Assignment 1 Ferrario Gabriele 817518

Durante l'analisi del dataset mi sono accorto della presenza di una colonna senza nome riportante gli indici ed ho provveduto alla sua eliminazione (nel X_test e nel X_train). Proseguendo con l'analisi ho eliminato le istanze che presentavano un valore del campo price maggiore o uguale 9999. Analizzando i campi Private_room e Entire_home/apt ho notato che ci sono istanze con entrambi i valori a zero, questo potrebbe significare che ci sono problemi nei dati poiché per ogni istanza mi aspetto un 1 in almeno uno di questi campi. Ho comunque considerato queste istanze perché ho notato che sono presenti anche nel file X_test e non ho la certezza sul fatto che sia un errore dei dati. Ho verificato che non ci siano valori mancanti (nan) ed ho trasformato le features ridimensionandole in un intervallo [0, 1] tramite il MinMaxScaler[1] (ho utilizzato questo scaler poiché in fase di valutazione del modello tra quelli utilizzati é stato quello che mi ha fornite le performance migliori).

Per l'allenamento e la valutazione del modello ho deciso di suddividere i dati a disposizione per il training in 90% training e 10% validation.

Il modello é una rete neurale composta da 4 layer:

1. layer di input che accetta vettori di dimensione pari a 9 (dove 9 il numero di features utilizzate per la predizione);

2. hidden layer denso con un numero di unitá pari a 13 e come funzione di attivazione usa la relu;

3. hidden layer denso con un numero di unitá pari a 6 e come funzione di attivazione usa la relu;

4. output layer denso con un numero di unitá pari a 1 e come funzione di attivazione usa la funzione lineare (non l'ho specificata e quindi prende questa di default);

Ho usato questa configurazione della rete poiché nelle prove che ho fatto é stata quella che si comportata meglio ottenendo le performance migliori.
Come optimizer ho utilizzato adam mentre come loss function e come metrica MSE. Ho scelto di usare MSE e non MAE poiché quest'ultima metrica é meno sensibile agli outliers.

[1]Questo estimator scala e traduce ogni features individualmente in modo tale che si trovi nell'intervallo dato sul set di addestramento, ad esempio tra zero e uno.

Il modello é stato allenato tramite 100 epoche e con la dimensione dei batch pari a 16. Con il passare delle epoche il valore della loss function si abbassa e raggiunge cifre vicine a 5.65e-04.