CS3601 MACHINE LEARNING LAB

EX.NO .1 LAB PROGRAMS

Date :05.08.24

Understand your dataset:

1. Open diabetes dataset.arff dataset

1. How many instances (examples) contained in the dataset?
2. How many attributes used to represent the instances?
3. Which attribute is the class label?
4. What is the data type (e.g., numeric, nominal, etc.) of the attributes in the dataset?
5. For each attribute and for each of its possible values, how many instances in each class have the attribute value (i.e., the class distribution of the attribute values)?

2. Visualise the Data

Take any one dataset

1. What is the ***mean***, Its ***standard deviation,*** Its ***min*** and ***max*** for the age attribute?
2. b Provide the ***summary*** of this attribute. Is this figure provided in ***Weka***?
3. Specify which attributes are numeric, which are ordinal, and which are categorical/nominal.
4. What do the red and blue colors mean in the graphic available in the right low corner? What does this graphic represent?
5. Visualize all the attributes in graphic format. Paste a screenshot.
6. Comment on what you learn from these graphics.
7. Does any pair of different attributes seem correlated?
8. Use different options in ***Weka*** for selecting attributes, with a short explanation about the corresponding method.

3. Preprocess the data

1. Remove the missing values with the method of your choice, explaining which filter you are using and why you make this choice.
2. List the methods seen in class for dealing with noisy data, and which ***Weka filters*** implement them – if available
3. List the methods seen in class for detecting outliers. How would you detect outliers with ***Weka***? Are there any outliers in this

4.Attribute Transformation

1. ***Create a new attribute*** – for example adding an attribute representing the sum of two other ones. Which ***Weka filter*** permits to do this?
2. ***Normalize*** an attribute. Which ***Weka filter*** permits to do this? Can this filter perform Min-max normalization? Z-score normalization? Decimal normalization? Provide detailed information about how to perform these in ***Weka***.

3. Save the normalized dataset into **regno.arff**, and paste here a screenshot showing at least the first 10 rows of this dataset – with all the columns.
4. How to perform sampling with ***Weka filters***? Can it perform the two main methods: ***Simple Random Sample Without Replacement***, and ***Simple Random Sample With Replacement***?

5.Supervised Algorithms

5. Run any two classifiers and observe the results shown in the "Classifier output" window.

. Choose three ratios for "Percentage split" (e.g.66 ,70 and 80% for training) test mode.

Choose three "Cross-validation" (10 folds) test mode. Run the **Id3** classifier and observe the results shown in the "Classifier output" window.

- How many instances are incorrectly classified?

- What is the MAE made by the classifier?

- Visualize the classifier errors.

- Compare these results with those observed for the **ZeroR** classifier in the cross-validation test mode. Which classifier, **ZeroR** or **Id3**, shows a better prediction performance for the current dataset and the cross-validation test mode?

- What can you infer from the information shown in the Confusion Matrix?

- Visualize the classifier errors. In the plot, how can you differentiate between the correctly and incorrectly classified instances? In the plot, how can you see the detailed information of an incorrectly classified instance?

- How can you save the learned classifier to a file?

6.Unsupervised Algorithms

Implement K-means Clustering Algorithms for iris data set.