Jahnavi Nuthalapati

MACHINE LEARNING-2208-CSE-6363-001
INSTRUCTOR : DAJIANG ZHU
FALL 2020

ASSIGNMENT 1 - REPORT

IMPLEMENTATION
Environment/IDE : Google Colab/Jupyter Notebook in Anaconda-Navigator
Programming Language : Python

DESCRIPTION
Given is an Iris Dataset. So the task is to train the model using Linear Regression, use this trained model to classify on the test data and use Cross Validation.

ABOUT IRIS DATASET

- So the data consists of three different type of flowers namely Iris-Setosa, Iris-Versicolor and Iris-Virginica and it's attributes are sepal length (cm), sepal width (cm), petal length (cm), petal width (cm).
- It has 150 samples pertaining to three of them as length and width of petals and sepals.
- So this dataset will be used so as to create a linear differentiation among them in order to know which belongs to which class.

So what's Linear Regression? It's simply used to find the relation between an independent and a dependent variable. Like how important a variable is in order to find the target result and the precise accuracy.

So in a standard way, a regression equation is given as $Y= \alpha + \beta X + E$, where Y is the dependent variable, α is the population Y intercept, β is the population slope coefficient, X is an independent variable and E is the error in equation (Equation courtesy : google images Pinterest)

Now the most important is the regression line, which tells that how a parameter changes with respect to another.

So in if we regression, we can know the relation between two parameters, provided those have been provided and it's behavior by plotting on graph and finally if a new set is given, we can also predict the accuracy of fit in the given data.

IMPORTING THE REQUIRED LIBRARIES

- Firstly, we have to load the given data “iris.data” which has One Hundred and Fifty rows and five attributes
- Numpy, Pandas and Matlplotlib will be imported so as to support the working of functions.
- Matplotlib.pyplot is used to depict the information in the form of figures by plotting necessary points on the graphs.
- Seaborn is used in making visualizations of random distributions of data
- Sklearn is a library which is one of the most prominent imports in the code as it’s a tool which supports modeling like regression, clustering, etc.. which helps in setting up various ML models.

```
In [10]: import numpy as np
         import matplotlib.pyplot as plt
         import pandas as pd
```

```
In [11]: %matplotlib inline
         import numpy as np
         from sklearn import datasets
         import seaborn.apionly as sns
         import matplotlib.pyplot as plt
```

LOADING THE IRIS DATASET INTO THE TOOL

- This step involves the import of data into the tool. This will done by making use of seaborn library with load_dataset() function.

COVARIANCE AND CORRELATION OPERATION

- Covariance is defined as a measure on how two parameters are related or relationship between them.
- So inorder to use this concepts, values [3,5,7], [2,6,9] and [3,5,7], [2,6,9] have been considered. This has been done using Numpy library.
- Note here both covariance and correlation have been used in order to determine how they’re related and also the direction of their relation between them.

