DLI Accelerated Data Science Teaching Kit

Module 19 Lab:
Cancer Recognition on Genomics Data via Decision Tree Algorithm

OBJECTIVE

Microarray technology has achieved significant developments in the field of molecular biology. It allows us to monitor the expression of hundreds of genes at the same time just in one hybridization test. Thus, it is possible to analyze the pattern and gene expression level of different types of cells or tissues, where it has an important application in medicinal and clinical research. For example, microarray data plays an important role in identification and classification of the cancer tissues. The goal of this project is to implement cancer classification on microarray data via decision tree algorithm.

PREREQUISITES

Install Python packages below.

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

INSTRUCTIONS

- Download Cancer Data including actual.csv, data_set_ALL_AML_train.csv and data_set_ALL_AML_independent.csv here (https://www.kaggle.com/crawford/gene-expression)
- Remove nonnumerical columns
- Prepare labels
    - "ALL (acute lymphoblastic leukemia)" for 0 class (non-cancer) AND "AML (acute myeloid leukemia)" for 1 class(cancer)
    - First 38 rows for training and the rest of rows for testing
- Data Preprocessing: Replacing INF values with mean values
- Train a cancer classifier with Decision Tree algorithm and evaluate the testing results with **Accuracy**
- Plot the Tree (the cancer classifier)