ELG 20225: Applied Machine Learning

Assignment 2

Due date posted in Bright Space

Submission

You must submit two documents. First, a report of the solutions including important code snippets as a PDF file. Second, the whole code should be in a separate python file (Notebooks are accepted). The file name must include your group number and assignment number, for example **Group1_HW2.pdf and Group1_HW2.py**.

Assignment must be submitted on-line with Bright Space. This is the only method by which we accept assignment submissions. We do not accept assignments sent via email, and we are not able to enter a mark if the assignment is not submitted on Bright Space! The deadline date is firm since you cannot submit an assignment passed the deadline. It is your responsibility to ensure that the assignment has been submitted properly.

Part 1:

1. Suppose we have Car data provided on Page 2 collected and the dataset contains three features. The first feature is the color, the second feature is Type, and the third feature is Origin. The target attribute is marked Stolen, which indicates whether a specific car is stolen or not. Suppose we have the following training data including 14 training samples or examples. Using Naive Bayes Classifier to classify a new instance which follows a condition ***New Instance = (Blue, SUV, Domestic)*** into **(Yes or No)**. Please include the detailed calculation process. **(20 Marks)**

2. Consider the following loss table, which contains three actions and two classes. Calculate the expected risk of three actions, and determine the rejection area of P(Class1| x). **(20 Marks)**

Target	Class 1	Class 2
a1 (Choose Class1)	0	6
a2 (Choose Class 2)	3	0
a3 (Rejection)	2	2

Example No.	Color	Type	Origin	Stolen
1	Red	Sports	Domestic	Yes
2	Red	Sports	Domestic	No
3	Red	Sports	Domestic	Yes
4	Yellow	Sports	Domestic	No
5	Yellow	Sports	Imported	Yes
6	Blue	SUV	Imported	No
7	Yellow	SUV	Imported	Yes
8	Yellow	SUV	Domestic	No
9	Red	SUV	Imported	No
10	Blue	Sports	Imported	Yes
11	Red	SUV	Domestic	No
12	Red	SUV	Domestic	No
13	Blue	Sports	Imported	No
14	Red	SUV	Imported	Yes

Part 2:

1. Use scikitlearn or other python packages to compare the performance of types of Naïve Bayesian classifiers with the help of accuracy scores. Use Spambase dataset in the question, which you can access using the URL: `https://archive.ics.uci.edu/ml/machine-learning-databases/spambase/spambase.data`

   Spam feature should be selected as output and there are 2 classes in this output **spam (1) or not (0)**, Each sample in this dataset has **58** features including spam.

   (a) Split the dataset into two parts as training data and test data. **first_80_percent** samples should be selected as training data and **last_20_percent** samples should be selected as test data. Please save these training and test datasets for remaining parts. Compute the confusion matrix and the accuracy of test data for **Gaussian** and **Multinomial Naive Bayes Classifiers** **(16 Marks)**

   (b) Use train_test_split function on input and output of the whole data and utilize **80%** of samples as train and **20%** of samples as test data. Please save these training and test datasets for remaining parts. After selecting the training and test dataset, compute the confusion matrix and the accuracy of test data for **Gaussian** and **Multinomial Naive Bayes Classifiers**. **(16 Marks)**

   (c) Use another Naive Bayes classifier of your choice to check for the improvement in terms of accuarcy_score of test data in (c) over Gaussian and Multinomial asked in (c) and provide an explanation for the improvement in performance (if any). Also, provide classification_report in terms of **precision**, **recall** and **F1-score** and display the confusion_matrix for only the selected classifier. **(12 Marks)**

   (d) Take same **first_80_percent** as asked in (b) training samples and split the data into four equal parts according to order such as the first 25% of training data (subset_1), the second 25% of training data (subset_2), the third 25% of training data (subset_3) and the fourth 25% of training data (subset_4). Train **selected**

classifier chosen in (d) for each subset and predict the accuracy_score by evaluating on **last_20_percent** of test data assumed in (b). Plot bar chart to show all subsets' accuracy on the figure. Add your comment **(16 Marks)**

Important Note

Report should include answers for all question briefly. All plots must have titles and proper axis labels. **Otherwise, you will lose one point for each missing item.** The code file is requested in case of need to verify.