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.

Introduction

This assignment aims to explore the performance of Support Vector Machines (SVM) and Perceptron that is a single neuron consists of adder and one threshold function on a given dataset (provided in bright space in txt format).

Problems

1.

   (a) Default SVM: Implement a default SVM classifier. Train the model using the dataset and obtain confusion matrices for training and test datasets. Visualize decision surfaces for multi-calss classification, using blue, red, and orange colors to plot each class. Training data points should be marked different than test data points for visualization of decision surfaces. Keep same colors and same markers for each class for all results. **(10 Marks)**

   (b) One-vs-Rest SVM and Perceptron: Extend the analysis using the one-vs-rest strategy for SVM and Perceptron algorithms. Training and test data should be labeled for one vs rest as binary classifier. For example, class 0 vs rest means class 0 will be labeled as 1 and the rest (classes 1 and 2) will be labeled as 0. Obtain confusion matrices for both training and test datasets. Compare and analyze SVM and Perceptron results. Visualize decision surfaces using the color scheme mentioned earlier. **(36 Marks)**

(c) Aggregate results from the one-vs-rest strategy for SVM and Perceptron. Calculate the confusion matrix and visualize the decision surface for the aggregated results. Analyze performance and compare with section (a) solution. **(12 Marks)**

(d) Determine the reason why SVM performance in section (a) is different than aggregated performance of SVM in section (c). Refine the default SVM by selecting the appropriate parameter. Train the SVM model with selected parameters and evaluate its performance. Obtain confusion matrices and decision surfaces for multi-class classification. Compare results with the default SVM and discuss the impact of parameter selection. **(12 Marks)**

2. 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 magentatheir official website or magentaKaggle:

(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.