

CS3481 Fundamentals of Data Science

Assignment 1

Objective

In this assignment, you will implement a predictive modeling approach based on the decision tree using Python.

Detailed requirement

We have introduced a predictive modeling approach based on the decision tree in the lecture. In this assignment, you will implement and evaluate this approach on the *Forest type mapping* data set, which you can access from the UCI Machine Learning Repository: archive.ics.uci.edu/ml/datasets/Forest+type+mapping.

The objective of this classification task is to determine the specific forest type at a particular geographic location, based on a set of 27 input attributes. There are four classes of forest types: Sugi ('s'), Hinoki ('h'), mixed deciduous ('d'), and non-forest land ('o'). The first nine input attributes correspond to remote sensing measurements from multiple spectral bands, and the remaining attributes correspond to interpolated spectral information. The data set is partitioned into a training set and test set.

You can implement a decision tree model using the Python package `scikit-learn`, and visualize the model using the package `graphviz`.

Assignment Submission

You should submit a report to summarize your work. For the above data set, the following tasks are to be performed:

- (a) Construct multiple decision trees based on the default training set/test set partition using different parameter settings. Compare the structures and performance of these decision trees. (25%)
- (b) Exchange the training and test set and repeat the tasks in (a). (25%)
- (c) For selected trees, observe the classification performance associated with the different classes, and determine which pair(s) of classes are likely to be confused with each other. (25%)
- (d) For selected confused class pairs in (c), identify the corresponding leaf node(s) and analyze the sequence of decisions that lead to the misclassification. (25%)

Please provide a detailed description of the results of the above tasks in your report.