UECS3213 / UECS3453 Data Mining

SESSION: 2019/1

Lab 7: Decision Tree Implementation Using scikit-learn

Introduction

A decision tree is one of most frequently and widely used supervised machine learning algorithms that can perform both regression and classification tasks. Decision trees also provide the foundation for more advanced ensemble methods such as bagging, random forests and gradient boosting.

In this tutorial, you will discover how to implement the Classification and Regression Decision Tree algorithm with Python. Remember from the previous practical that the steps involved in any classification tasks involve:

- Loading the dataset.
- Exploring the dataset.
- Visualizing the dataset.
- Preprocessing the dataset.
- Constructing models and making some predictions.
- Evaluating the algorithm (see in https://stackabuse.com/decision-trees-in-python-with-scikit-learn)

Objectives

At the end of this lab, you are expected to acquire the following:

a) How to calculate and evaluate candidate split points in a data.
b) How to arrange splits into a decision tree structure.
c) How to apply the classification and regression tree algorithm to a real problem.

Instruction

1. Go through Section 1.10.1 and 1.10.2 of https://scikit-learn.org/stable/modules/tree.html. Read the other sections of the tutorial on theory to understand as well.

2. Also, go through the "Decision Tree Classification in Python" at the following link: https://www.datacamp.com/community/tutorials/decision-tree-classification-python

3. Follow the step-by-step instructions in the tutorial.

4. Finally, try out the example in https://stackabuse.com/decision-trees-in-python-with-scikit-learn/

The End