Airline Departure
Data Analysis and Regression

Lucchi Manuele & Tricella Davide

August 25, 2022

Instructors: Professor Cesa-Bianchi & Professor Malchiodi

Abstract

The purpose of this paper is to evaluate the usage of a Logistic Regression model on a airlines dataset to predict flight cancellation or diversion, in a scalable and time/space efficient implementation.

1 Definitions

Label

Model

2 Dataset

The initial dataset, [Airline Delay and Cancellation Data] is made of 9 years of airlines flights data, composed by 10 files (one for each year from 2009 to 2018) of around 6 milions records each. The files presents 28 columns, of which we only took the 9 more relevant

FL_DATE The flight date.

OP_CARRIER The carrier code.

ORIGIN The departure place.

DEST The destination place.

CRS_DEP_TIME The.

CRS_ARR_TIME The.

CANCELLED If the flight has been canceled.

DIVERTED If the flight has been diverted.

CRS_ELAPSED_TIME

DISTANCE The distance the flight has to cover.
In the case the prediction is about the cancellation, the DIVERTED column will be ignored, while if the prediction is on if the flight would be diverted or not, the CANCELLED column will be ignored.

The carrier code is a two characters alphanumeric code, the origin and destination places are a three characters alphanumeric code.
Flight date, departure time and arrival time are dates, while the elapsed time and the distance are real numbers.
Cancelled and diverted are either 0 or 1.

One milion of records equally distributed between the files were taken to perform the training.

3 Preprocessing Techniques

Multiple preprocessing techniques were used.
First, the dataset has been balanced in regard of the evaluated property, be it being canceled or diverted, so that there are an equal number of uniformly drawn positives and negatives. MIGLIORARE
Then the data not already represented as real numbers has been converted; places and carriers, that were alphanumeric codes, had a number assigned based on the code, dates were splitted between the year and the rest, with the latter being hashed MIGLIORARE.
The data (now completely composed of real numbers) was then normalized between 0 and 1, to avoid exploding values.
Lastly, the data was splitted between the training set (75%) and the test set (25%).

DIVISO PER MEDIA E VARIANZA???

4 Preprocessing Parallelization

5 Model

5.1 Parameters initialization

Parameters such as Weights and Bias are initialized using a uniform distribution between 0 and 1, with the first one having the same length as the number of columns and the second being a scalar value.
DA VEDERE ITERAZIONI, BATCH SIZE E LEARNING RATE

5.2 Algorithm

5.3

- descrizione inizializzazione parametri - descrizione batching e iterazioni - descrizione forward propagation - descrizione loss - descrizione calcolo gradiente - descrizione update

6 Performances

- come scalano le operazioni di numpy effettuate

7 Experiments

- nostro modello dopo v1 iterazioni con learning rate v1 - nostro modello dopo v1 iterazioni con learning rate v2 - nostro modello dopo v2 iterazioni con learning rate v1 - nostro modello dopo v2 iterazioni con learning rate v2 - confronto modello di sklearn con i valori migliori

8 Results and Conclusions

The atomic weight of magnesium is concluded to be $24\,\mathrm{g\,mol^{-1}}$, as determined by the stoichiometry of its chemical combination with oxygen. This result is in agreement with the accepted value.