ITESM Campus QRO

Linear Regression for Power Consumption Electricity

David Reyes Núñez – A01209004

Abstract

The goal of the project is to implement a linear regression model with python and use a data base in csv format with more than 1,000 instances to predict the amount of electricity transfer between two states of Australia, New South Wales and Victoria.

Introduction

Linear regression is a model that look for a relationship between one or more features or independent variables and a continuous target variable. In the case that we only have one feature, we talk about a Univariate Linear Regression model, if there are multiple features, we call of a Multiple Linear Regression Model.

To represent a Linear Regression Model we use a hypothesis function. This function mark the relationship of the independent variables of the database and the target.

$$Y = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_n x_n$$

Training the model means to find the parameters or weights that makes the fit the best possible with the data. To measure the performance of the model we need to be able to measure the error between the hypothetic data found with the model and the real value of the target and we define the loss function to compute the square – mean error of the model result.

$$J_\theta = \frac{1}{2m}\sum_{i=1}^{m}(h(x^i) - y^i)^2$$

Where $h(x^i)$ is the hypothesis function.

To improve the model is necessary to modify the values of the parameters at any iteration. To accomplish this task we use the Gradient Decent. It is a generic optimization algorithm used in many machine learning algorithms. It consist of iteratively tweaks the parameters of the model in order to minimize the cost function an improve the accuracy. Here it is the mathematical definition of the function to update one parameter.

$$\theta_{1\,update} = \theta_{1\,last} - \frac{\alpha}{2m}\sum_{i=1}^{m}(h(x^i) - y^i){x_1}^i$$

Database

The database used in this project was obtained from OpenML, a website dedicated to sheare free databases for machine learning applications in different topics. This time is used a database of 45,312 instances with five features that are describe as follows:

- **Date**: date between 7 May 1996 to 5 December 1998. Here normalized between 0 and 1
- **Day**: day of the week (1-7)
- **Period**: time of the measurement (1-48) in half hour intervals over 24 hours. Here normalized between 0 and 1
- **NSWprice**: New South Wales electricity price, normalized between 0 and 1
- **NSWdemand**: New South Wales electricity demand, normalized between 0 and 1
- **VICprice**: Victoria electricity price, normalized between 0 and 1
- **VICdemand**: Victoria electricity demand, normalized between 0 and 1
- **transfer**: scheduled electricity transfer between both states, normalized between 0 and 1

As is mentioned in the abstract, the goal is to predict the energy transferred between New South Wales and Victoria state. In this case, the features needed are the NSW price, NSW demand, VIC price and VIC demand.

date	day	period	nswprice	nswdemand	vicprice	vicdemand	transfer	class
0	2	0	0.056443	0.439155	0.003467	0.422915	0.414912	UP
0	2	0.021277	0.051699	0.415055	0.003467	0.422915	0.414912	UP
0	2	0.042553	0.051489	0.385004	0.003467	0.422915	0.414912	UP
0	2	0.06383	0.045485	0.314639	0.003467	0.422915	0.414912	UP
0	2	0.085106	0.042482	0.251116	0.003467	0.422915	0.414912	DOWN
0	2	0.106383	0.041161	0.207528	0.003467	0.422915	0.414912	DOWN
0	2	0.12766	0.041161	0.171824	0.003467	0.422915	0.414912	DOWN
0	2	0.148936	0.041161	0.152782	0.003467	0.422915	0.414912	DOWN
0	2	0.170213	0.041161	0.13493	0.003467	0.422915	0.414912	DOWN

Table 1. Sample of the database.

To make the test of the algorithm faster, it is used just 2000 random features to train the model and other 100 to test it.

Python implementation

The source code can be found in following Githun repository:

https://github.com/DavidReyNu28/Machine_Learning.git

However, there are some punctual notation about the code and how it works.

Libraries

- `matplotlib.pyplot`: This library is very useful to plot data in python.
- `numpy`: This library makes easier the task to work and deal with arrays.
- `pandas`: This library is critical in the program because it facilitates the reading of the database in csv format and save in arrays the features and target instances.

Read the dataset with pandas library

CSV format (Comma Separated Value) is very common to save databases and, to process this data, it is necessary a tool that allow obtain the information of the database in an order form and panda deal with this topic. Pandas is a data manipulator package in Python that use data frames as data type for storing tabular 2D data.

