Machine Learning: Programming Exercise 3: Clustering

1 Explore the sklearn.datasets package for generating random data sets. In particular, use make_blobs, make_cicles and make_moons with different parameter values to generate 10 different datasets. These data sets will be used to evaluate the various clustering algorithms to be implmented within this exercise.

2 Implement the following algorithms using numpy. You should attempt to vectorized the code to whatever extent possible.

   - K-means: This should include code for ensuring no empty clusters are produced. Also implement atleast three methods for initialising clusters.
   - Expectation Maximization: Implement code to obtain a train/test split (without for loops) of the data and use maximum linkelihood estimate to choose the optimal number of clusters.
   - Agglomerative Clustering: Use reduction in dispersion to propose a good cut point.
   - CURE: Provide a parameter to choose the number of representative objects
   - BIRCH: The algorithm should be able to take different values of the radius threshold, B and L parameters
   - DBScan: Parameters should be Eps and NumPoints

3 Now implement the following methods for choosing the number of clusters:

   - Elbow Curve using Dispersion
   - Sillhoutte Score
   - Gap Statistic: Use the uniform distribution, random number generator in python for this exercise

Please note that you should extend the object oriented code base from Exercise 2 with this code and reuse as much code as possible between these algorithms.