CS60050 MACHINE LEARNING

Assignment 2

Report

Group Members:

Rahul Mandal (20CS30039)

Sailada Vishnu Vardhan (20CS10051)

Q2. Files:

- main.py

This file contains operations that are asked in part 2 of assignment.

Some of the defined functions in it are:

1. splitting_data(): this function splits dataset into a specified ratio for training and testing purpose.
2. If_null_value(): this function detects any empty data and fills it with the mean of the column.
3. standard_scalar_normalize(): this function performs standard scalar normalization on while data (except label column).
4. def forward_sel(): this function performs the forward selection method to get best set of features.

- lung-cancer.data

This file contains our dataset. There is another file lung-cancer.names which describes our data.

Standard Scalar Normalization

Initially the standard scalar normalization is performed using the function standard_scalar_normalize(). No standard library was used to perform this operation.

Function basically performs

$$z = \frac{x - \mu}{\sigma}$$

Where, z = normalized value

μ = mean of column

σ = standard deviation

Binary SVM classifier

The SVM(support vector machines) classifier is used to find a hyperplane in an n-dimensional space that separates the data points to their potential classes.

SVMs has some kernels which converts the input data space into a higher-dimensional space.

Kernel's we have used are: Linear, Quadratic, Radial Basis.

For SVM we used sklearn's library to implement this classifer.

Result of SVM classifier with required kernels for a random split of data in to training/test are as follows:

```
Accuracy for SVM Classifier using linear kernel is:
83.33333333333334
Accuracy for SVM Classifier using quadratic kernel is:
83.33333333333334
Accuracy for SVM Classifier using radial kernel is:
16.666666666666664
```
d

MLP Classifier

MLP (Multi-layer Perceptron) classifier unlike SVM classifier relies on neural network to perform classification.

sklearn library is used to implement this.

Stochastic gradient descent optimiser was used to implement MLP with keeping learning rate as 0.001 and batch size of 32. Then the accuracy was calculated by varying number of hidden layers with respective nodes.

Result for same are as follows for a random split of dataset.

```
Accuracy for MLP Classifier using Stochastic Descent Gradient with 1 hidden layer with 16 nodes is:
33.33333333333333
Accuracy for MLP Classifier using Stochastic Descent Gradient with 2 hidden layers with 256 and 16 nodes is:
83.33333333333334
```

Graph for learning rate vs accuracy

By varying learning rate as 0.1, 0.01, 0.001, 0.0001, 0.00001.

We see how accuracy changes as follows through a graph

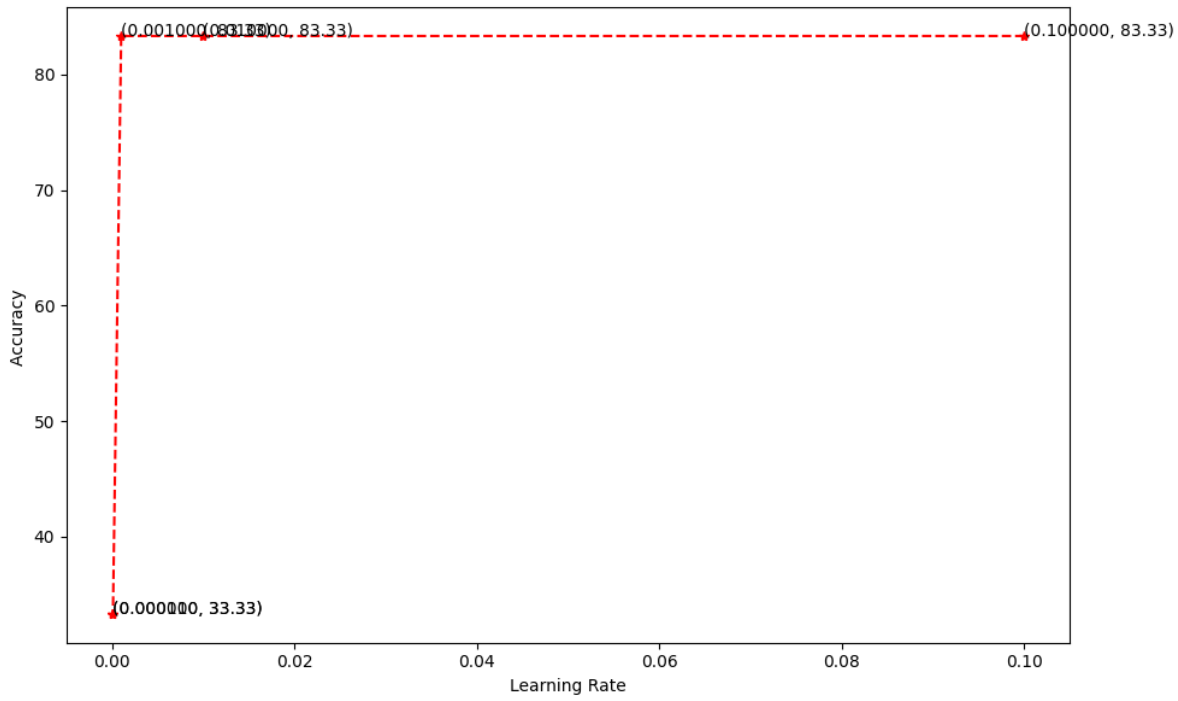
For Learning Rate = 0.00001 we have accuracy = 33.334,

Learning Rate = 0.0001 we have accuracy = 83.334,

Learning Rate = 0.001 we have accuracy = 83.334,

Learning Rate = 0.01 we have accuracy = 83.334

Learning Rate = 0.1 we have accuracy = 83.334



Forward selection method:

This method uses best model from MLP classifier models to select best set of features which gives us high accuracy.

The data contains 57 columns and 32 rows.

The best set of final feature's index is printed in output.txt file.

Ensemble learning (max voting technique)

For applying ensemble learning using SVM kernels of quadratic, radial basis function and best accuracy model of MLP is used.

The final accuracy is as follows:

```
Final Accuracy after ensemble learning with max voting method is:
83.33333333333334
```