Challenge - Software Development

Chloé Serre-Combe

April 1, 2021

1 Introduction

The aim of this prediction project is to predict the number of bicycle passing next to "Albert 1er" counter in Montpellier on Friday, 2021, April 2^{nd}, between 00:01 AM and 09:00 AM.

Link to the git repository : `https://github.com/chloesrcb/bike_challenge`

2 Data treatment

First of all, we use the data from "Albert 1er" counter. These have recorded the number of bike passing since 2020, March 12^{th}. In these data, there are three columns, one for the dates and hours, one for the total record of bikes since 2020, March 12^{th}, and the last for the daily record of bikes since 00:01 at every record hour.

However, 2020 was a special year because of health crisis. So, we have to consider lockdown dates. To this end, we add a column `confinement` containing 0 and 1. Moreover, in 2020 and in 2021, we have to considers the curfews dates and hours. To do this, we add a column `couvre-feu` containing 0 and 1.

Then, in the aim to have a better prediction, we add some columns about weather in Montpellier. Indeed, when there is lot pf wind or rain it's more complicated to ride a bike. And people ride bikes more often when the weather is nice. So, we collect 2020, January and February 2021 data from Historique Météo and we add manually March 2021 weather data by using the website Weather Underground. Thanks to these weather data, in our dataframe with bikes, we add a column for the maximum wind speed, a column for the rain and a column for the global impression of the weather. In the latter, there are 5 values from 0 (bad weather) to 4 (ideal weather).

Since the data recorded at time t corresponds to the total bikes of the day, knowing the previous record on the same day is a valuable piece of information, so we add the previous recorded value and the timestamp of it.

Finally, we add a column for holidays and statutory holidays because it can have an influence on the prediction. But afterward, we remove it because it has a bad influence on the prediction.

3 Prediction

To make the prediction we will use a regression model based on the correlated data we have. With a sufficient amount of records, we are able to construct a model by making links between inputs (weather, previous record, weekday...) and their effects.
We can write a linear regression like this :

$$Y = \beta X + \epsilon$$

with :

- Y: the prediction
- β: the weight associated to X
- X: the covariates (columns in our dataframe)
- ϵ: the residual error of the model

To do this, we use a software library for machine learning named `Tensorflow`. This library allows us to build a regression model based on neural networks.

A neural network is a set of nodes linked to each others based on the human brain functioning. It is made up of layers, one is the input layer, one is the output layer and others are the hidden layers.

In our case, we have in input 15 nodes corresponding to the 15 columns of our dataframe and we only have one output node corresponding to our prediction. Moreover, we choose 2 hidden layers, with 128 nodes, arbitrarily because it seemed to give the best prediction with a globally low computational time.

Then we have to train the model. The training of a model is the way to adjust the influence of each node to its successors in the next layer. It's done by giving the data in input of a corresponding record, establish a prediction and compute the mean squared error between the real value and the prediction then the influence is modified to better match the real value. This process need to be repeated a huge amount of time to reach a convergence in predictions.

4 Conclusion

Thanks to this process, we are now able to predict the number of bikes on a given timestamp knowing the input data such as weather, working day or not, weekday... So, for Friday, 2021, April 2^{nd}, between 00:01 AM and 09:00 AM we predict a number of $\mathbf{334.779 \approx 335}$ bikes.

To go further, to improve this prediction code it would be nice to have better weather data, because those we have are for every day and not every hour of each day and we know that the weather can change in a day. Moreover, by adding some columns one by one and changing them coherently at the call of the prediction function, we can see the influence of every covariates. Indeed, the lockdown and curfew covariates were crucial in the prediction and the weather data had improve it.