Practical 5 Exercise:

Part I: Supervised Learning

1. Please classify iris dataset into respective classes (Iris dataset is given as iris.txt file). Compare the results between K-Nearest Neighbour (KNN) and Linear Support Vector Machine (SVM) classifier.

   Iris dataset:
   - Number of Instances: 150 (50 for each classes)
   - Number of Attributes: 4 numeric, predictive attributes and the class
   - Attribute Information:
     1. sepal length in cm
     2. sepal width in cm
     3. petal length in cm
     4. petal width in cm
     5. class: Iris-Setosa, Iris-Versicolour, Iris-Virginica
   - Data in the txt file is presented as "sepal length, sepal width, petal length, petal width, class". Example: "5.1,3.5,1.4,0.2,Iris-setosa"

2. Use "***score***" function to evaluate the performance. (Try different portion of training data and repeat the evaluation, explain the difference.)
3. Perform prediction using "***predict***" function.

Hints:

(a) Read the iris dataset from txt file.

(b) Preprocess the iris dataset into the suitable representation for classification (2D Array).
   - To initialize a 2D array: dataset = np.zeros((row, column))
   - The output should be converted to numeric representation (e.g. Iris-Setosa = 0, Iris-Versicolour = 1, and Iris-Virginica = 2)

(c) Separate the dataset into training and testing data.

(d) Refer to the practical to initialize KNN and SVM classifier.

Part II: Unsupervised Learning

4. Comparison of different clustering algorithms:
   a. Open "plot_cluster_comparison.py" in Python IDE, run it and understand the code and output.
   b. Include the code to perform Kmeans clustering into the same Python file.
   c. Visualize the Kmeans output as first result and following by the others.
5. Perform k-means clustering on iris dataset. You may reuse the code from practical 4(i) for data preprocessing.
   a. Cluster the iris data with 2 attributes at each time, plot and compare the distribution of data between original label and cluster label for:

1)First and second attribute (sepal length & sepal width)
2)Second and third attribute (sepal width & petal length)
3)Third and fourth attribute (petal length & petal width)

b. Explain why it is very difficult to evaluate the score.

References:

Scikit-learn supervised learning documentation is available at http://scikit-learn.org/stable/supervised_learning.html#supervised-learning

Scikit-learn unsupervised learning documentation is available at http://scikit-learn.org/stable/unsupervised_learning.html