University of Stirling
Computing Science and Mathematics

CSCU9YE - Artificial Intelligence

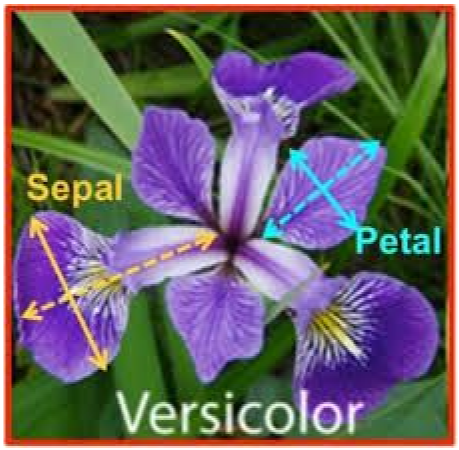
Lab 5: Unsupervised Machine Learning

For machine learning we will be using the Python library Scikit-learn, which implements a wide range of algorithms, and has wonderful documentation, examples and tutorials.

Clustering is an unsupervised learning method that allows us to group a set of objects based on similar characteristics. In general, it can help you find meaningful structure among your data, group similar data together and discover underlying patterns.

The dataset we will be using in the lab is provided by the Scikit-learn library. It is a well-known dataset named IRIS flowers. The data set consists of 50 samples from each of three species of Iris flowers: setosa, virginica and versicolor. For each flower, 4 features are stored: the length and the width of the sepals and petals, in centimeters. The dataset contains both the features and the target (species of flowers). For clustering, however, we will only consider the features (and not the targets or labels) as the idea of clustering is to work with unlabelled data.

Image of an Iris flower, showing their petal and sepal width and length.



Activities

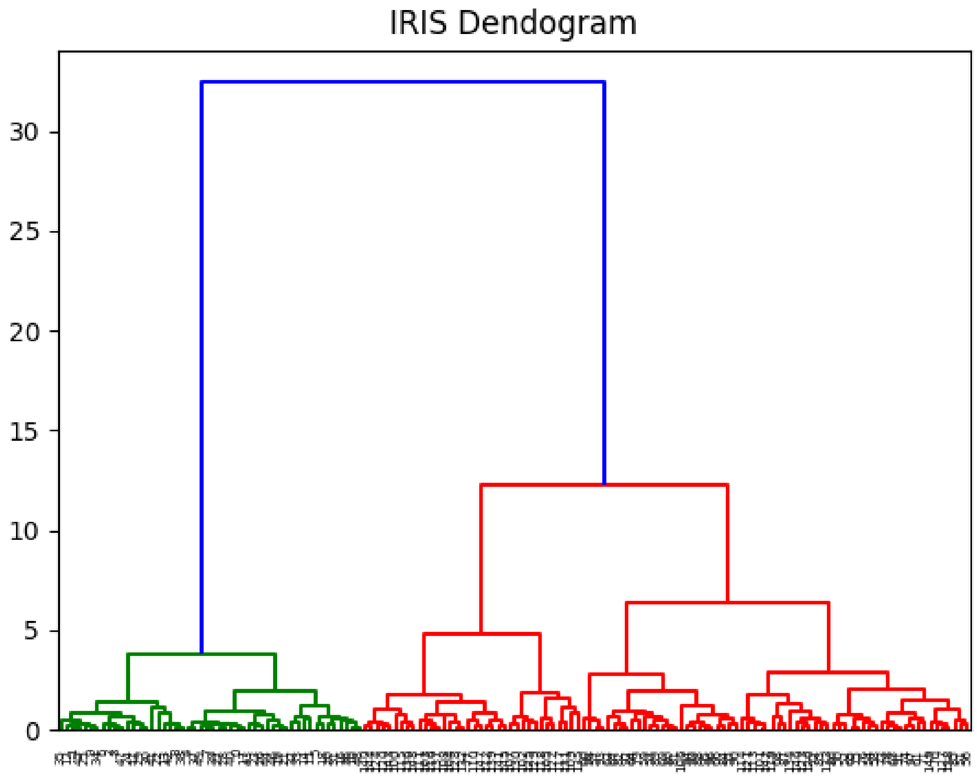
Your tasks in this lab are the following:

- Browse the Scikit-learn documentation at: http://scikit-learn.org/stable/documentation.html
- Browse the structure and content of the IRIS dataset at: https://gist.github.com/netj/8836201
- Run the python script provided in Canvas/Units/Lab sheets: lab5_unsupervised.py. In this script the K-Means algorithm is used model and cluster the Iris dataset.
- Analyse and understand each line of code and the outputs/printouts of running the script

- Three scatter plots are created by running the script, where all of them visualises the first two attributes (columns 0 an 1).

    1. IRIS - No-Labels: assume that we do not have the labels, so all the dots will appear in the same color.
    2. IRIS - Real Labels: Uses the real labels from the IRIS dataset
    3. IRIS - KM Cluster Labels: Uses as labels those produced by the K-Means clustering algorithm

- How the scatter plots compare? Does the K-Means clustering reproduces the real labels?
- Create another set of three plots that visualise another pair of attributes, for example columns 1 and 3.
- Using the Scikit-learn documentation for AgglomerativeClustering implement an agglomerative algorithm to cluster the IRIS dataset.
- Produce a scatter plot with title "IRIS - AC Cluster Labels" where the colors (c parameter) is now given by the clustering produced by agglomerative clustering
- How this new scatter plot compares to both the scatter plot with real labels and the scatter plot with KM cluster labels?
- Using the documentation for Scientific Python library 'scipy.cluster.hierarchy' visualise the Dendongram of the IRIS Dataset, you should see an image similar to this:


IRIS Dendogram