Machine Learning (CS60050)

Assignment-3 Report

Supervised Learning
Group 50
Gaurav Malakar - 20CS10029
Atishay Jain - 20CS30008

Dataset

Title of Database: Abalone data
Predicting the age of abalone from physical measurements. The age of abalone is determined by cutting the shell through the cone, staining it, and counting the number of rings through a microscope -- a boring and time-consuming task. Other measurements, which are easier to obtain, are used to predict the age. Further information, such as weather patterns and location (hence food availability) may be required to solve the problem.
Number of Instances: 4177
Number of Attributes: 8
Attribute information:

Given is the attribute name, attribute type, the measurement unit and a brief description. The number of rings is the value to predict: either as a continuous value or as a classification problem.

Name	Data Type	Meas.	Description
Sex	nominal		M, F, and I (infant)
Length	continuous	mm	Longest shell measurement
Diameter	continuous	mm	perpendicular to length
Height	continuous	mm	with meat in shell
Whole weight	continuous	grams	whole abalone
Shucked weight	continuous	grams	weight of meat
Viscera weight	continuous	grams	gut weight (after bleeding)
Shell weight	continuous	grams	after being dried
Rings	integer		+1.5 gives the age in years

Tasks

In this asignment our task is to create Support Vector Machine and Multilayer Perceptron classifier models that can predict the age of abalone from other given attributes. A more detailed description of the tasks is as follows:

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
2) Implement the binary SVM classifier using the following kernels: Linear, Quadratic, Radial Basis function. Report the accuracy for each.
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 nodesb.
    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.

Formulas Used

$$\text{Accuracy} = \frac{No.\,of\ corrrectly\ classified\ examples}{Total\,no.\,Of\ examples}$$

$$= 1 - \frac{No.\,of\ misclassified\ exapmles}{Total\,no.\,Of\ examples}$$

$$\text{Standard Deviation} = \sqrt{Variance}$$

Prodcedure

ENCODING:

The data was encoded by removing the column with categorical variable (‘Sex’) and adding a new column(‘Sex_encoded’) containing numeric values corresponding to each category in the original column. The mapping of category to a numeric value was as follows: ‘I’ ⟶ 0 ‘M’ ⟶ 1 ‘F’ ⟶ 2

NORMALIZATION:

Normalization was performed on the non-target columns (i.e., all columns except ‘Rings’). To normalize a column, we find the mean and standard deviation of all elements in that column. After that, we modify all elements in that column by subtracting the mean from them and dividing them by the standard deviation.

SAMPLING:

The data was divided manually without the use of any external library. For this, we created a NumPy array of length equal to the number of rows in the data set. The elements of the array are row indices going from zero to (no. of rows - 1). This NumPy array was randomly shuffled using numpy.random.shuffle() function. After this, the first 80% of rows were used for training our models, and the rest 20% of rows were used for testing our models.

BACKWARD ELIMINATION:

The Backward Elimination algorithm goes as follows,

- For all columns in the current data set, remove that column from the data set, check the model's accuracy on the new data set, and note the column whose removal produces maximum accuracy for the model.
- If the best accuracy achieved by removing a column is greater than the accuracy without removing any column, then remove that column from the current data set and run backward elimination on the newly formed dataset.
- The algorithm terminated is there is only one column left in our dataset or the best accuracy achieved by removing a column is not greater than the accuracy achieved without removing any column.

ENSEMBLE LEARNING:

Ensemble Learning was performed by making a class with three models as member variables. The class contains two member functions predict( ) and accuracy( ):-

- predict ( ): This function takes an instance as input, find the predicted value of all three models corresponding to that input, and returns the prediction which was made a maximum number of times. Since, there are only three predictions, if any prediction occurs >=2 times, then it will have maximum occurrence. If the three models make three different predictions, then we return the second-largest prediction(median in this case)

- accuracy( ): This function takes a set of examples as input and uses them to find the accuracy of the model. It simply calls the predict( ) function for each instance in the example set, matches the prediction with the actual result, and counts the number of misclassifications. The number of misclassifications is then used to find the accuracy using the formula mentioned above.

