Machine Learning

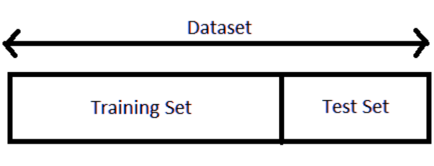
- ML is a technique followed to make a computer learn from the previous experience and make an assumption for the future outcome.
- It can learn and adapt to the new data without any human intervention.
- It needs prior training so that it can be tested to the new data.

Credits - Image from Internet (www.favouriteblog.com)

In [ ]:

ML Dataset

In machine learning, we divide the dataset into two.



- **Training Data** - Here we train the machine learning model by showing both inputs and outputs.
- **Testing Data** - Here we test the model where inputs (new data) are not mapped with outputs. We will check the model performance in terms of accuracy.

Credits - Image from Internet

In [ ]:

Supervised Learning

The computer is presented with both example inputs and their respective outputs. The algorithm learns a general rule to map the inputs to the outputs.

Training stage

We show the model a set of inputs along with the respective outputs. The task of the model is to learn by mapping the inputs and outputs.

The model is trained from the dataset that is fed. It will completely learn from it.

Testing stage

We show the model a set of new inputs without the respective outputs. The aim of the model is to predict based on the learning it had undergone.

The model predicts the category based on the previous training or learning.

Images by Author

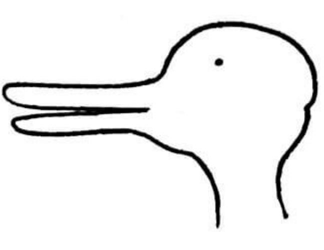
Note

- Algorithms learn from data.
- They find -
  - relationships
  - develop understanding
  - make decisions
  - evaluate their confidence from the training data they are given.
- The better the training data is, the better the model performs.

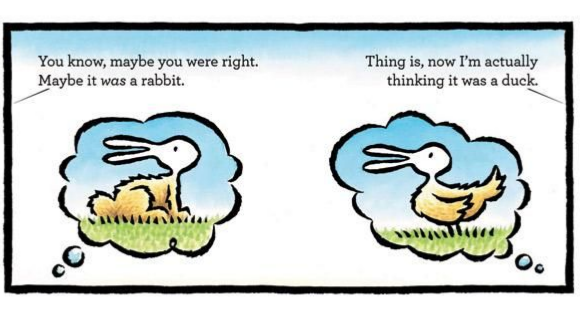
Exception case for the above example

1. Suppose I have trained my model to identify/separate `duck` and `rabbit` images from the large dataset.

1. Now, I need to test my model with the new data.
   - New image



   - Is it duck or rabbit?



2. What can you say about this?

Credits - Images from Internet

In [ ]:

Regression

- Regression is a process of predicting the dependant variable based on the independant variable.
- Dependant variable is always considered as Y.
- Independant variable is always considered as x.
- For doing regression analysis, there needs to be a strong relationship betweem x and Y.

Example - Predicting the expenses of an employee based on his/her income.

- Here,
  - expenses are dependant variable $\rightarrow Y$
  - income is independant variable $\rightarrow x$

How can we predict Y from x?

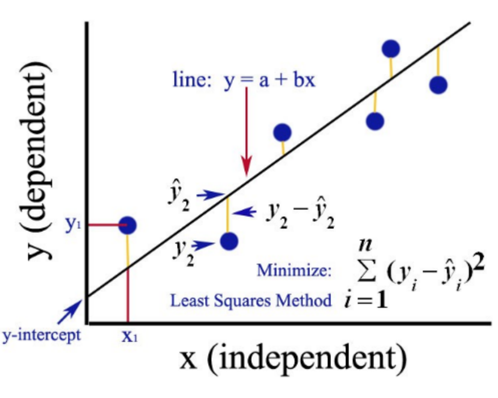
To predict Y from x, we follow a mathematical model -

$$Y = bx + a$$

- Here,
  - $b \rightarrow$ Co-efficient parameter
  - $a \rightarrow$ Bias parameter
- Other terms
  - $\hat{Y}$ (Y hat) $\rightarrow$ Predicted Y value
  - $Y \rightarrow$ Actual value

What is Error?

- The distance between $\hat{Y}$ and Y is called Error.
- Our aim while building the Regression model is to minimize the Error.



Credits - Image from Internet

In [ ]:

In [ ]:

Let's predict something ... !