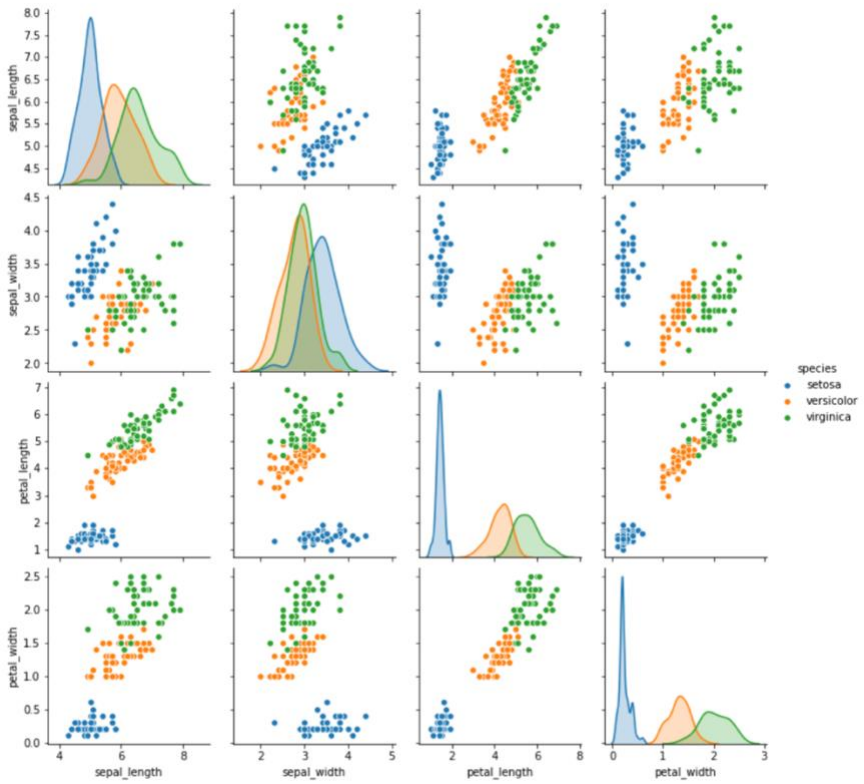
CROSS VALIDATION IMPLEMENTATION

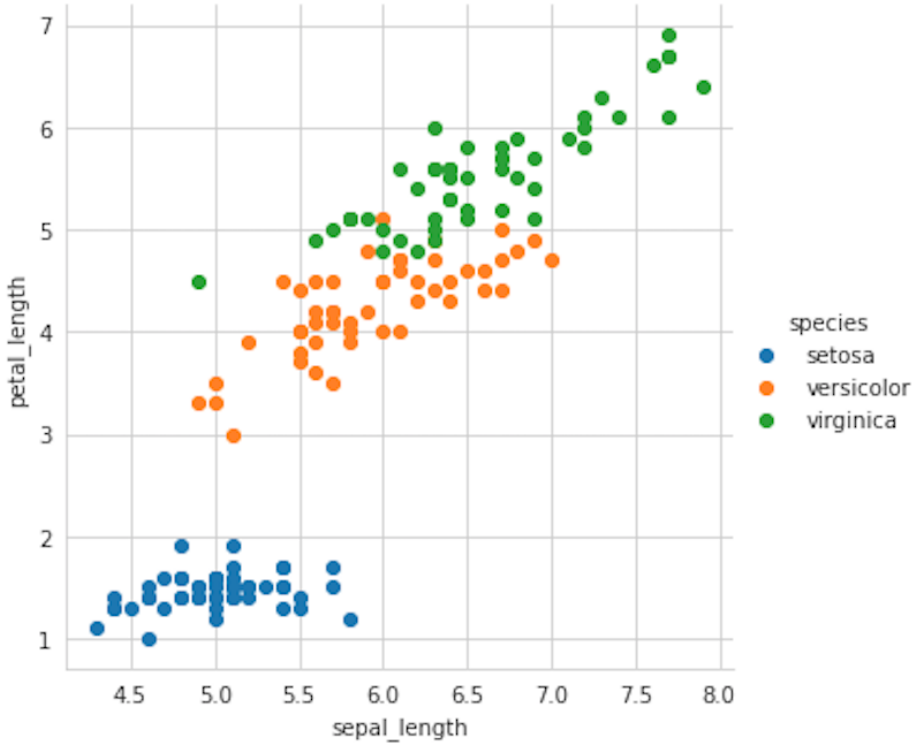
- Cross validation is used in guessing or predicting the outcome on the parameters which aren't present during the training model i.e unknow data, so it's represented as an equation like linear regression where error function is separately coded. This gives the preciseness of the validation
  def predict(a, b, f): return b * f + a