Here is the line used in the source file of the project to read the database stored in local way.

```
data=pd.read_csv("Datasets/Electricity_Norm/electricity-
normalized.csv")
```

Results

After running the code, in the console are shown the different values of the Error Mean at any epoch running (3000 in this project).

```
E_meam = 0.033637
E_meam = 0.033630
E_meam = 0.033624
E_meam = 0.033617
E_meam = 0.033610
E_meam = 0.033604
E_meam = 0.033597
E_meam = 0.033591
E_meam = 0.033584
E_meam = 0.033577
E_meam = 0.033571
E_meam = 0.033564
E_meam = 0.033558
E_meam = 0.033551
E_meam = 0.033545
E_meam = 0.033538
E_meam = 0.033532
E_meam = 0.033525
E_meam = 0.033519
E_meam = 0.033512
E_meam = 0.033506
E_meam = 0.033500
E_meam = 0.033493
E_meam = 0.033487
```

Figure 1. Training Error

When the training process has finished, it is shown the Test Error mean with the 100 test instances. This result is very important because it allows to say if the linear regression model is overfitting, meaning that the model works just for the training data because it memorized the features and not the patron of the data, or if it is capable to predict the target with new data different to the training one.

```
Weights:

[0.3407803152166932, 0.016310137760272764, 0.12600752373100094,
 0.0009140085768203154, 0.12125297347523385]

Test Errormean = 0.022957
```

Finally, it is a little query that allow the user to get a prediction of the energy transfer respect the prices and demand of both Australian states.

```
Enter nswprice:0.041161

Enter nswdemand:0.13493

Enter vicprice:0.003467

Enter vicdemand:0.422915
0.40973672211908213
```

In the next graph printed with the program we see how the training error is decreasing while the number of epochs or iterations are increasing until the parameters stop updated or the training process get the 3,000 epochs.

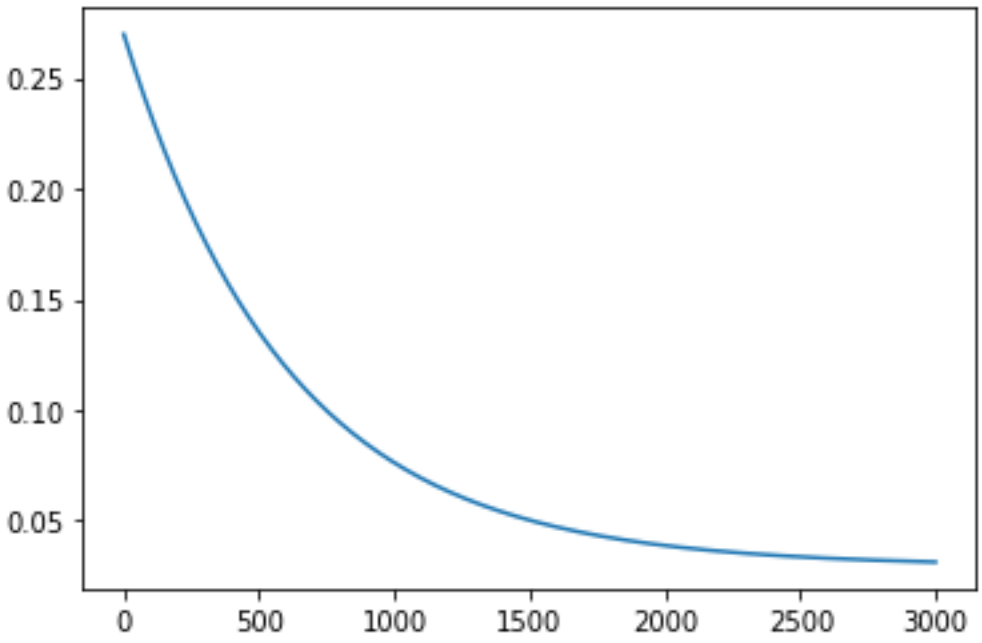

Figure 2. Graph of the training error respect to the number of epochs.

References

Linear Regression using Python. (2018). *Medium*. Retrieved 25 May 2020, from https://towardsdatascience.com/linear-regression-using-python-b136c91bf0a2

Vanschoren, J. (2020). *OpenML*. *OpenML: exploring machine learning better, together.*. Retrieved 25 May 2020, from https://www.openml.org/d/151

Python Pandas read_csv – Load Data from CSV Files. (2018). *Shane Lynn*. Retrieved 25 May 2020, from https://www.shanelynn.ie/python-pandas-read_csv-load-data-from-csv-files/