`import` packages

In [1]:
```
import warnings
warnings.filterwarnings("ignore")
```

To ignore warnings that arise while executing the library method

In [2]:
```
import pandas as pd
import numpy as np
import seaborn as sns
from matplotlib import pyplot as plt
```

If you do not have `seaborn`

```
pip install seaborn --user
```

Data exploration

In [3]:
```
# read data
df = pd.read_csv('brain_body_weight.csv')
```

In [4]:
```
# shape
df.shape
```

Out[4]:
```
(54, 3)
```

In [5]:
```
# head
df.head()
```

Out[5]:

	Index	Brain Weight	Body Weight
0	1	3.385	44.5
1	2	0.480	15.5
2	3	1.350	8.1
3	5	36.330	119.5
4	6	27.660	115.0

In [6]:
```
# remove Index
df = df.drop(columns=['Index'], axis=1)
```

In [7]:
```
# shape
df.shape
```

Out[7]:
```
(54, 2)
```

In [8]:
```
# head
df.head()
```

Out[8]:

	Brain Weight	Body Weight
0	3.385	44.5
1	0.480	15.5
2	1.350	8.1
3	36.330	119.5
4	27.660	115.0

- Dependant variable $(Y) \rightarrow$ Brain Weight
- Independant variable $(x) \rightarrow$ Body Weight

In [9]:
```
# Create Y and x (array)
Y = np.array(df['Brain Weight'])
x = np.array(df['Body Weight'])
```

How x is related to Y?

In [10]:
```
# correlation
bb_corr = df.corr()
```

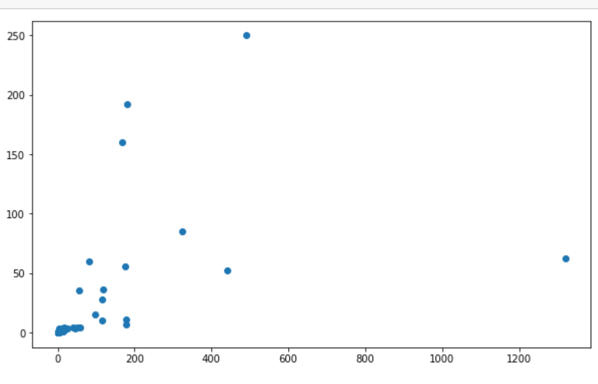
In [11]:
```
bb_corr
```

Out[11]:

	Brain Weight	Body Weight
Brain Weight	1.000000	0.490261
Body Weight	0.490261	1.000000

From above, we can say `Brain Weight` and `Body Weight` are highly correlated.

In [12]:
```
# scatter plot (10, 6)
plt.figure(figsize=(10, 6))
plt.scatter(x, Y)
plt.show()
```



There are outliers in our data.

Outlier detection function

In [13]:
```
def calculate_zscore(data_values):
    # mean
    mean_vals = np.mean(data_values)
    # standard deviation
    std_dev = np.std(data_values)
    # applying the formula for all the values
    zscore = np.array([(i - mean_vals)/std_dev for i in data_values])
    return zscore
```

Model Implementation

- More info → https://towardsdatascience.com/gradient-descent-animation-1-simple-linear-regression-e49315b24672

In [14]:
```
# b, a - polyfit
b, a = np.polyfit(x, Y, 1)
```

In [15]:
```
# b
b
```

Out[15]:
```
0.1192489488025367
```

In [16]:
```
# a
a
```

Out[16]:
```
10.709996660625379
```

Broadcasting

In [17]:
```
f = [1, 2, 3, 4] # → [3, 6, 9, 12]
f = [i*3 for i in f]
f
```

Out[17]:
```
[3, 6, 9, 12]
```

In [18]:
```
f = [1, 2, 3, 4] * 3
f
```

Out[18]:
```
[1, 2, 3, 4, 1, 2, 3, 4, 1, 2, 3, 4]
```

In [19]:
```
g = np.array([1, 2, 3, 4])
g
```

Out[19]:
```
array([1, 2, 3, 4])
```

In [20]:
```
g * 3
```

Out[20]:
```
array([ 3,  6,  9, 12])
```

In [ ]:

In [21]:
```
# y_preds = bx + a
y_preds = b*x + a
```

In [22]:
```
# y_preds
y_preds
```

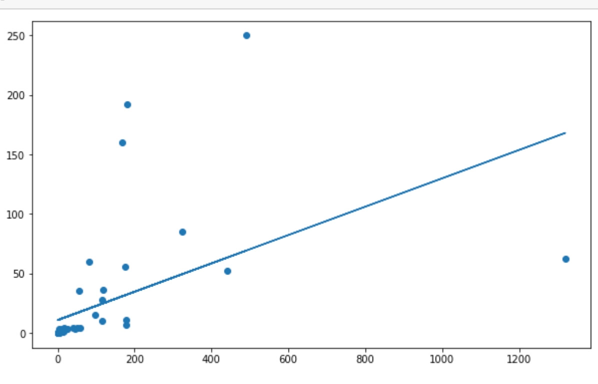
Out[22]:
```
array([ 16.01657488,  12.55835537,  11.67591315,  24.96024604,
        24.42362577,  22.42024343,  11.36586588,  17.62643569,
        11.47318993,  11.18699246,  11.38971567,  11.49703972,
        10.72669151,  10.82924561,  11.99788531,  12.17675873,
        11.46126504,  10.74577135,  11.12736798,  24.42362577,
        13.76276975,  11.3062414 ,  12.79685326,  49.46590502,
        12.17675873, 168.11860908,  11.17506756,  32.0555585 ,
        17.38793779,  12.73722879,  10.82924561,  10.75769624,
        10.7398089 ,  12.20060852,  69.14198157,  12.15290894,
        31.5785627 ,  63.17953413,  32.11518297,  10.99619414,
        20.36916151,  13.21422459,  15.38455545,  10.93656966,
        10.8530954 ,  11.06774351,  10.74934881,  32.17480745,
        13.69122038,  30.86306901,  11.02004393,  12.06943468,
        11.00811903,  16.72014368])
```

