Machine Learning Programming Task P04

Team
Alishiba D'souza
Anustup Das

08.01.2018

Abstract

The task given in this assignment was to create a Naive Bayes classifier for the Car data from the UCI Lab. Also to print the confusion matrix and give the mean error rate for the classifier.

1 Algorithm

1. Get the data.
2. Divide the data into training and testing.
3. Build a model on training data
   until all the training data
   (a) Get the attributes and classes
   (b) Calculate probability for each attribute and class
4. Use the above calculated probabilities to classify the test data using the naive bayes formula
5. Calculate the error rate over 100 samples.
6. Create confusion matrix

2 Description of the Program

The program has 5 classes.

- **NaiveBayesLauncher.java**

  This is the class with the main method in it. It takes the 'cardata.txt' file as a command line input. it passes the file to the LoadFile.java class. The partitioning parameter for train and test data i.e 2/3 is passed over here.

  The method also calls Matrix.java class for the confusion matrix calculation and at the end gives the min error rate over the 100 samples.

- **LoadFile.java**

  The class loads the file and stores into Arraylist of String.

  – **public void load (String datafile)**
    This function is used to load the file passed as the first command line argument when the program is executed.

- **Data.java**

  This class is used for partitioning the data into testing and training data using the partitioning given in the main method.

- **public void generateTrainingAndTestData()**

  This function takes the percentages value for the partitioning and randomly partitions data by generating random numbers.

- **NaiveBayes.java**

  The class implements the Naive Bayes Classifier. It has constructors and three important methods

  - **private double countDistinctClass(String value, int position)**

    This function is used to get the distinct class values in that position of the training data and returns number of distinct classes.

  - **public double countAttribteClass(String attributeValue, int attribPosition, String classValue, int classPosition)**

    This function is used to count the number attribute values belonging to that position given the class value at the class position and returns the value.

  - **private ArrayList<String[]> createClassification()**

    This function is used to make the model using the training data and then classify the test data using the training data model.

- **Matrix.java**

  The class makes and plots confusion matrix, also generates the error rate of the algorithm.

  - **private ArrayList<String> getAllDistinctClasses()**

    This function is used to get all the distinct classes from the data ArrayList as the classes will be true/false positives, true/false negative.

  - **public void confusionMatrixCreation()**

    This method is used to create the confusion matrix.

  - **confusionMatrixPrint()**

    This function is used to print the confusion matrix in the output along with the error rate. Misclassification Rate: Overall, how often is it wrong? = 1 - Accuracy Accuracy: Overall, how often is the classifier correct = (TruePossitive+TrueNegetive)/total

The Program prints the confusion matrix in following format.

Confusion Matrix	unacc	acc	vgood	good
unacc	388	16	0	0
acc	28	95	0	3
vgood	0	9	11	0
good	0	15	0	6

Table 1: Confusion Matrix.

3 Comparison with ID3 Algorithm

In the ID3 program the whole data was used as the training data and the tree was generated. The data was fully classified and the tree was specific. We didnt have any testing data in ID3 to calculate error rate.

In this program we are dividing the data into training and testing data and then calculating the error we are also printing the confusion matrix which shows the misclassification for the classes . The mean error calculated is around 11-12 percent which shows that naive bayes is a one of the best classifiers.