ELG 5255: Applied Machine Learning

Assignment 1

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_HW1.pdf and Group1_HW1.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.

Dataset

During this assignment, Data User Modeling Dataset (DUMD) is used. Training and test splits are provided in csv file format.

Problems

1. (a) Load the DUMD dataset and convert categorical class labels under the "UNS" column to numerical values by using the LabelEncoder. **(5 Marks)**

   (b) Choose two features from DUMD dataset to apply SVM and Perceptron algorithms for classification. Plot the data by showing classes separately. Explain how and why you chose the two features? **(10 Marks)**

   (c) Classify testing data by using SVM and Perceptron classifiers. Provide accuracies, confusion matrix and decision boundaries for both classifier. **(10 Marks)**

2. (a) Build OvR-SVM (One vs Rest or One vs All) , test on DUMD testing dataset with obtained features from Problem 1. **(30 Marks)**

   For each binary classifier:

   - Obtain the binarized labels (OvR) (3 Marks)
   - Obtain the SVM's accuracy (1 Marks)
   - Plot SVM's decision boundary (2 Marks)

- Make comments on model's performance on each binary classification problem. (1.5 Marks)

Do not forget to store probability values for each classifier!

(b) Use argmax to aggregate confidence scores and obtain the final predicted labels and obtain the performance (i.e., confusion matrix, accuracy, plotting correct and wrong prediction points) of OvR-SVM. You can check MBC_Simple_Data example in lab 2 for aggregation of confidence scores. **(10 Marks)**

(c) Provide a conclusion section on your report. Include overview of what you have done and learnt during the assignment. Aim no less than one third of a page and no more than half page. **(5 Marks)**

- Models (Perceptron, SVM)
- OvR and Aggregated results

An illustrative diagram is provided in Figure 1 for Problem 2.

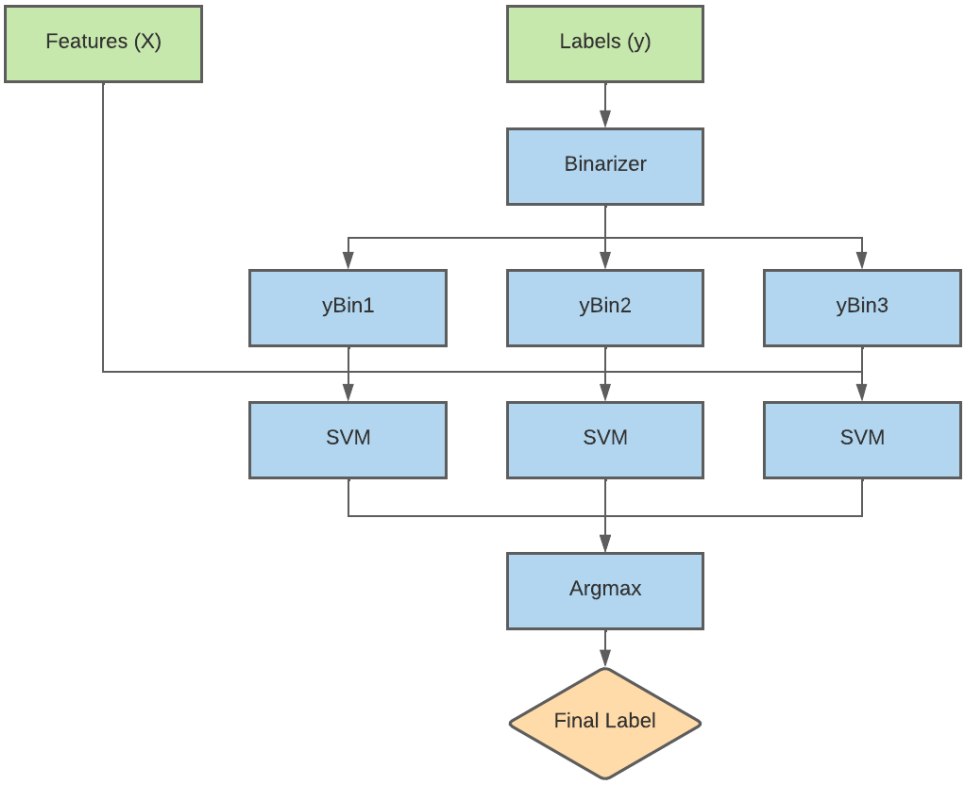

Figure 1: Converting multi-class classification to binary classification problem for OvR approach

3. Use scikit-learn or other python packages to implement a KNN classifier (redKNeighborsClassifier). In this question, we use car-evaluation-dataset, which can be downloaded from their official website or Kaggle:

(a) In this dataset, there are 1728 samples in total. Firstly, you need to shuffle the dataset and split the dataset into a training set with 1000 samples and a validation set with 300 samples and a testing set with 428 samples. Use python to implement this data preparation step. **(5 Marks)**

(b) Since some attributes are represented by string values. If we choose a distance metric like Euclidean distance, we need to transform the string values into numbers. Use python to implement this preprocessing step. **(5 Marks)**

(c) Try to use different number of training samples to show the impact of number of training samples. Use 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90% and 100% of the training set for 10 separate KNN classifiers and show their performance (accuracy score) on the validation set and testing set. You can specify a fixed K=2 value (nearest neighbor) in this question. Notably, X axis is the portion of the training set, Y axis should be the accuracy score. There should be two lines in total, one is for the validation set and another is for the testing set. **(10 Marks)**

(d) Use 100% of training samples, try to find the best K value, and show the accuracy curve on the validation set when K varies from 1 to 10. **(5 Marks)**

(e) Provide your conclusions from the experiments of question (c) and (d) in this question. **(5 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.

Similarity check will be applied for each assignment. All assignments must be original and are prepared by group members only, otherwise cheating activities in an assignment are not tolerated.