LINEAR REGRESSION

From scratch

SEOUL BIKE SHARING SYSTEM

Deep Learning

Alejandra Verónica López Chiquito

Universidad Autónoma de Querétaro

Querétaro, Qro. A 19 de septiembre de 2023

Abstract

One of the most used machine learning models is linear regression, however, it cannot be the correct one for all the problems. This model is beneficial to find the relationship between independent variables and one dependent variable.

The paper's data set is categorized or was designed as a regression task. The results of this exercise will define if the regression is the best machine learning model to solve this problem.

Dataset information

The file is published in the University of California Irvine (UCI) machine learning repository, and it is classified as a regression tasks.

This dataset contains a count of public bicycles rented per hour in the Seoul Bike Sharing System, with corresponding weather data, the season, and a column to know if that day was a holiday or not.

Data Set link:
https://archive.ics.uci.edu/dataset/560/seoul+bike+sharing+demand

Variable information

Date: year-month-day

Rented Bike count - Count of bikes rented at each hour.

Hour - Hour of the day

Temperature-Temperature in Celsius

Humidity - %

Windspeed - m/s

Visibility - 10m

Dew point temperature - Celsius

Solar radiation - MJ/m2

Rainfall - mm

Snowfall - cm

Seasons - Winter, Spring, Summer, Autumn

Holiday - Holiday/No holiday

Functional Day - NoFunc(Non Functional Hours), Fun(Functional hours)

(UC Irvine Machine LEarning Reporsitory, 2020)

Definitions

Machine learning is part of the artificial intelligence that allows to the models learn from the data.

Multiple linear regression is a model of the machine learning used to modelling the relationship between more than 2 variables. Formula:

$$y = \beta_0 + \beta_1 X_1 + \ldots + \beta_n X_n + \epsilon$$

Mean Square error (MSE): It is used as statistic metric to evaluate how accurate is a model. Formula:

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{Y_i}\right)^2.$$

Theory

As Seoul is a very populated city from South Korea, they have a very modern bike circuit into the downtown. Currently, the population in the Urban area of Seoul is of 9,977,776 citizens. This is a big indicator that the bikes are very good choice as transportation option.

This bike rental system can be used by all the citizens and foreign people. In 2022, there was more than 2000 bikes to rent in the public stations in the city of Seoul and the number of stations is over 800 across Seoul.

The data included in the 'SeoulBikeData.csv' file is sufficient to create a model of machine learning and make predictions of different areas.

Methodology

As mentioned, the goal of this exercise is to find the relationship between the independent variables (hour, Temperature, Humidity, Wind speed, Visibility, Dew point temperature, Solar Radiation, Rainfall, Snowfall, Seasons, Holiday) and the independent variable (Rented Bike count) and try to make some prediction in how many bikes will be rented in different weather conditions or days.

The steps to find this relationship will be:

- Prepare data.

- Visualize data.
- Split data
- Fit the model.
- Make predictions.
- Evaluate the model.

In the first section of the notebook, no libraries will be used and in the second one the sklearn library will be used. In this way, the short and long process will be showed.

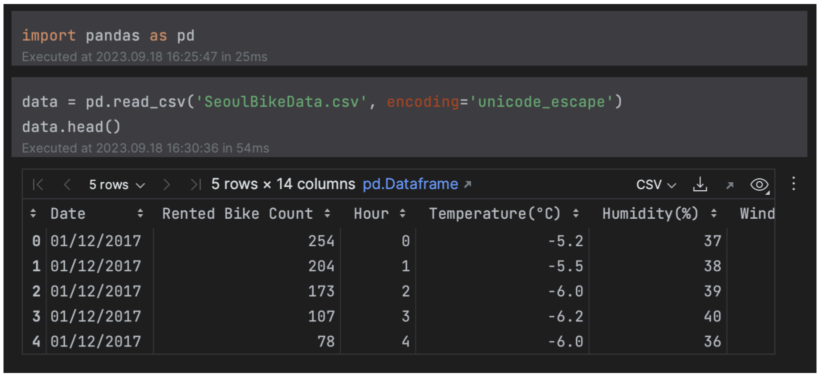
Hypothesis

With all the information in the data set, the goal is to know if there is relationship between different conditions and the rented bikes in Seoul. Also, it is important to verify if the multiple linear regression model is the correct technique to solve this question. This will be validate training a linear regression model from scratch and applying the regression formulas to the data set.

