Own Understanding of Scikit-learn and Analysis of the Code Block

Scikit-learn, also known as sklearn, is a machine learning algorithm package for python. And it is very useful for processing data. It is mainly used for classification, clustering, regression and dimensionality reduction. It also contains a module that includes utilities to load data sets, including methods to load and fetch popular reference data sets. In the DataSeparation.py file, which contains the code block for separating the original data into training set and testing set, I used the train\_test\_split function to do the task.

The following part is the step-by step analysis of DataSeparation.py.

At the very beginning, we import the necessary libraries - numpy and sklern. Numpy is used for reading and writing csv files which contains the data as the data sets. Then we define the location of our data sets. The datapath is the original data set path. Trainpath is for the generated training set path, while testpath is for the generated testing set path. Next, we load data into a matrix use the function numpy.loadtxt. Since the data stored in csv file uses comma as delimiter, and the first line is the label, so we skip it. Then we generate the two sets using train\_test\_split form sklearn. X and y is our matrix. Test\_size defines the size of the testing set. And random\_state defines how the original data is separated. Different random\_state will generate different result. Finally, we write the two generated two sets back to two files, and use their original format.