IE343: Statistical Machine Learning

Assignment 3

Professor: Seyoung Yun

(TA : Mingyu Kim)

1 Instruction

This assignment is to implement a "Decision tree with a CART algorithm". The task is binary classification $x \in \mathbb{R}^2, y \in \{0, 1\}$. You should write a code for implementing the `gini_index` function in the `(./App/metric.py)` and run the algorithm for solving two tasks. The first task is to predict synthetic data used in the assignment 2. The second task is to predict banknote authentication data in the UCI dataset. Both tasks can be executed by running `(./decision_tree.py)`. Some of details of the source codes are as follows :

* There are "App", "Data", "Data2" and "Result" directories and `(./decision_tree.py)`. Students have to run this '.py' files to make sure that decision tree algorithm have no errors. After running this file, you can get a accuracy and its standard deviation.
* When it comes to "App" directory, there are "Pre_processing" directory and several source codes such as data_import, evaluation. The important thing is that you should complete the "gini_index" methods which can properly work `(./App/metric.py)`
* The decision tree method requires the hyper-parameters "max_depth" and "min_size". 'max_depth" is the maximum depth of the tree. "min_size" is the minimum data points to split the node. In order for the decision tree to work properly, these hyper parameters are very important. In this assignment, you should look for the appropriate hyper parameters by trial and error approach. Prior to running `./decision_tree.py`, you must manually determine these values.
* To help your implementation, there is a boolean variable to choose a dataset between synthetic data and banknote data by "synthetic" in (`./decision_tree.py`)

In this assignment, we must describe the model that best fits the entire data set. For evaluation, we provide the accuracy metric. Like Assignment 2, the average accuracy and its standard errors for the entire data set are also provided. Furthermore, you need to visualize the training and testing data with the decision boundary. `(./Results/result_(i).png)` Therefore, you must answer the questions in English and submit a report as a PDF file with your completed source code. (No limit of pages) The report should includes theoretical background for decision tree and your opinion about this model. (Please write a report concisely). In addition, you have to complete this source codes in the attached python files.

2 Data

As we mentioned before, there are two datasets. First, the synthetic data is generated by a specific data generation process, but it is not released for students `.App/Data`. Second, it comes from a banknote task in UCI datasets `.App/Data_2`. In both problems, there are 10 data folds. In synthetic data, the first 5 data folds consists of 50 training points and 20 testing data points. it was generated without outliers, but last 5 data folds consists of 60 training points and 25 testing points, which have outliers. In this assignment, students can study how sensitive the model is to outliers. The classifiers should be trained by only one of data sets. After training your model, you should compare the binary 'y' values to fitted values in terms of "accuracy"

Synthetic Data

* The first 5 data sets have 50 training data and 20 test data, but the last 5 data sets have 60 training data and 25 test data with outliers.
* This data comes from a specific data generation process, but it is not released.

* You shoud the use same numbered data sets while training and testing.

Banknote Data

* The all data sets have 960 training data points and 412 test data points.
* You shoud the use same numbered data sets while training and testing.

3 Implementation

This code should be written using the 'Numpy' package. Other packages are not allowed. (Scipy, Tensorflow, pytorch and etc.) You must write down appropriate methods ('gini_index' method) in `(./App/metric.py)`. Except for the `"gini_index"` methods, you don't need to modify any source code. Furthermore, you should submit all source codes to run your codes.

4 Question

A. Complete the `"gini_index"` method in `(./App/metric.py)`. You should take a picture of your source code `"gini_index"` and attach it.

B. Write down all metrics used in the decision tree algorithm. You also need to describe mathematical equations of these metrics.

C. After applying decision tree, write the average accuracy and its standard error and make a plot with train, test data and the decision boundary in any data fold in the synthetic data. Remember that you need to attach two kinds of results and plots, such as, with outliers as well as without outliers. You need to check how the decision tree works when you train and test the model using data set $5 \sim 9$. Compare the result from data set $0 \sim 4$ with the result from data set $5 \sim 9$. Also explain your opinion about the decision tree method in terms of sensitivity.

D. Compare the results from logistic regression and LDA with the results from the decision tree using the synthetic data set. In particular, you should attach plots from each method and explain what shape of the decision boundary has.

G. Elaborate the accuracy and its deviation in the banknote data set. You should describe why decision tree can work well in this data set.