Advanced Machine Learning

Lorenzo Mauri matr 807306

October 29, 2020

Assignment 1

The assignment consists in the prediction of the price (same name as the target variable in the dataset) of a private room/entire apartment using a neural network.

The dataset includes all the information you need to learn more about hosts, geographic availability, and necessary metrics of Airbnb apartments in New York City in 2019.

The provided data comprises the training set that can be used for the training (and eventually for the validation) and the unlabelled test set.

1 Data Processing

1.1 Descrizione del dataset

Il dataset fornito comprende complessivamente 33884 osservazioni di training, 9 attributi e la variabile target :

- **Features** : latitude, longitude,minimum nights, number of reviews, reviews per month, calculated host listings count, availability 365, Private room, Entire home/apt, price

1.2 Encoding coordinate geografiche

Per sfruttare al meglio l'informazione contenuta nelle coordinate geografiche ho deciso di raggruppare le osservazioni per distretto : in una prima fase ho convertito le coordinate (latitudine e longitudine) in punti geometrici e tramite una *spatial left join* (con un dataframe fornito da Geopandas) ho potuto integrare i nomi dei distretti, in modo tale da assegnare a ciascun record il proprio distretto di appartenenza.
Noto che questa procedura non introduce "rumore" nei dati in quanto non dipende da nessuna forma di casualità (sarebbe stato diverso, invece, affidarsi ad un algoritmo di clustering) : in questa fase arricchisco solamente il dataset per ottenere miglioramente sulle performance finali del modello.
Fatto questo, tramite *one-hot-encoding* creo 5 attributi corrispondenti alle 5 modalità esistenti (Manhattan,Brooklyn,Bronx,Staten Island e Queens) riportanti il valore 1 o 0 a seconda che l'osservazione in esame appartenga ad un distretto piuttosto che ad un altro : in questo modo il modello non darà maggior peso ad un'osservazione rispetto ad un'altra per il solo valore di etichetta del distretto (ad esempio, il distretto con valore 4 non peserà il doppio del distretto 2 e il quadruplo del distretto 1).

1.3 Analisi esplorativa : criticità

L'analisi esplorativa effettuata sul dataset ha portato alla luce alcune criticità riguardanti gli attributi :

1. **longitude e latitude**
   Osservo un massimo di 40894 per la latitudine e un minimo di -74124 per la longitudine.
   Google Maps identifica la città di New York con le coordinate (40.730610,-73.935242), pertanto posso ritenere ragionevole che errori di input abbiano spostato il separatore decimale di tre cifre.
   Divido quindi i valori inputati in modo errato per 1000, spostando così il separatore.

2. **price**
   Osserviamo valori pari a 0 . Per qualche ragione il proprietario è disposto ad affittare l'abitazione gratuitamente.
   Decido di eliminare le osservazioni per cui il prezzo è pari a zero.

   Al fine di migliorare le performance del modello sono stati eliminati eventuali outlier, riducendo in tal modo la variabilità.

1.4 Normalizzazione

Al fine di ridurre le differenze di scala applico una normalizzazione ricordandomi che, una volta ottenute le previsioni, dovrò riapplicare la trasformazione per ottenere le previsioni sui dati originali

$$x_{norm} = \frac{x - min}{max - min}$$

Le differenze di scala potrebbero infatti indurre il modello a dare maggiore importanza ad un attributo piuttosto che ad un altro (è la stessa differenza riscontrabile tra le

variabili Età e Reddito, che hanno scale totalmente diverse ma non per questo sono una più rilevante dell'altra).

2 Data Modelling

Al fine di valutare correttamente le performance del modello (e in particolare avere maggiori informazioni circa la generalizzazione del modello) ho deciso di dividere il dataset di training in due parti distinte (90% dei dati per il training set e il restante 10 % per il test set, in modo pseudo-casuale). Ottengo in tal modo 28587 osservazioni per il training e 3177 per il test.

2.1 Struttura della Neural Network

La Neural Network implementata possiede le seguenti caratteristiche :

- **Numero di hidden layers** : 1
  Considerato il volume di dati ritengo sufficiente un solo hidden layer in quanto una configurazione con più di un hidden potrebbe far cadere il modello in overfitting
- **Numero di unità/neuroni per hidden layer** : 32
  Strati con maggiori unità impediscono al modello una buona generalizzazioni sui nuovi dati
- **Numero di unità/neuroni per input**: 12
  Corrispondente al numero di attributi presenti nel dataset normalizzato e codificato
- **Numero di unità/neuroni per ouput**: 1
  Corrispondente alla singola previsione del modello. Il valore di output sarà compreso tra 0 e $+\infty$ (qualora vi fossero valori negativi, la funzione ReLU provvederà ad annullarli)
- **Funzioni di attivazione** : Rectifier Linear Unit (ReLU)
  Si è rivelata la funzione di attivazione più performante per il problema in esame pertanto ho deciso di inserirla in ogni strato della rete
- **Funzione di perdita** : Errore quadratico medio (MSE)
  E' la funzione di perdita più utilizzata nei problemi di regressione.Ho deciso di utilizzarla in quanto penalizza fortemente gli errori grandi e di meno quelli più piccoli aiutando pertanto il modello durante la fase di training.
  Rispetto ad altre funzioni quali la funzione Mean Absolute Error, non ha problemi di derivabilità.

3 Risultati e conclusioni

Di seguito i grafici di performance del modello, nei quali è possibile notare l'andamento del RMSE durante la fase di training e validation. Questi ultimi indicano una buona generalizzazione del modello. Alla fine della fase di training e validation ottengo le seguenti metriche :

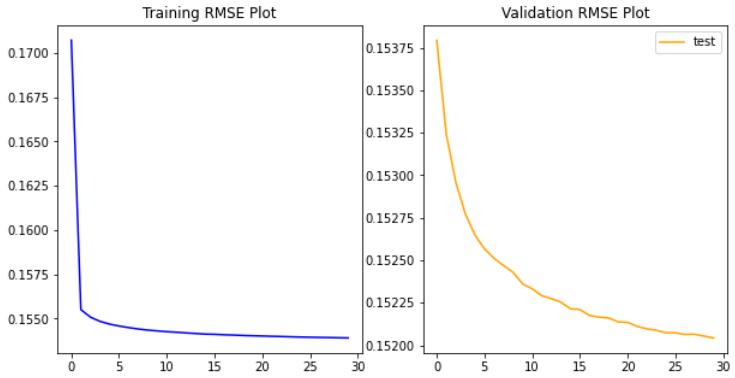
$$trainingRMSE = 0.1539$$

$$validationRMSE = 0.1520$$

Applicando la trasformazione inversa delle previsioni ottenute sui dati di test ottengo :

$$RMSE = 49.11$$

$$MAE = 36.164$$



Se consideriamo le metriche sulle trasformazioni degli output possiamo notare che i prezzi previsti non si discostano di molto da quelli reali, come mostrato in tabella.

predicted	actual	MAE	RMSE
92.950163	80.0	12.950163	167.706709
176.507061	200.0	23.492939	551.918171
143.881747	140.0	3.881747	15.067963
192.188214	150.0	42.188214	1779.845386
154.480245	250.0	95.519755	9124.023670