le features piu importanti ai fini della qualità del modello.

Ripetere tutti i passaggi di valutazione, con i nuovi input, e giustificare il perché avete scelto proprio quegli input rispetto agli altri.

11 Istruzioni per codice e la relazione

Il codice DEVE essere adeguatamente commentato. Il codice deve essere accompagnato da una relazione in cui si devono discutere i risultati ottenuti dal codice sia per quanto riguarda la parte descrittiva (EDA) che per quanto riguarda la parte predittiva (classificazione), anche

discutendo le scelte fatte sia come grafici nella parte EDA che come parametri nella parte di classificazione.