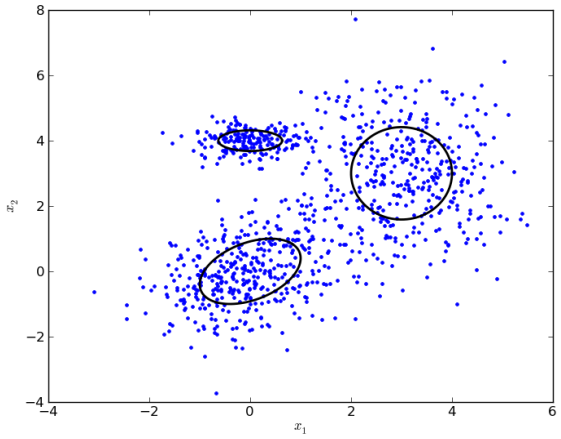
Department of Computer Science and Engineering
CSE-454: Data Warehousing and Data Mining Sessional
Assignment-5 (Clustering Analysis)

Description of the Assignment

Let's assume you have been hired by Bangladesh Navy as a data analyst for their newly acquired Ming-class submarines. They are planning to use these submarines to secure national maritime border. One of their potential targets is pirate ship. They use active acoustic sonar to find floating pirate ships. By transmitting a sound pulse and receiving the echo on an array, the sonars can determine the direction of the echoes that return from objects hit by the sound. They can also measure the time it takes for echoes to return and calculate the distance to the object causing the echo. Unfortunately, if there are several pirate ships nearby, the signals cause interference. Using periscope visuals, it is possible to determine the number of ships but exact location of each ship is difficult to locate. Without precise location, a torpedo hit is not guaranteed. The 2D plot of a sample sonar data is given in the figure below. There are three pirate ships in surface water, one inside each black circle. Each blue dot represents a single location measurement of one of the three ships. You have been instructed to build a system which takes the sonar data as well as the number of pirate ships as input and provide the estimated location of each ship. *The centroid of each cluster is the estimated location of a pirate ship.*



Since you are a new recruit, you are not yet well versed in applying machine learning model to real world problem. So, you plan to consult Dr. Google who is an expert of virtually every field. He informs you that you should use *K-means Clustering Model* in the above-mentioned case. One possible implementation for this model is an EM (expectation-maximization) algorithm. He is also generous enough to give you a pseudocode for this which is given below.

Algorithm-1 *k-means* algorithm

1: Specify the number *k* of clusters to assign.
2: Randomly initialize *k* centroids.
3: **repeat**
4: **expectation:** Assign each point to its closest centroid.
5: **maximization:** Compute the new centroid (mean) of each cluster.
6: **until** the centroid positions do not change.

Now you are excited to implement the above pseudocode and estimate the location of enemy ships. However, since this is an issue of national security you want to first generate data by your own and then train the model with the actual data. As the model will be used in a small RASPBERRY PI device where you cannot use heavy libraries, so you have to implement the algorithm with basic python programming.

Special Instructions

1. Although you are allowed to consult Dr. Google, don't copy anything! If you do copy from internet or from any other person or from any other source, your grade will be badly hampered and it is obvious. More than that, we expect fairness and honesty from you. Don't disappoint us!
2. Generate random data (which will look like the figure above) and use these data in your model. You may use Gaussian distribution (with 3 different mean and standard deviation) to create those data.
3. Implement the above pseudocode and estimate the location of enemy ships.
4. Use Jupyter notebook to create your project. Rename the notebook file with your student ID. File name must be your student ID (<std id>.ipynb). Do not add your name or your section in the file name.
5. Submit only one Jupyter notebook (<std id>.ipynb). file. Do not compress it nor include any other file with your submission.
6. Marks distribution will be as follows:

Ser	Description	Marks
1	Generating dataset	5
2	Implementing learning algorithm	10
3	Printing estimated location and comparing with actual location	5
Total		**20**

7. Deadline for submission is **Tuesday 27 October, 2020 11:55pm**
8. Contact your course teachers for any typos or discrepancies in this document.