Machine Learning Assignment-2

Group no - 35

Identification of different species(Iris-virginica, Iris-versicolor) of iris-flowers.

Logistic Regression:

Generally logistic regression is used for classification purposes. Unlike linear regression , the dependent variable can take limited number of values only i.e, the dependent variable is categorical.

Here in our iris data set we can observe that the number of possible outcomes is only two. So we can say that we can identify the species using binary logistic regression.

Implementation:

We implemented the algorithm by broking down into 5 steps for both the questions a and b

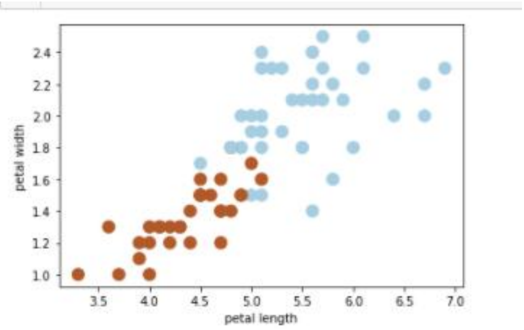
1. Load the train data and test data
2. Observing the data
3. Training the model(estimating coefficients)
4. Predicting
5. Calculating accuracy of model

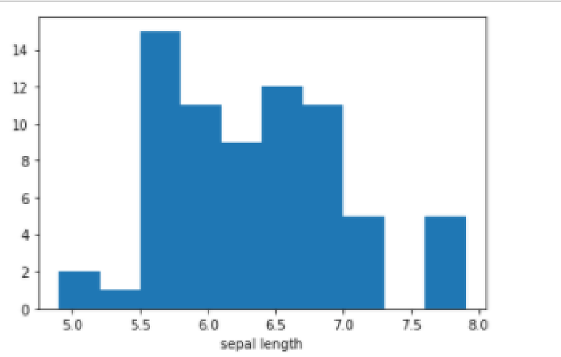
1) Loading the train data and test data:

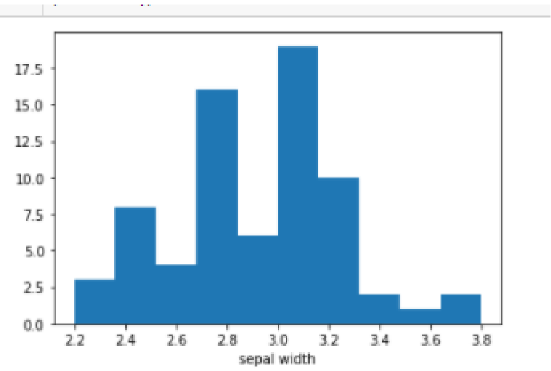
We had used pandas library to read the train data and test data. We had used read_excel for reading the train data where as read_csv for reading the test data.

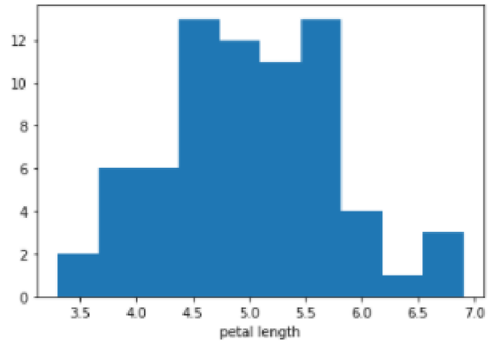
2) Observing the data:-

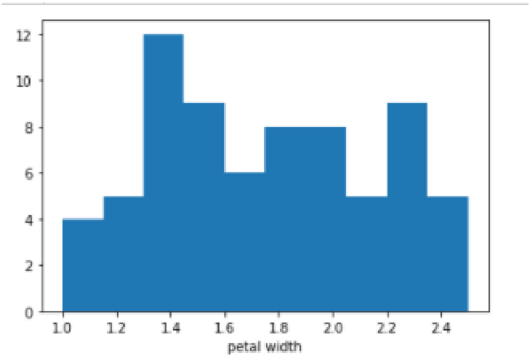
We observed the graphs as shown below and understood the data.











3) Training the model:

In this we had first developed a function to predict which will be used in the evaluation of coefficient values in stochastic gradient descent and after the model is finalized and we wish to start making predictions on test data or new data.

We can estimate the coefficient values for our training data using stochastic gradient descent.

Generally stochastic gradient descent requires two parameters in second part

- **Learning Rate:** Used to limit the amount each coefficient is corrected each time it is updated.
- **Epochs:** The number of times to run through the training data while updating the coefficients.

Whereas we haven't used epochs in the first part we stopped with the help of Euclidian norm of vector

We have used the formula **b = b + learning_rate * (y - yhat) * x** **.** Which is discussed in the class for updating of co-efficients .

Where Where **b** is the coefficient or weight being optimized, **learning_rate** is a learning rate that you must configure (e.g. 0.02), **(y – yhat)** is the prediction error for the model on the training data attributed to the weight, **yhat** is the prediction made by the coefficients and **x** is the input value.

4) Prediction:

After loading the train data and test data and training the model. We had passed the features to predict function with the predicted coefficients and the returned data is collected and passed through accuracy_metric function to calculate the accuracy.

5) Accuracy:

We passed the predictions along with actual to the accuracy_metric function to calculate the accuracy.

Observations:

By training the model we have got 100 percent accuracy for both the methods. We observed that the a part a have taken more time than that of part b .

Prepared by:

E Pavan Sai Santhosh	S20180010053
M Sai Nikhil	S20180010095
B Raj Kumar	S20180010026