BIG DATA AND TEXT ANALYSIS
26/07/2024

Nome:	Cognome:
Matricola:	

Parte 1	
Parte 2	
Totale	

Regole:

1. E' vietato comunicare con altri durante la prova.
2. Al termine della prova, lo studente rinomina il notebook con il proprio nome e cognome, lo scarica nel formato **.ipynb** e lo manda via email alla docente **federica.rollo@unimore.it** con oggetto: **BDTA: 26-07-2024**.
3. I risultati saranno pubblicati entro il giorno 07/08/2024.

Note:
Durata della prova: 2 ore. Il file csv che si trova al link https://bit.ly/4bUbfSx.

Parte 0: Il Dataset

Il dataset (https://www.kaggle.com/datasets/saurabhbadole/housing-price-data) contiene dati relativi ad alcune case in vendita, tra cui:

- `price`: prezzo di vendita della proprietà
- `area`: area totale in square feet
- `bedrooms`: numero di camere da letto
- `bathrooms`: numero di bagni
- `stories`: numero di piani
- `mainroad`: indica se la proprietà si trova su una strada principale
- `guestroom`: indica se la proprietà dispone di una camera per gli ospiti
- `basement`: indica se la proprietà ha un seminterrato
- `hotwaterheating`: indica se la proprietà dispone di riscaldamento acqua calda
- `airconditioning`: indica se la proprietà dispone di aria condizionata
- `parking`: numero di posti auto disponibili
- `prefarea`: indica se la proprietà si trova in un'area preferita
- `furnishingstatus`: lo stato di arredo della proprietà.

La variabile da predire è `price`.

Parte 1: Analisi (10 punti)

1. Quante sono le istanze contenute nel dataset? _____ Il dataset è completo (cioè per ogni istanza tutti i valori di ogni attributo sono sempre correttamente specificati – non esistono "missing values")? _____ (punti 1).

2. Verificare se è vero che le case situate su una strada principale sono quelle più costose (punti 1). Considerando soltanto le case su strada principale, realizzare una pivot table attraverso la quale mostrare il prezzo medio delle case con / senza guestroom (sulle righe) con / senza seminterrato (colonne). Quale caratteristica tra guestroom e seminterrato comporta un aumento maggiore del prezzo della casa? Motivare la scelta (punti 2).

3. Rappresentare in un istogramma la distribuzione dei valori della feature `area`. Poi, raggruppare i valori secondo questa suddivisione: gruppo1 (1649-3660]; gruppo2 (3660-6150]; gruppo3 (6150-9150]; gruppo4 (9150-13150]; gruppo5 (13150-16200], e visualizzare in un nuovo istogramma la distribuzione nei gruppi. Infine, indicare per ogni gruppo il numero di istanze per ogni valore di `bedrooms`. Si può dire che le case più grandi hanno più camere da letto? (punti 3)

4. Considerare solo le case con almeno 2 bagni e almeno 2 camere da letto, rappresentare in uno scatterplot i valori di `price` (ascisse) e `area` (ordinate). Colorare i punti nel grafico in base alla presenza dell'aria condizionata. Usare i nomi dei due attributi come etichette di ascisse e ordinate. (punti 3)

Parte 2: Trasformazione e Predizione (20 punti)

1. Si vuole predire il prezzo di vendita delle case. Ricaricare il dataset originale, eliminare eventuali attributi inutili (giustificare la scelta) ed eventuali istanze con valori nulli. Convertire i valori delle colonne `mainroad`, `guestroom`, `basement`, `hotwaterheating`, `airconditioning`, `prefarea` in modo che "yes" sia sostituito con 1 e "no" con 0. Trasformare anche i valori testuali della colonna `furnishingstatus` in valori numerici a piacere. Dividere il dataset in modo che 3/4 degli elementi siano contenuti in un nuovo dataset "train" e 1/4 nel dataset "test".

Allenare il train con il modello LinearRegression e valutare il Mean Squared Logarithmic Error (MSLE) e R2 sia sul dataset train sia sul dataset test. Confrontare i risultati ottenuti con quelli ottenuti con una predizione basata sul modello SGDRegressor. Effettuare alcune considerazioni sui risultati ottenuti. (punti 4)

2. Confrontare i valori di R2 ottenuti nel punto precedente con il valore di R2 che si ottiene con una 5 Fold cross validation. (punti 1)

3. Trovare i parametri migliori di SGDRegressor, agendo sui parametri `loss` e `penalty`. Scegliere alcuni valori da testare e riportare i valori di MSLE e R2 ottenuti con la migliore configurazione, confrontare questi valori con quelli ottenuti al punto 1. (punti 3)

4. Studiare la correlazione tra le feature del dataset, creare un dataframe che contiene, oltre alla colonna target, le 5 feature più correlate (positivamente) al target. Ripetere la predizione sul nuovo dataset e verificare se il MSLE ottenuto con LinearRegression e SGDRegressor migliora (punti 3).

5. Considerare il dataset usato al punto 1, creare una pipeline in cui al dataset normalizzato si aggiunga una colonna che contiene i valori della colonna `area` discretizzati in 5 intervalli. La pipeline deve applicare il modello SGDRegressor con i parametri migliori trovati al punto 3. Valutare MSLE e R2 della predizione. (punti 3)

6. Aggiungere in coda alla pipeline la funzione SelectKBest. Utilizzare la funzione di gridSearchCV per selezionare il K migliore e anche gli intervalli migliori in cui discretizzare i valori di `area`. Ignorare eventuali warning (punti 3).

7. Considerare il dataset usato al punto 1, creare una nuova pipeline in cui gli attributi `bedrooms`, `bathrooms` e `stories` sono discretizzati in 2 intervalli, l'attributo `area` è trasformato con uno StandardScaler e tutti gli altri attributi sono lasciati invariati. La pipeline deve applicare il modello SGDRegressor con i parametri migliori trovati al punto 3. Valutare i risultati ottenuti e confrontarli con quelli ottenuti al punto 5. (punti 3).