Results

Support Vector Machine (SVM) Classifier

Kernel	accuracy (in %)
Linear	27.27
Linear	26.31
Linear	25.83
Quadratic	23.20
Quadratic	22.96
Quadratic	25.59
Radial Basis Function	27.75
Radial Basis Function	24.40
Radial Basis Function	27.27

Average accuracy = 25.62%

Multilayer Perceptron (MLP) Classifier

Hidden Layer 1	Hidden layer 2	Accuracy (in %)
16 nodes	---	29.18
16 nodes	---	26.67
16 nodes	---	26.07
256 nodes	16 nodes	29.90
256 nodes	16 nodes	25.71
256 nodes	16 nodes	26.67

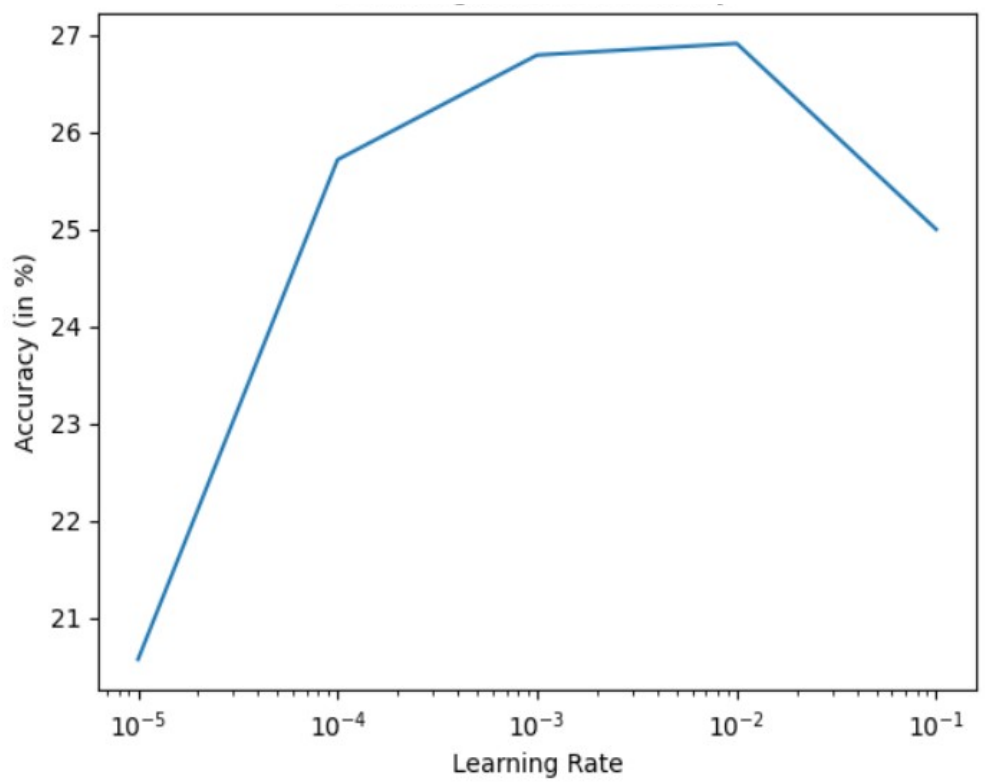
Average accuracy = 27.36%

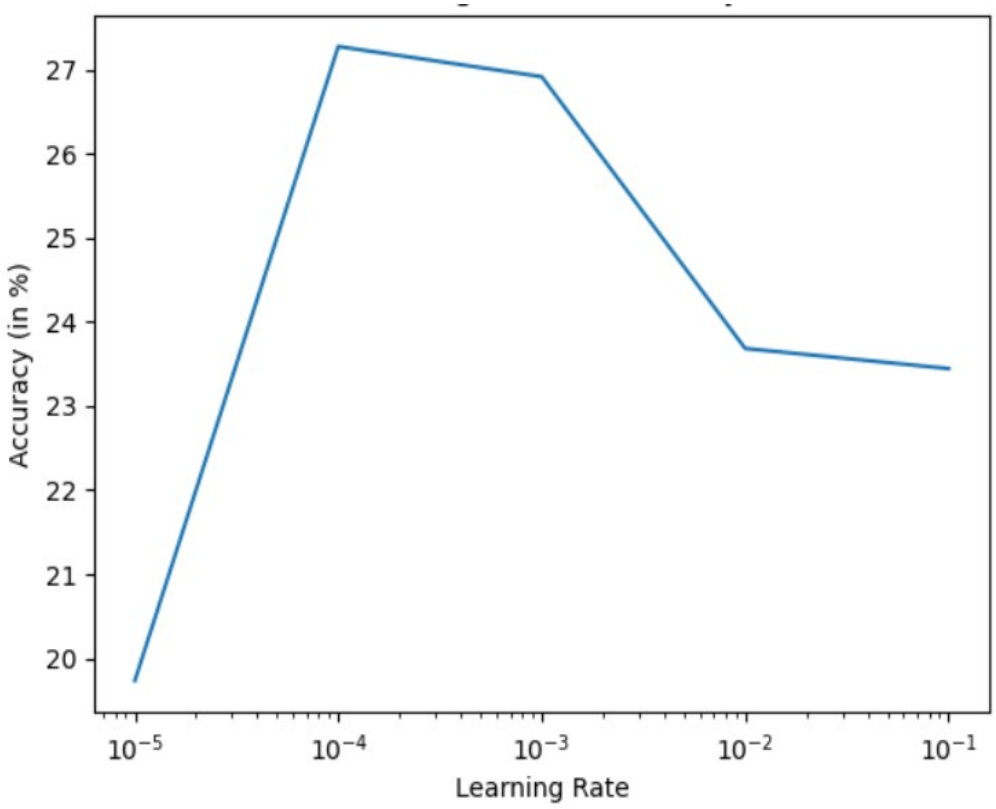
Ensemble Learning Classifier

model_1 kernel	model_2 kernel	model_3 hidden layers	accuracy (in %)
quadratic	rbf	(256, 16)	27.39
quadratic	rbf	(16)	24.40
quadratic	rbf	(256, 16)	25.83
quadratic	rbf	(256, 16)	27.63
quadratic	rbf	(256, 16)	26.67
quadratic	rbf	(256, 16)	25.95
quadratic	rbf	(16)	27.27

Average accuracy = 26.44%

Learning Rate vs Accuracy Plots





Example Outputs

Example-1

SVM classifiers
Accuracy for SVM with Linear Kernel = 27.27272727272727%
Accuracy for SVM with Quadratic Kernel = 23.205741626794257%
Accuracy for SVM with RBF Kernel = 27.751196172248804%

MLP classifiers
Accuracy for MLP-1 (1 hidden layer with 16 nodes) = 29.1866028708134%
Accuracy for MLP-2 (2 hidden layers with 256 and 16 nodes) = 29.904306220095695%