In [23]:
```
Y
```

Out[23]:
```
array([3.385e+00, 4.800e-01, 1.350e+00, 3.633e+01, 2.766e+01, 1.483e+01,
       1.040e+00, 4.190e+00, 4.250e-01, 1.010e-01, 9.200e-01, 1.000e+00,
       5.000e-03, 6.000e-02, 3.500e+00, 2.000e+00, 1.700e+00, 2.300e-02,
       7.850e-01, 1.000e+01, 3.300e+00, 2.000e-01, 1.410e+00, 8.500e+01,
       7.500e-01, 6.200e+01, 3.500e+00, 6.800e+00, 3.500e+01, 4.050e+00,
       1.200e-01, 2.300e-02, 1.000e-02, 1.400e+00, 2.500e+02, 2.500e+00,
       5.550e+01, 5.216e+01, 1.055e+01, 5.500e-01, 6.000e+01, 3.600e+00,
       4.288e+00, 2.800e-01, 7.500e-02, 1.220e-01, 4.800e-02, 1.920e+02,
       3.000e+00, 1.600e+02, 9.000e-01, 1.620e+00, 1.040e-01, 4.235e+00])
```

Best fit line

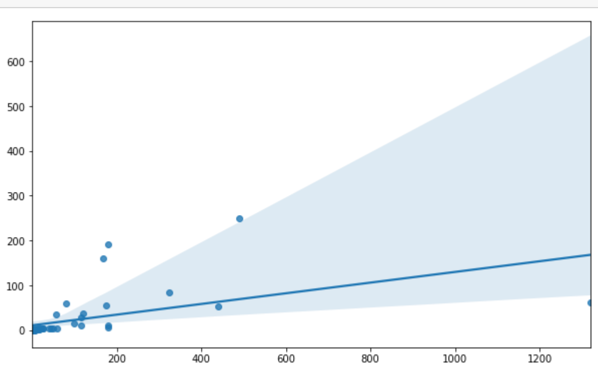
In [24]:
```
# scatter and line
plt.figure(figsize=(10, 6))
plt.scatter(x, Y)
plt.plot(x, y_preds)
plt.show()
```



In [ ]:

Seaborn plot with `confidence` level

In [25]:
```
# confidence level
plt.figure(figsize=(10, 6))
sns.regplot(x, Y)
plt.show()
```



In [ ]:

Plotly plot for interactivity

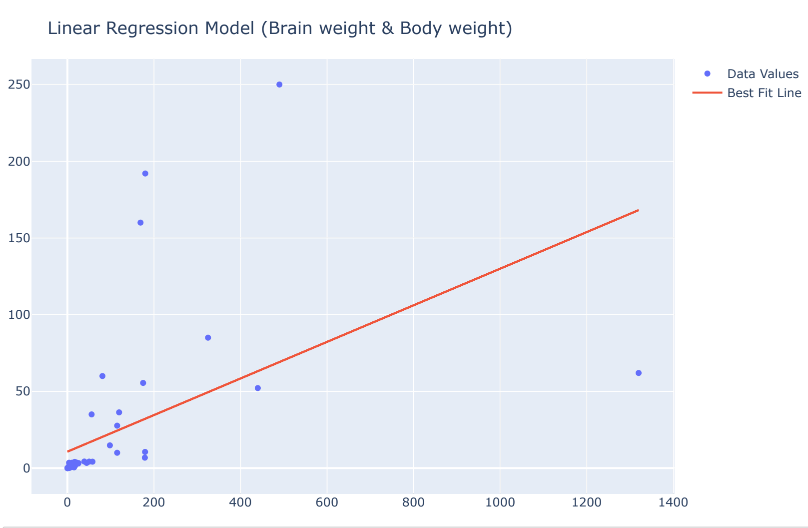
In [26]:
```
import plotly.graph_objects as go
```

In [27]:
```
trace1 = go.Scatter(
    x=x,
    y=Y,
    mode='markers',
    name='Data Values'
)

trace2 = go.Scatter(
    x=x,
    y=y_preds,
    mode='lines',
    name='Best Fit Line'
)

layout = go.Layout(
    title='Linear Regression Model (Brain weight & Body weight)',
    height=500,
    width=800,
    margin=dict(l=0, b=10, t=50, r=0)
)

fig = go.Figure(data=[trace1, trace2], layout=layout)
fig.show()
```



In [ ]:

What about Multiple Linear Regression model?

Dependant variable is depends on mutiple Independant variables.

Model -

$$Y = b_1x_1 + b_2x_2 + b_3x_3 + \cdots + b_nx_n + a$$

- $x_1, x_2, x_3, \ldots, x_n$ are independant variables.
- $b_1, b_2, b_3, \ldots, b_n$ are coefficient variables.
- a is bias variable.

The main aim of any machine learning model is to minimize the error.