Lab Class ML:XII

(This exercise sheet is for Digital Engineering students only)

By 2019-01-30 11:59 am, solutions for the following exercises have to be submitted: 1, 2, 3.

Exercise 1 : Running Time

Which of the following statement are true?

- ☐ The running time of single link and complete link are of the same order.
- ☐ The dimensionality of the data affects asymptotically the running time of link-based algorithms.
- ☐ The dendrogram shows the order or building the cluster in the hierarchical algorithms

Exercise 2 : Density-based clustering

Which of the following statements are true?

- ☐ DBSCAN is unsupervised, even though you need to manually define some parameters like the value of neighborhood radius.
- ☐ In DBSCAN, once a point is labled as noise, the algorithm will not include it in any clusters later.
- ☐ Both MajorClust and DBSCAN work on small subsets of nodes.
- ☐ MajorClust is more effective than DBSCAN in high-dimensional data.

Exercise 3 : [P] Density-based clustering

Your task is to use DBSCAN to cluster the instances in the Fisher's Iris dataset and evaluate its effectiveness. DBSCAN is a density-based algorithm that requires the configuration of hyper-parameters. Sklearn offers an implementation of the algorithm which we recommend to use.

(a) Download the dataset and write a Python program that reads the data. Alternatively, you can read the dataset in Scikit-learn using the command:

```
import sklearn
iris = sklearn.datasets.load_iris()
```

, and split the dataset into instances and their corresponding labels.

(b) Use DBSCAN (in sklearn.cluster.DBSCAN) to cluster the instances with different epsilon values ranging from 0 to 1 with a step of 0.01.

(c) Use the evaluation measure `adjusted_rand_index` to evaluate the produced clusters based on the ground-truth labels for each epsilon value. Which epsilon value is the best? how many clusters get produced?

(d) Use the `matplotlib` library, and draw a scatter plot that shows the *petal length* attribute on the x-axis and *sepal width* on the y-axis. Use different colors for the three different species based on the ground-truth label.

(e) Draw a similar scatter plot with the same dimensions but color the instances according to their predicted clusters.