Backward Eliminaltion
Best set of features = Length, Diameter, Whole weight, Shucked weight, Viscera Weight, Shell weight, Sex_encoded

Ensemble classifier
Accuracy of Ensemble Classifier = 27.392344497607656%

Example-2

SVM classifiers
Accuracy for SVM with Linear Kernel = 26.31578947368421%
Accuracy for SVM with Quadratic Kernel = 22.966507177033492%
Accuracy for SVM with RBF Kernel = 24.401913875598087%

MLP classifiers
Accuracy for MLP-1 (1 hidden layer with 16 nodes) = 26.674641148325357%
Accuracy for MLP-2 (2 hidden layers with 256 and 16 nodes) = 25.717703349282296%

Backward Eliminaltion
Best set of features = Length, Whole weight, Shucked weight, Viscera Weight, Shell weight

Ensemble classifier
Accuracy of Ensemble Classifier = 24.40191387559809%

Example-3

SVM classifiers
Accuracy for SVM with Linear Kernel = 25.837320574162682%
Accuracy for SVM with Quadratic Kernel = 25.598086124401913%
Accuracy for SVM with RBF Kernel = 27.27272727272727%

MLP classifiers
Accuracy for MLP-1 (1 hidden layer with 16 nodes) = 26.076555023923444%
Accuracy for MLP-2 (2 hidden layers with 256 and 16 nodes) = 26.674641148325357%

Backward Eliminaltion
Best set of features = Viscera Weight, Shell weight

Ensemble classifier
Accuracy of Ensemble Classifier = 25.837320574162682%