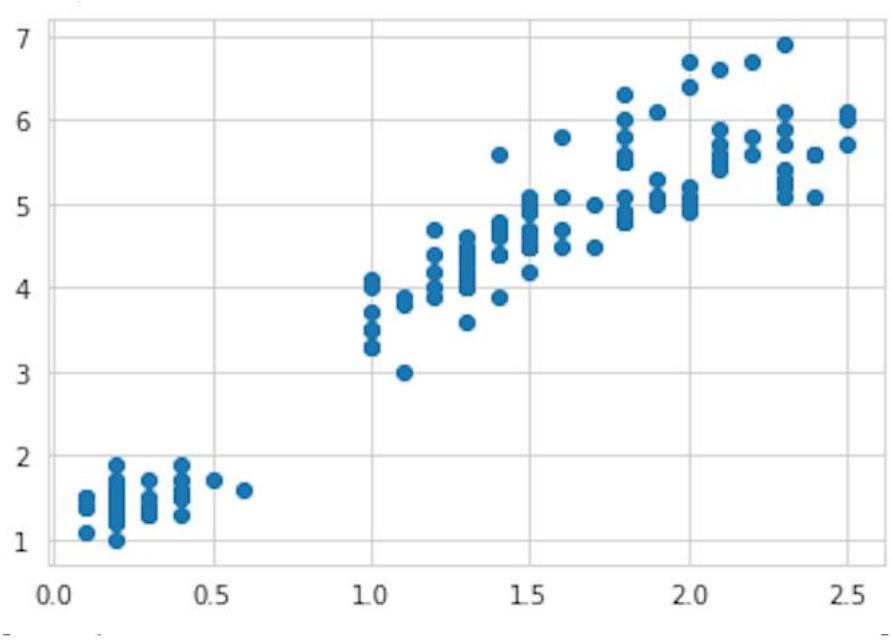
- Now, for error, this is used in determining the divergence of guessing the outcome and it's accuracy. So two values are taken 'a' and 'b'. Because that difference can be linear or more than that i.e exponential, double, etc.. So def error is difference of two values and def sum is the variant.
  def error(a, b, f, j):
  def sum_sq_e(a, b, i, j):
  return sum(error(a, b, f, j) ** 2 for f, j in zip(i, j))

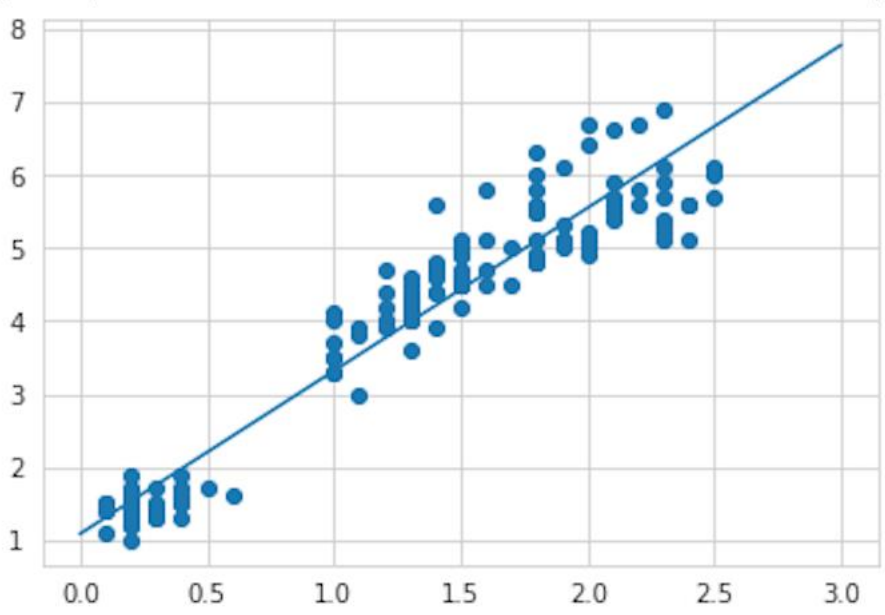
GRAPH PLOT

- In order to represent and classify the given data more effectively, few of the different graph representations, scatter plots have been used, which can be found below which have been done using Matplotlib and seaborn.
- So the plotting will be done by providing the computed values, so that classification using those functions and methods can be clearly seen on the graphs.


sepal_length
sepal_width
petal_length
petal_width
species
setosa
versicolor
virginica


petal_length
sepal_length
species
setosa
versicolor
virginica





- So we take two parameters, find if there's any error and fit them (if needed) then store in them in two variables then depict then in the form of a scatter plot which is fit linearly.

References

- https://mc.ai/visualization-and-understanding-iris-dataset/
- https://medium.com/@randerson112358/python-logistic-regression-program-5e1b32f964db
- https://www.youtube.com/watch?v=BrFEmO-zPuA
- https://stackoverflow.com/questions/54317168/plotting-a-dataframe-with-seaborn-pairplot-in-multiple-colors