# COMP 4331 Assignment 3

## Possible Learning Objectives:

1. Learn how to convert the pseudocode in the tutorial into functional codes.
   1. You can get the pseudocode in the tutorial note 6
   2. You can also refer to the implementation found in the resource to understand the details of the pseudocode

## 2 DBSCAN via Python (30 points)

Given the dataset (https://github.com/ZengqiangYan/COMP4331/tree/master/ Assignment3-Dataset/DBSCAN-Points.mat), implement the DBSCAN algorithm for clustering.

### 2.1 Dataset Description

The dataset contains 500 2D points totally.

### 2.2 DBSCAN Implementation

You are required to implement the DBSCAN clustering algorithm:

* You are not allowed to use any existing DBSCAN method.
* Run your implemented DBSCAN on the dataset by setting epsilon = 0.12 and MinPts = 3.
* Use the euclidean distance as measurement.
* Draw the clustering results and compare your results with the corresponding results generated by the DBSCAN model in scikit-learn library.
* Adjust the parameters 3-5 times, draw the corresponding results and analyze the inﬂuence of the parameters.

### Resources:

1. This is the implementation of DBSCAN by a user. It is very readable and its implementation is pretty similar to the pseudocode given in the tutorial 6 note:

<http://mccormickml.com/2016/11/08/dbscan-clustering/>

1. To learn the way to plot the sklearn cluster, you can refer to this link:

<http://scikit-learn.org/stable/auto_examples/cluster/plot_dbscan.html#sphx-glr-auto-examples-cluster-plot-dbscan-py>

1. You can use this to double check your output with the output produced by sklearn algorithm using this webpage:
2. <https://www.diffchecker.com/diff>

## Possible Objective:

1. Learn to use Numpy:
   1. Learn to implement mathematical equations using Numpy library
   2. Learn to put your linear algebra knowledge into practice.
   3. You can find the pseudo-code in the tutorial 7

## 3 EM-GMM via Python (20 points)

Given the training data (https://github.com/ZengqiangYan/COMP4331/tree/ master/Assignment3-Dataset/GMM-Points.mat), implement the GMM algorithm for clustering.

### 3.1 Dataset Description

The dataset contains 400 2D points totally with 2 clusters. Each point is in the format of [X-coordinate, Y-coordinate, label].

### 3.2 EM-GMM Implementation

You are required to implement the GMM clustering method by using the EM algorithm (reference to slides No. 29-41):

* You are not allowed to use any existing EM-GMM method.
* Run your implemented GMM on the dataset.
* In your report, draw the clustering results of your implemented algorithm and compare with the original labels in the dataset.

Hint: For simpliﬁcation, during the M step, you can directly calculate the mean and the std of points assigned to each cluster for updating.

### Resource:

1. Multivariate Gaussian Mixture Models
2. <https://www.youtube.com/watch?v=qMTuMa86NzU&t=351s>

### Note

One exemplar way to draw the clustering results on the DBSCAN dataset is shown as below.

Figure 1: One exemplar way to draw the clustering results where points are assigned with colors according to the corresponding clusters.