CS60050 : Machine Learning

Assignment 2 Report

Group 21

Sharannya Ghosh (20CS10054)
Aritra Mitra (20CS30006)

Problem 1

1. Normalise the data using Standard Scalar Normalisation. Randomly divide the Dataset into 80% for training and 20% for testing. Encode categorical variables using appropriate encoding method (in-built function not allowed for normalization, sampling and encoding).
2. Implement the binary `SVM` classifier using the following kernels: Linear, Quadratic, Radial Basis function. Report the accuracy for each. (in-built function allowed).
3. Build an `MLP` classifier (in-built function allowed). for the given dataset. Use stochastic gradient descent optimiser. Keep learning rate as 0.001 and batch size of 32. Vary the number of hidden layers and number of nodes in each hidden layer as follows and report the accuracy of each:
   (a) 1 hidden layer with 16 nodes
   (b) 2 hidden layers with 256 and 16 nodes respectively.
4. Using the best accuracy model from part 3, vary the learning rate as 0.1, 0.01, 0.001, 0.0001 and 0.00001. Plot the learning rate vs accuracy graph.
5. Use backward elimination method on the best model found in part 3 to select the best set of features. Print the features.
6. Apply ensemble learning (max voting technique) using `SVM` with quadratic, `SVM` with radial basis function and the best accuracy model from part 3. Report the accuracy.
7. Prepare a report including all your results.

Solution

1 Dataset

The relevant information regarding the `Iris` dataset is as follows:

Number of Attributes: 4 numeric, predictive attributes and the class

Attribute Information:
1. sepal length in cm
2. sepal width in cm
3. petal length in cm
4. petal width in cm
5. class
   - Iris Setosa
   - Iris Versicolor
   - Iris Virginica

Number of Instances: 150 (50 in each of three classes)

2 Encoding

There was only one non-numerical attribute, the `class`. It was encoded as follows:

Iris-setosa: 0

Iris-versicolor: 1

Iris-virginica: 2

3 Support Vector Classifier

The *sklearn* library was used for this task, and the classifier was imported by calling an instance of the class `sklearn.SVM.SVC`. It was called thrice, once for the classification of each class from the rest.

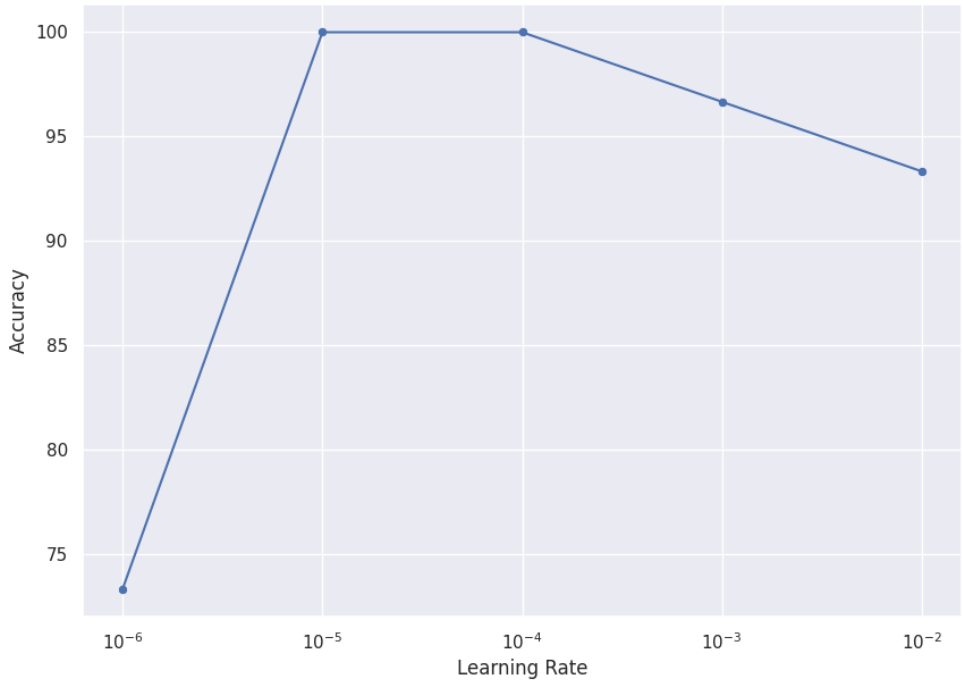
4 Multi-Layer Perceptron

The *sklearn* library was used once again for this task, and the classifier was imported by calling an instance of the class `sklearn.neural_network.MLPClassifier`. It was called twice, with different parameters as follows:

1. `hidden_layer_sizes=(16),batch_size=32,learning_rate='constant',learning_rate_init=0.001`
2. `hidden_layer_sizes=(256,16),batch_size=32,learning_rate='constant',learning_rate_init=0.001`

5 Different Learning Rates

The learning rate was varied as $0.1, 0.01, 0.001, 0.0001, 0.00001$ and the **accuracy vs. learning rate** graph is as follows:



6 Backwards Elimination

A function for backwards elimination was written and applied the best model from section 4. No feature was eliminated in this process, and all the 4 features were still there.

7 Ensemble Learning

As the SVMs are binary classifier, whereas the MLP classifier is ternary, we cannot perform ternary ensemble learning. We chose two alternatives:

- Using the ternary-to-binary functions used for the SVMs on the MLP predictions, and perform ensemble learning.
- Training binary MLP classifiers by applying the ternary-to-binary functions to the dataset, and perform ensemble learning.

8 Output

```
   SVM Accuracy Score for binary classification of setosa from rest using linear kernel ->
100.0
SVM Accuracy Score for binary classification of setosa from rest using radial basis function
kernel -> 96.66666666666667
SVM Accuracy Score for binary classification of setosa from rest using quadratic kernel ->
80.0
SVM Accuracy Score for binary classification of versicolor from rest using linear kernel ->
63.33333333333333
SVM Accuracy Score for binary classification of versicolor from rest using radial basis function
kernel -> 100.0
SVM Accuracy Score for binary classification of versicolor from rest using quadratic kernel
-> 100.0
SVM Accuracy Score for binary classification of virginica from rest using linear kernel ->
100.0
SVM Accuracy Score for binary classification of virginica from rest using radial basis function
kernel -> 100.0
SVM Accuracy Score for binary classification of virginica from rest using quadratic kernel
-> 46.666666666666664
MLP Classifier accuracy score for Part a.  -> 96.66666666666667
MLP Classifier accuracy score for Part b.  -> 100.0
Features:
Sepal Width, Sepal Length, Petal Width, Petal Length,
Accuracy Score for Ensemble Learning of setosa -> 96.66666666666667
Accuracy Score for Ensemble Learning of versicolor -> 100.0
Accuracy Score for Ensemble Learning of virginica -> 100.0
Accuracy Score for Ensemble Learning of setosa using alternative method-> 80.0
Accuracy Score for Ensemble Learning of versicolor using alternative method-> 100.0
Accuracy Score for Ensemble Learning of virginica using alternative method-> 100.0
```