DLI Accelerated Data Science Teaching Kit

Module 14 Lab:
Decision Tree Classification Clustering

OBJECTIVE

As an unsupervised learning algorithm, Decision Tree is used to build a tree structure model for resolving classification and regression problems. The goal of this lab is to implement the classification with Decision Tree and evaluate the classification performance with various evaluation methods.

PREREQUISITES

Install Python packages below.

- **numpy** is the fundamental package for scientific computing with Python.
- **pandas** is a fast, powerful, flexible and easy to use open source data analysis and manipulation tool via Python.
- **sklearn** is contains a lot of efficient tools for machine learning and statistical modeling including classification, regression, clustering and dimensionality reduction.

INSTRUCTIONS

- Downloading *Balance Scale Data Set* employed for validating the model, which can be downloaded UCI Machine Learning Repository (https://archive.ics.uci.edu/ml/machine-learning-databases/balance-scale/balance-scale.data).
- Splitting the *Balance Scale Data Set* into training and testing data
- Building function for decision tree classifier based on information gain and training the classifier on training data
- Implementing classification on testing data
- Evaluating the prediction results with evaluation methods including confusion matrix (https://scikit-learn.org/stable/modules/generated/sklearn.metrics.confusion_matrix.html) and accuracy (https://scikit-learn.org/stable/modules/generated/sklearn.metrics.accuracy_score.html).