Development

To start the preprocessing process, it is better to remember that the objective of this notebook is to know the relationship between the factors and the number of bikes rented, and then try to predict the number of bikes rented in a determinate condition is important to analyze the columns that is important to drop, use or modify to train the model.

The csv is imported using pandas library.



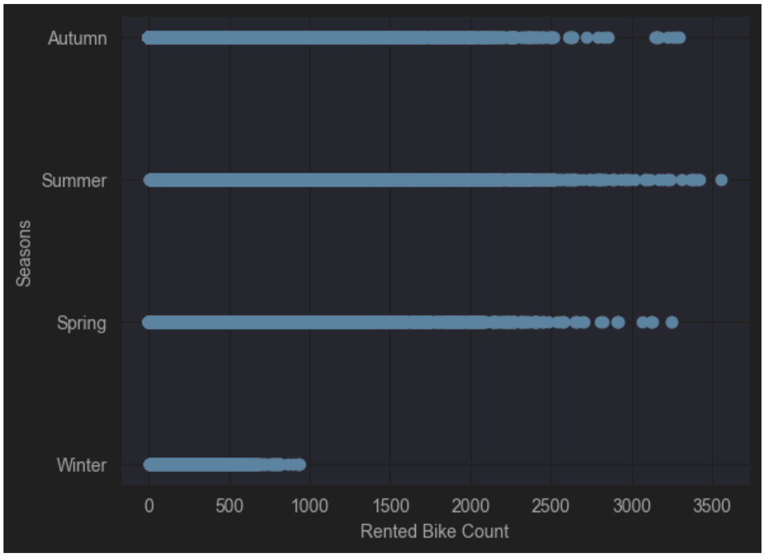
```
import pandas as pd
Executed at 2023.09.18 16:25:47 in 25ms
```

```
data = pd.read_csv('SeoulBikeData.csv', encoding='unicode_escape')
data.head()
Executed at 2023.09.18 16:30:36 in 54ms
```

```
  Date        Rented Bike Count  Hour  Temperature(°C)  Humidity(%)  Wind
0 01/12/2017                254     0             -5.2           37
1 01/12/2017                204     1             -5.5           38
2 01/12/2017                173     2             -6.0           39
3 01/12/2017                107     3             -6.2           40
4 01/12/2017                 78     4             -6.0           36
```

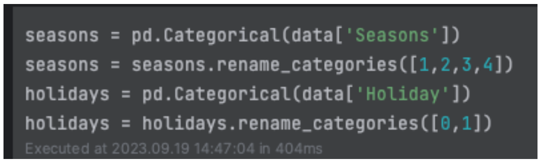
The first column, Date, is not important due to it has a lot of null values and the rest of the columns are factors that provide more information to the model.

Most part of the people have a better understanding of a problem using tools to visualize the data. The introduction of the library matplotlib is one of those tools and the first use to it is see the plot of the dependent variable (DV) with the independent variables (IV).



Following the objective, the Y of this data set is the column ‘Rented bike Count’ and the rest columns are part of the X matrix.

In the last two columns (Seasons and holidays) of the X, the values are discrete, and to train the model all the data need to be continuous. To fix this little issue, the function rename_categories was used to pass from the string values to integer ones.



```
seasons = pd.Categorical(data['Seasons'])
seasons = seasons.rename_categories([1,2,3,4])
holidays = pd.Categorical(data['Holiday'])
holidays = holidays.rename_categories([0,1])
Executed at 2023.09.19 14:47:04 in 404ms
```

Other preprocessing operations is to insert the one column in the first place on the X matrix and the create the transpose matrix of X. Also, the creation of the thetas array with random number between 0 and 1. In this exercise, 12 thetas will be needed.

The hypothesis is represented by the h variable and the m variable was get form the shape of X, it will be used into the cost function.

