ELG 5255: Applied Machine Learning

Assignment 4

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, Pen-Based Recognition of Handwritten Digits Data Set is used. **Please use the provided Training and test set files.**

Problems

Part 1: Numerical Questions

Part 1 is not programming questions and should be solved manually! Please show the whole process. You will not receive any marks if you only show the final results.

Let's assume that TAs would go hiking every weekend, and we would make final decisions (i.e., Yes/No) according to weather, temperature, humidity, and wind. Please create a decision tree to predict our decisions based on Table 1.

1. (a) Please build a decision tree by using Gini Index (i.e., $Gini = 1 - \sum_{i=1}^{N_c}(p_i)^2$, where N_C is the number of classes). **(15 Marks)**

   (b) Please build a decision tree by using Information Gain (i.e., $IG(T,a) = Entropy(T) - Entropy(T|a)$, More information about IG). **(15 Marks)**

   (c) Please compare the advantages and disadvantages between Gini Index and Information Gain. **(5 Marks)**

Table 1:

Weather (F1)	Temperature (F2)	Humidty (F3)	Wind (F4)	Hiking (Labels)
Cloudy	Cool	Normal	Weak	No
Sunny	Hot	High	Weak	Yes
Rainy	Mild	Normal	Strong	Yes
Cloudy	Mild	High	Strong	No
Sunny	Mild	High	Strong	No
Rainy	Cool	Normal	Strong	No
Cloudy	Mild	High	Weak	Yes
Sunny	Hot	High	Strong	No
Rainy	Cool	Normal	Weak	No
Sunny	Hot	High	Strong	No

Part 2: Programming Questions

In this part, Use Pen-Digits dataset with provided splits to solve questions.

2. Apply decision tree to classify testing set, display accuracy and Confusion Matrix. **(5 Marks)**

#Bagging

Bagging is to generate a set of bootstrap datasets, create estimators for each bootstrap dataset, and finally utilize majority voting (soft or hard) to get the final decision.

3. (a) Apply bagging strategy to classify test set samples by using SVM and Decision Tree algorithm as base estimators. Display accuracy and Confusion Matrix. **(15 Marks)**

   (b) Find the best number of estimators as taking Decision Tree base estimator. Try 5 different values within the interval of [**10, 200**]. Plot accuracy vs. number of estimators. **(10 Marks)**

#Boosting

4. (a) Use GradientBoosting classifier to classify test set samples. There are 2 important hyperparameters in GradientBoosting, i.e., the number of estimators, and learning rate. First, tune number of estimators parameter by trying 4 values

in the interval of [**10, 200**]. Then by using the tuned value for number of estmators, tune the learning rate parameter by trying 4 values within the range of [**0.1, 0.9**]. Display accuracy and Confusion Matrix separately for the best value of both parameters (Number of estimators and learning rate). **(15 Marks)**

(b) Build XGBoost classifier with the same parameters that you obtained in question (4-a). Provide accuracy and Confusion Matrix. **(10 Marks)**

(c) Compare XGBoost classifier and GradientBoosting classifier performance. Which metric is the best to compare performance, accuracy or confusion matrix? Comment on Bagging and Boosting approches based on question 3 and 4. **(10 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.