ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING
BIG DATA ANALYTICS
POLITECNICO DI BARI

A.A. 2019/2020
13/12/2019

Una libreria online ha realizzato un grande database con alcune informazioni sui libri in vendita. Il database è contenuto nel file `books/train.csv` e la sua struttura è rappresentata di seguito.

Title	Author	Edition	Reviews	Ratings	Synopsis	Genre	BookCategory	Price
…	…	…	4.2	62	…	…	…	9,99

L'obiettivo di questa challenge è quello di utilizzare le informazioni raccolte per addestrare un modello di Machine Learning in grado di predire il prezzo di un libro a partire dalle informazioni disponibili su di esso.

I dati. Il modello dovrà lavorare sull'insieme di feature che si riterranno più adeguate allo scopo. Alcune colonne potranno necessitare di pulizia o estrapolazione dell'informazione. Per alcune feature potrebbe essere necessaria la trasformazione in categorie (ad esempio, con one hot encoding o label encoding). Per altre feature, invece, potrebbe essere necessario l'embedding da testo a "vettore numerico" (in questo caso si ricordi che un testo necessita di essere pulito, ad esempio con rimozione della punteggiatura, e su di esso sono indispensabili operazioni di stemming e tokenization). Qualunque tecnica illustrata a lezione (normalizzazione, PCA, …) potrà essere utilizzata, laddove essa risulti utile.

Il modello. Sarà possibile realizzare qualunque modello di Machine Learning capace di raggiungere gli obiettivi richiesti. Sarà possibile fare qualunque scelta riguardante architettura e iperparametri, purché giustificata.

Le valutazioni. Si effettuino le scelte più adeguate per effettuare la validazione del modello realizzato. Esso, infine, dovrà essere valutato sui dati presenti in `books/test.csv`. Saranno ritenuti adeguati i modelli che raggiungono uno score, calcolato come segue, di almeno 0.65

$$Score = 1 - RMSLE,$$

dove *Root Mean Squared Log Error* (*RMSLE*) $= \sqrt{\frac{1}{2n}\sum_{i=1}^{n}(log_{10}(\hat{y}_i + 1) - log_{10}(y_i + 1))^2}$

ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING
BIG DATA ANALYTICS
POLITECNICO DI BARI

A.A. 2019/2020
13/12/2019

An online book shop has built a big database with data about on sale books. The database is contained in `books/train.csv` and it is structured as follows.

Title	Author	Edition	Reviews	Ratings	Synopsis	Genre	BookCategory	Price
…	…	…	4.2	62	…	…	…	9,99

The goal of this challenge is to exploit the collected data to train a Machine Learning model capable of predicting the price of a book starting from the available information about it.

The dataset. The model will work on the features you think are suitable for the final purpose. Some columns might require cleaning or information-extraction operations. Some features might need to be transformed into categorical ones (e.g., with one hot encoding or label encoding). Other features might require a text-to-numerical-array embedding (in such situation, please remember that text is supposed to be cleaned - e.g., by removing punctuation marks - and it is essential to perform stemming and tokenization operations on it). You can adopt any of the techniques explained during the lectures (normalization, PCA, …), if it is useful.

The model. You can build any Machine Learning model capable of reaching the given goal. You can take any decision about the architecture and the hyperparameters, provided that you explain why you made this choice.

Evaluating. You will choose how to correctly validate the built model. The model will eventually be evaluated on the dataset contained in `books/test.csv`. Only the models which reach a score of at least 0.65 will be positively evaluated (the score is calculated as follows):

$$Score = 1 - RMSLE,$$

where *Root Mean Squared Log Error* (*RMSLE*) $= \sqrt{\frac{1}{2n}\sum_{i=1}^{n}(log_{10}(\hat{y}_i + 1) - log_{10}(y_i + 1))^2}$