Fundamentals of Data Science

Assignment 1

Objective

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

Detailed Requirement

We have introduced a predictive modeling approach based on the decision tree in the class. In this assignment, you will implement and evaluate this approach on the ***Vertebral Column*** dataset from the UCI Machine Learning Repository: **http://archive.ics.uci.edu/ml**.

You should partition the dataset into two subsets: one for training and the other for evaluation. The partitioning should be performed in such a way that the proportions of data records belonging to the different classes in the training set and test set should be similar to those of the original dataset.

Please note that there are two versions of the ***Vertebral Column*** dataset. Please use the version in which the orthopedic patients are categorized into three classes (disk hernia (DH), spondylolisthesis (SL) or normal (NO)).

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

You may refer to the following references for more details about Python and its packages.

- Data mining tutorials using Python (**http://www.cse.msu.edu/~ptan/dmbook/software**)
- Scikit-learn website (**https://scikit-learn.org/**)

Assignment Submission

You should submit a report to summarize your work. The following tasks are to be performed:

a. Construct multiple decision trees based on different partitions of the dataset into a training set and a test set. You should clearly specify which impurity measure you have used for tree construction, and the parameters you have selected. (25%)

b. Compare the structures and classification performances of these different trees. (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.