In the name of God

Machine Learning Course (Spring 2021)

Assignment #2: Classification and Regression Decision Tree Algorithms

Due date: 25th of Farvardin

In this assignment, you need to implement the ID3 and regression decision tree algorithms.

Part A

You got familiar with ID3 decision tree algorithm in class. Implement a basic algorithm for learning the ID3. (The ID3 algorithm is mentioned in table 3.1 of the Tom Mitchel's book.)

- Implement your own code from scratch. Use information gain to find best attribute at each node. Use 10-fold cross validation for evaluation, and report the average accuracy and standard deviation (STD) of the 10-folds. For example, 5.89 ± 0.02 is the results for X data set, in which 5.89 is the average accuracy over 10-folds and 0.02 is the STD of these ten accuracy values.
- You need to use the data set in 'Classification' folder. This dataset includes descriptions of hypothetical samples corresponding to 23 species of gilled mushrooms. Each species is identified as definitely edible or definitely poisonous. There are missing values for some attributes that you should handle them in your implementations. In order to do this, you should substitute the missing values by the mode (most common value of an attribute) of the corresponding attribute of the samples of the same class. In categorical (Nominal) attributes, one way is to substitute the missing values with the mode. Mode should be computed form the samples which have the same label with the considered sample's label.

Part B

Please read about the regression decision tree algorithm in the enclosed file, which explains the algorithm with a simple example.

The algorithm is just like the ID3 algorithm, except for some minor changes. For example, you need to calculate the *Standard Deviation (S)* instead of information gain; or use the average of the values in a leaf node rather than finding the majority of the labels.

You need to implement the tree from scratch and calculate the MSE of the predicted target values for the data sets enclosed in the 'Regression' folder.

In the name of God

Machine Learning Course (Spring 2021)

Assignment #2: Classification and Regression Decision Tree Algorithms

- To get familiar with the implementation of the algorithm, read the enclosed file carefully and completely follow its example.
- Implement the regression decision tree in general. Make sure to use the formulas in the attached file in order to compute different values. You can use the termination threshold of 10% or three samples according to the results.
- Use the dataset 'EnjoySport' in order to test your code. You have to report the MSE (Mean Square Error) of the train set for this data set. In other words, train your decision tree with the given data set and then use the trained tree to predict the values of the target variable for all instances and calculate the MSE.
- Once your code is executed properly, use the real-world data set 'Automobile' to evaluate its performance. Split your data set into two disjoint sets, namely train and test sets, with 70% and 30% ratios. This splitting process should be random, so you need to run your code for ten individual runs and report the average MSE and standard deviation (STD). For example, 5.89 ± 0.02 is the result for X data set, in which 5.89 is the average MSE over ten individual runs and 0.02 is the STD of these ten MSE values.

Important Notes:

- You need to implement the algorithms from scratch. Using the built-in functions and algorithms is not allowed.
- Feel free to use your preferred programming languages.
- Pay extra attention to the due date. It will not be extended.
- Be advised that submissions after the deadline would not grade.
- Provide a report for your assignment and explain your features, code, and results in part B.
- The name of the uploading file should be your **Lastname_Firstname.**
- Using other students' codes or the codes available on the internet will lead to zero grades.

Reference:

www.saedsayad.com