EE6435 Homework 5
Comparing the clustering performance of k-means and EM clustering algorithm

Out: April 16th, 2019
Due: **11:59PM, April 28 (Sunday night).** Note that the turn in function will **close at 11:59PM sharp**. **No late work will be accepted** unless there is documented crisis such as sickness. Total points: 120.

This is a group project. Each group can have at most two students. You can also choose to work individually. In that case, you get 15 bonus points automatically. For example, if you work on this by yourself and get 100 points, your total points will be 115 for this homework. This will also be counted towards your total homework points.

If you have a partner, clearly describe the contributions of each of you in the readme file. In addition, each of you must work on either k-means or EM. It is not allowed to have one student do all the work.

Copying others' codes is not acceptable. All your codes will be checked using MOSS, a tool designed for detecting similarities between programs. We will run it against codes online and also codes between submissions. If high similarity is detected, both parties get zero.

Problem description:
In this homework, you will apply two different clustering algorithms to cluster data points that are randomly sampled from two normal distributions. In addition, you need to generate the input data yourself. The detailed process can be found below.

1. (15 pts) Generate random samples from two normal distributions. You can use Excel, Python, or Matlab for this. For example, provide the number of points (e.g. 100), the mean, and the standard deviation, let the function output randomly sampled data points. For the data points sampled from two different normal distributions, mix them and use all as input to the following clustering algorithms.
2. (65 pts) Implement k-means and EM on Gaussian-mixture model.
   2.1 Implement k-means. In the codes, clearly comment the following components: initialization, distance

computation, assignment of a data point to a cluster, and update of the centroid.

2.2 Implement the EM algorithm. In the codes, clearly comment the following components: initialization, computing the posterior probability (E step), updating the parameters of the normal distributions (M step), computing the weights of each distribution.

2.3 You can assume that k=2 for both clustering algorithms. But the initialization should not use the known membership of the data points (i.e. you actually know which distribution generates a data point).

2.4 You need to implement the algorithms yourself, rather than calling any known function.

2.5 Use Euclidean distance. You can decide the convergence criteria for the EM algorithm (e.g. the change of parameters < a threshold). BUT, you must clear describe this in your report.

3. (40 pts) Once you implement the two clustering algorithms, you need to compare and report the performance of the two clustering algorithms.

3.1 As you know which normal distribution produces a data point, compute the accuracy for each clustering algorithm (similar to classification)

3.2 Also compare the derived mean and standard deviation of the two clusters with the known ones.

3.3 Generate multiple input data sets with different mean and standard deviation. In particular, evaluate the performance change of the two clustering algorithm when the two normal distributions become closer (distance between the mean becomes smaller with relatively big standard deviation). Visualize this comparison (e.g. using curves show the performance change etc.)

3.4 Draw the conclusion about the running time and accuracy of the two clustering algorithms.

What to turn in:

1. One data set that is the mixture of data points sampled from two normal distributions. Specify the parameters of the two distributions.
2. The codes and a read me about how to run your program. The program must accept the input file as a parameter as below: Program < input data set>

The output should be two files, each containing data points in one cluster.

3. The report should include a clear description of your implemented method, part 2.5, part 3, and also any optimizations, thresholds, and variants you made for the standard algorithm.
4. The report must also include the first three steps of both k-means and EM for the following input data points:

   1 2 3 5 7
   Let k=2. Clearly describe the posterior probabilities, weights, and the mean/standard deviation of each iteration in EM.