The training of the model is focused in 2 functions, the optimizer function, and the cost function.

The MSE was calculated before and after the call of the train function.

	MSE
Value before train	4.57964787e+13
Value After train	1.79000044e+24

Once this part of this exercise is done, the second part will start, it is using sklearn library. The first step is to split the data to distribute all the information to train and test the model.

In this part, all the steps are faster due to sklearn can avoid a lot of lines of code and make it in one line, for example the creation of the model and the training process is done in 3 lines.

This machine learning library can show the values that the model creates after the training process, these values are equivalent to the thetas array from the first part of this exercise.

Once a prediction is done, there are two pending tasks. First one is plotting the results vs the training data to get a better vision of the results and second one is to use the MSE as a metric from Sklearn library and evaluate our model.

As the first part, we evaluate or model in two different moments. One evaluation was done with the prediction of the train data and the second one was done with the test data.

	MSE
Value before train	213665.79324284595
Value After train	221997.8281363019

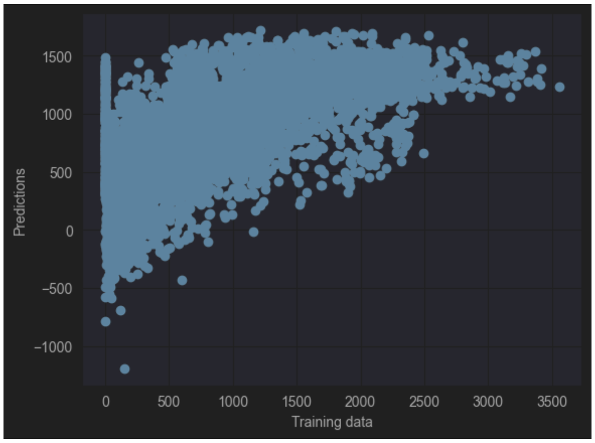
Conclusions

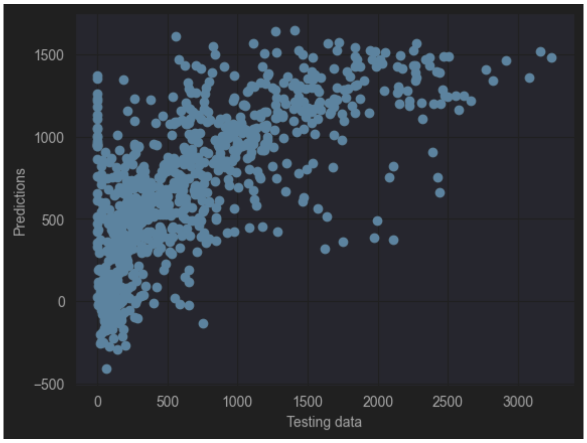
The use of 2 different tools of the regression model provides a different vision of the problems.

For the first part of the exercise, the MSE results were presented before, and the model improved after the training. This indicates that the mathematical procedures were successfully applied, the model find the correct thetas and it can be used to predict values for the future.

For the second part of the exercise, the results from the MSE were weird. A high value was obtained in the two measures. One possible reason can be the model was over fitting.

The images below are the comparison between the training data and the predictions, and the test data with the predictions. We can see the data is concentrated in one area.





Another reason for these weird results from MSE, can be that is the wrong metric. If we apply the R2 score metric, the result is reasonable.

As a conclusion, note that is very important choose the correct libraries, process, can be the difference in the results of the problems. Try different methods for the same problem can provide better vision of the data.

BIBLIOGRAPHY

MSalty. (2018). *Stackoverflow*. Retrieved from https://stackoverflow.com/questions/22216076/unicodedecodeerror-utf8-codec-cant-decode-byte-0xa5-in-position-0-invalid-s

Barrera, P. O. (2022). Ecobici y los sistemas de bicicletas públicas en las megaciudades, evolución y análisis comparativo. *Facultad de Ingeniería (UNAM)*.

UC Irvine Machine LEarning Reporsitory. (2020, 02 29). Retrieved from https://archive.ics.uci.edu/dataset/560/seoul+bike+sharing+demand

Population Stat. (2023). Retrieved from https://populationstat.com/south-korea/seoul