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_HW4.pdf and Group1_HW4.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.

Problems

Part 1: Numerical Questions

Part 1 is not programming questions and should be solved manually with detailed explanation! 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 Children (i.e., $Gini = \sum_{i=1}^{k} \frac{n_i}{n} GINI(i)$, 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	Hot	High	Strong	No
Sunny	Mild	High	Weak	No
Rainy	Cold	Normal	Strong	Yes
Cloudy	Mild	Normal	Strong	Yes
Sunny	Mild	High	Strong	No
Rainy	Cool	Normal	Strong	No
Cloudy	Mild	High	Weak	Yes
Sunny	Hot	High	Strong	No
Rainy	Hot	Normal	Weak	Yes
Sunny	Hot	High	Strong	No

Part 2: Programming Questions

In this part, use KDD Cup 1999 dataset and download the csv file provided with this assignment.

1. (a) Load the dataset which shows 39 columns and 494021 rows. View the dataset which must show 38 input feature variables and 1 target (marked as **target** on .csv file provided) variable Obtain input feature variables as **X** and target variable as **Y**. Normalize **X** using **MinMaxScaler** from sklearn library. Compute filter-based feature selection algorithm on dataset by reducing the number of feature variables to **10** (i.e. 9 input feature variables + 1 target variable) from 39 columns and show the first five rows again and name this dataset as **my_data** comprising 10 feature variables. **(5 Marks)**

   (b) Use **sklearn** to split **my_data** using **train_test_split** into three subsets, for instance, **my_data_1 with 70% train & 30% test data**, **my_data_2 with 60%train & 40% test data**, **my_data_3 with 50%train & 50% test data** and compute the performance of Decision tree in terms of **classification_report** for each subsets.

   (10 Marks)

   (c) Visualize the **best** split of the Decision tree by considering **Entropy** as a measure of node impurity and assuming parameters **max_depth=[4, 6, 8]** for each **my_data_1 with 70% train**, **my_data_2 with 60%train** and **my_data_3 with 50%train** data as asked in (b). [**NOTE: Make sure to also consider other parameters of Decision Tree which might**

improve the performance of classification] **(15 Marks)**

(d) Compute and compare the classification performance of tuned Decision Tree in (c) for each test size **my_data_1: 30% test data**, **my_data_2: 40% test data**, **my_data_3: 50% test data** in (b) and display the **accuracy_scores**, **classification_report**, and **confusion_matrix** respectively. **(15 Marks)**

(e) Train **DecisionTree** with parameters of your choice on **my_data_1 with 70% train & 30% test data** in (b) and display the **F1 scores** for both train and test data showcasing an issue of overfitting or overlearning. In addition, apply three mitigation strategies (pre-prunning, post-prunning and k-fold cross validation) to address the problem of overfitting and display the train and test **F1 scores** showing an improvement. **(20 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.