Traccia progetto

Corso di Statistica Numerica
Laurea Triennale in Informatica per il Management

a.a. 2024-25

Contents

1 Selezionare e Caricare il Dataset

Su `https://www.kaggle.com` scegliere il dataset di riferimento e scaricarlo.

Nella scelta del data set tenere presente che si dovrà fare una classificazione, quindi si deve individuare una caratteristica target adatta alla classificazione, e una regressione fra due variabili di cui una sarà indipendente e l'altra dipendente. Se un unico data set non consente entrambi i task, si possono utilizzare due dataset differenti.

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.

Una volta scaricato, caricare il Dataset sul proprio codice, ricordando di impostare come working directory quella in cui è presente il dataset, possibilmente in `.csv`, che si vuole

importare (facendo attenzione ad impostare la working directory):

2 Pre-Processing

In questa seconda fase, bisogna:

1. Ripulire il dataset da eventuali NaN (comando `na.omit(df)`).
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.

3 Exploratory Data Analysis (EDA)

In questa fase bisogna sfruttare gli strumenti grafici 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.
Utilizzare, fra i grafici, diagrammi in frequenza e boxplot.

Tenere in considerazione che la matrice di correlazione è uno strumento che può essere particolarmente utile nell'indagine della maggior parte dei dataset.

4 Splitting

Nella fase di Splitting, il dataset deve essere preparato per l'addestramento di modello di Classificazione. Per questo, si richiedere di dividerlo in training set, validation set e test set, seguendo le indicazioni presenti nelle slides del corso. Le dimensioni dei tre sottoinsiemi così ottenute sono arbitrarie. E' consigliato far sì che validation set e test set siano all'incirca grandi uguali, mentre il training set sia più grande degli altri due. Fare alcune prove fino a trovare una buona dimensione.

Ricordare che, per rendere valida l'analisi, da questo momento in poi dovremo far finta di non aver a disposizione il test set (che rappresenta il dataset contenente i dati futuri, che devono ancora essere collezionati), e quindi useremo solo training set e validation set.

5 Addestramento del Modello

In questa fase si richiede di utilizzare i modelli di Classificazione visti a lezione, cioè la (`Regressione Logistica` e `SVM`, con i vari `kernel`), addestrati sui dati presenti nel training set, con l'obiettivo di predire la **variabile di target**, scelta in fase di selezione del dataset.

6 Valutazione della Performance

Una volta definito un modello, bisogna valutarne la performance. Per farlo, possiamo utilizzare i modelli addestrati per predire il test set, e valutare la qualità della predizione, delinandone punti di forza e di debolezza.

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 `svm`, come il `kernel`, il `cost` e il `degree` / `gamma`.

Dobbiamo quindi identificare la combinazione ottimale per questi parametri, utilizzando il validation set (e sperando che questa combinazione sia ottima anche sul test set).

8 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%).

9 Regressione

Scegliere un nuovo dataset utilizzando su Kaggle le seguenti parole chiave: **linear regression simple** ed eseguire un aregressione lineare semplice fra le due varaibili del dataset (i dataset che avete trovato dovrebbero contenere solo due colonne, di cui una è la varaibile indipendente e l'altra quella dipendente).

Tale regressione deve includere almeno i seguenti punti:

- Stima dei coefficienti
- Grafico dei punti e della retta
- calcolo del coefficiente r^2
- calcolo del valore di MSE
- Analisi di normalità dei residui

10 Istruzioni per codice e la relazione

Il progetto DEVE essere svolto INDIVIDUALMENTE. Il codice DEVE essere adeguatamente commentato. Per l'esame si devono preparare alcune slides (una decina non di più) e i devono discutere i risultati ottenuti dal codice sia per quanto riguarda la parte descrittiva (EDA) che per quanto riguarda la parte predittiva (classificazione e regressione), anche discutendo le scelte fatte sia come grafici nella parte EDA che come parametri nella parte di classificazione.

Il codice e le slides devono essere compressi in formato .ZIP e denominati `cognome_nome.zip` e devono essere consegnati su Virtuale entro 24h prima dell'appello. Fare attenzione ad allegare TUTTI i file utili per la riproducibilità degli esperimenti nel file compresso (esempio: dati, eventuali file di supporto, ecc.). Il codice e la relazione saranno discussi durante la prova orale. La mancata consegna del progetto su Virtuale preclude la partecipazione all'esame orale.