Activity Unsupervised Learning

Subject: Machine Learning - MSc. Computational Biology

Introduction

In order to illustrate clustering with a real dataset, we will analyze the NCI-60 Human Tumor Cell Lines Screen provided by the National Cancer Institute[1]. It utilizes 60 different human tumor cell lines to identify and characterize novel compounds with growth inhibition or killing of tumor cell lines. The original dataset has been modified to minimise its noise. Genes with variability above a threshold have been filtered out to generate normalised RNA abundance data.

Description

NCI60 is a dataset of gene expression profiles of 60 National Cancer Institute (NCI) cell lines. These 60 human tumour cell lines are derived from patients with leukaemia, melanoma, along with, lung, colon, central nervous system, ovarian, renal, breast and prostate cancers. This panel of cell lines have been subjected to several different DNA microarray studies using both Affymetrix and spotted cDNA array technology.

Objectives

Cluster the data using hierarchical and partitional unsupervised learning models to analyze how similar these cancers are based on gene expression. Is their similarity related to tissue of origin?

To-DO

1. **Prepare Data**
    a. Load data from *nci_var_filtered.csv* into a Dataframe (df)
    b. Change the index from the numerical index (default when you load a pandas df) to the column 'gene'
    ```
    df = df.set_index('gene')
    ```
    c. Get a list of all genes by retrieving the column names
    ```
    genes = list(df.columns.values)
    ```

[1] https://dtp.cancer.gov/discovery_development/nci-60

2. **Hierarchical Clustering**
    a. Create several **dendrograms** by varying the *method* used (e.g. single linkage, or ward's method) and the *distance* metric (e.g. 'euclidean')
    b. Analyse the results to answer the following questions:
        i. What does each group mean?
        ii. What does each level mean?
        iii. How does the method or the distance influence?
3. **Partitional Clustering**
    a. Performs some **K-means** clustering by varying the number of clusters (i.e. k)
    b. Build an **Elbow plot** to gauge the number of clusters (k)
        i. What is the most appropriate number of groups?
        ii. Why?
    c. Calculate the **Silhouette score** to measure cluster compactness and cluster separation
        i. What is the highest Silhouette score?
        ii. Considering the scores, what can you conclude?
    d. Analyse the **results** to answer the following questions:
        i. How many clusters should we have?
        ii. Does cluster assignment match tissue of origin?

To Submit

- The Jupyter Notebook source code of the study with explanations for the different steps.
- A pdf file with the study that you have done.

Evaluation Rubric

The following criteria will be used for evaluating the activity:

- The quality of the explanations and source code of the different steps taken to make the study.
- The inclusion of plots and interpretations that help to understand the presented results.

Deadline

- The deadline to submit this activity is: end of the day (23:59pm), December the 22th