ENGINEERING DIVISION | NYU ABU DHABI

Applied Machine Learning

Mini Project 3 - Decision Tree

Due Date: Please Refer to the Brightspace

Introduction

Decision Tree

In computer science, decision tree learning uses a decision tree (as a predictive model) to go from observations about an item (represented in the branches) to conclusions about the item's target value (represented in the leaves). It is one of the predictive modeling approaches used in statistics, data mining and machine learning. Tree models where the target variable can take a discrete set of values are called classification trees; in these tree structures, leaves represent class labels and branches represent conjunctions of features that lead to those class labels. Decision trees where the target variable can take continuous values (typically real numbers) are called regression trees. In decision analysis, a decision tree can be used to visually and explicitly represent decisions and decision making.

Dataset

Classification Trees with Numerical Features (two data sets)

- **Iris**: has three classes and the task is to accurately predict one of the three subtypes of the Iris flower given four different physical features. These features include the length and width of the sepals and the petals. There are a total of 150 instances with each class having 50 instances.
- **Spambase**: is a binary classification task and the objective is to classify email messages as being spam or not. There are about 4600 instances.

Since both datasets have continuous features you will implement decision trees that have binary splits. For determining the optimal threshold for splitting you will need to search over all possible thresholds for a given feature (refer to class notes and discussion for an efficient search strategy). Use information gain to measure node impurity in your implementation.

Requirements

Decision Tree

1. Growing Decision Trees

Instead of growing full trees, you will use an early stopping strategy. To this end, we will impose a limit on the minimum number of instances at a leaf node, let this threshold be denoted as n_{min}, where n_{min} is described as a percentage relative to the size of the training dataset. For example, if the size of the training dataset is 150 and n_{min}= 5, then a node will only be split further if it has more than eight instances.

- For the Iris dataset use n_{min} E {5, 10, 15, 20}, and calculate the accuracy using 10 fold cross-validation for each value of min.
- For the Spambase dataset use n_{min} E {5, 10, 15, 20, 25}, and calculate the accuracy using 10 fold cross-validation for each value of n_{min}.

You can summarize your results in two separate tables, one for each dataset (report the average accuracy and standard deviation across the folds).

Deliverables

A .ipynb file containing the following:

1. source code
2. detailed description of the project
3. answers to the programming questions.

Before submitting your project, please make sure to test your program on the given dataset.

Notes

You may discuss the general concepts in this project with other students, but you must implement the program on your own. ***No sharing of code or report is allowed.*** *Violation of this policy can result in a grade penalty.*

Late submission is acceptable with the following penalty policy:
10 points deduction for every day after the deadline