CSE 6363 – Machine Learning

Project 1

Done By-

Aniket Gade
UTA ID - 1001505046

1) Problem

Classification of Iris plants dataset.

Number of Instances: 150 (50 in each of three classes)
Number of Attributes: 4
Attribute Information:
a) sepal length in cm b) sepal width in cm c) petal length in cm d) petal width in cm

Class Information:
Iris-Setosa (0) Iris-Versicolour (1) Iris-Virginica (2)

Summary Statistics:

	MIN	MAX	Mean	SD	CORRELATION
SEPAL LENGTH	4.3	7.9	5.84	0.83	0.7826
SEPAL WIDTH	2.0	4.4	3.05	0.43	-0.4194
PETAL LENGTH	1.0	6.9	3.76	1.76	0.9490
PETAL WIDTH	0.1	2.5	1.20	0.76	0.9565

2) Method

The method used here is the least square estimator for linear regression.
The $p \times 1$ vector containing the estimates of the p parameters of the regression function can be shown to equal:

$$b = \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_{p-1} \end{bmatrix} = (X^{'}X)^{-1}X^{'}Y$$

where:

- $(X'X)^{-1}$ is the **inverse** of the $X'X$ matrix, and
- X' is the **transpose** of the X matrix.

Once the beta matrix is calculated, we multiple it with the test set to predict the test data set.

Before we use linear regression to classify, we first must split the data set into training and testing data. For this we use k-fold cross validation. The procedure has a single parameter called k that refers to the number of groups that a given data sample is to be split into. Steps for k-fold cross validation are as follows:

1) Shuffle the dataset randomly.
2) Split the dataset into k groups
3) For each unique group:
    - a) Take the group as a hold out or test data set
    - b) Take the remaining groups as a training data set
    - c) Fit a model on the training set and evaluate it on the test set
    - d) Retain the evaluation score and discard the model
4) Summarize the skill of the model using the sample of model evaluation scores

3) Results

The following table summarizes the different k-fold cross validations:

K-Fold (n)	Accuracy
2	96.66666666666667 %
3	96.66666666666667 %
4	96.62162162162162 %
5	97.33333333333334 %
10	96.66666666666666 %
Leave one out	96.64429530201343 %

As seen above, for this machine learning model, 5-fold CV performs the best in terms of accuracy. Also, as the data set is relatively small (150 instances) the computation is not high for a 5-fold CV. Hence in the code, I have selected k=5.

The code gives output as follows:

```
Actual Labels - [2 2 2 2 1 1 2 1 1 0 2 1 0 2 2 2 2 0 1 0 2 0 1 1 2 0 2 0 2 1]
Predicted Labels -  [2 2 2 2 2 2 2 1 1 0 2 2 0 2 2 2 2 0 1 0 2 0 1 1 2 0 2 0 2 1]
Accuracy =  0.9

Actual Labels - [2 1 1 0 0 0 2 2 1 0 0 1 0 0 0 0 2 2 0 2 0 0 1 1 0 0 0 2 2 1]
Predicted Labels -  [2 1 1 0 0 0 2 2 1 0 0 1 0 0 0 0 2 2 0 2 0 0 1 1 0 0 0 2 2 1]
Accuracy =  1.0

Actual Labels - [2 2 0 2 2 0 2 2 1 0 0 1 2 2 1 1 2 1 0 1 2 2 1 0 2 1 1 2 1 1]
Predicted Labels -  [2 2 0 2 2 0 2 1 1 0 0 1 2 2 1 1 2 1 0 1 2 2 1 0 2 1 1 2 1 1]
Accuracy =  0.9666666666666667

Actual Labels - [1 2 0 0 0 0 1 0 0 2 1 2 2 1 0 1 1 1 2 0 0 1 2 1 0 2 1 2 1 2]
Predicted Labels -  [1 2 0 0 0 0 1 0 0 2 1 2 2 1 0 1 1 1 2 0 0 1 2 1 0 2 1 2 1 2]
Accuracy =  1.0

Actual Labels - [1 1 0 2 1 1 0 2 0 0 1 2 0 1 1 2 0 2 0 0 1 0 1 1 1 0 2 1 0 0]
Predicted Labels -  [1 1 0 2 1 1 0 2 0 0 1 2 0 1 1 2 0 2 0 0 1 0 1 1 1 0 2 1 0 0]
Accuracy =  1.0

Average Accuracy for 5 fold CV : 0.9733333333333334
```

In this instance the training data set has 120 instances and the testing data set has 30 instances.