## Machine Learning – Comparing Two Cluster Analysis Algorithms

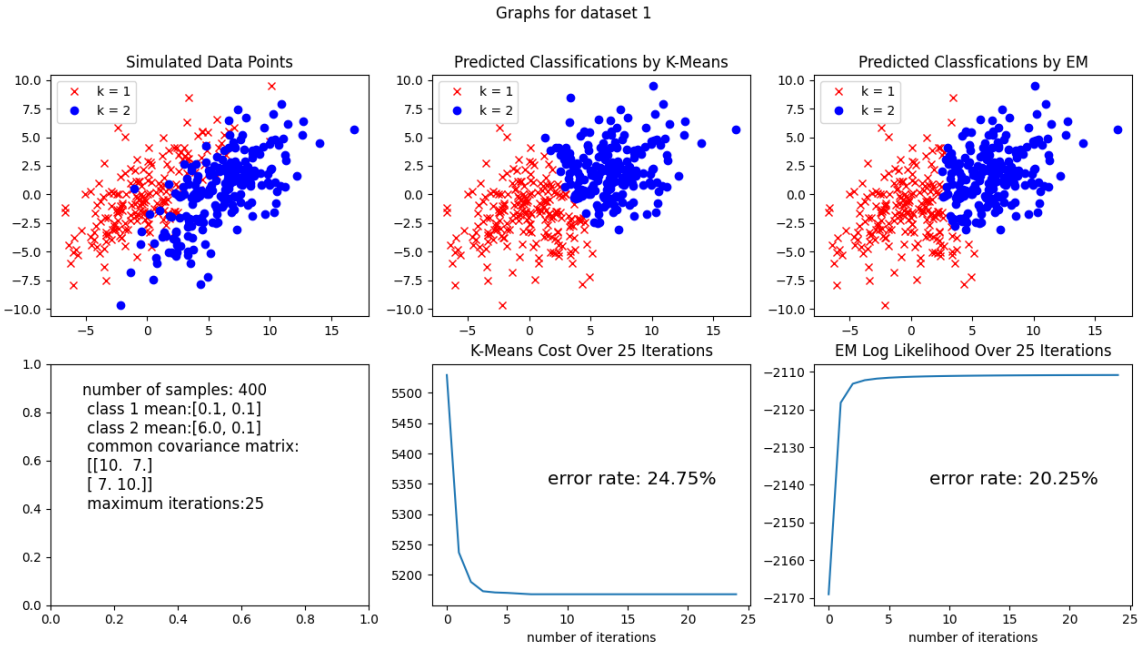
### Overview:

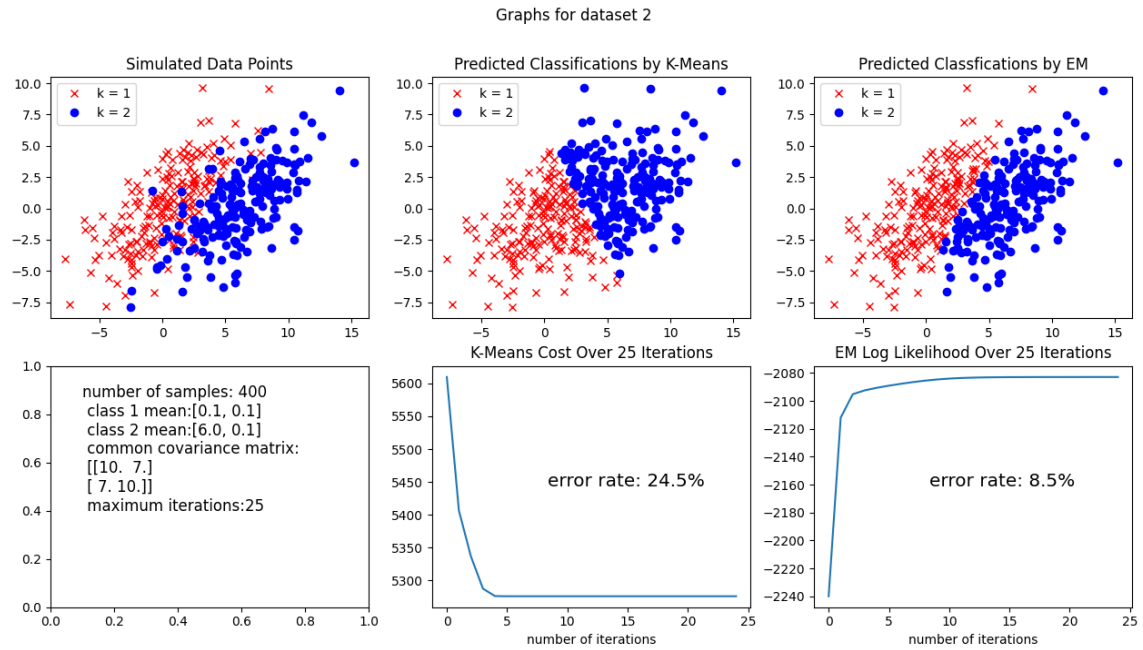
The goal of this project is to compare the accuracy and efficiency of the following two cluster analysis algorithms: K-Means algorithm and EM (Expectation Maximization) algorithm. Datasets were randomly generated and algorithms were implemented by hand using Python Pandas.

### Steps:

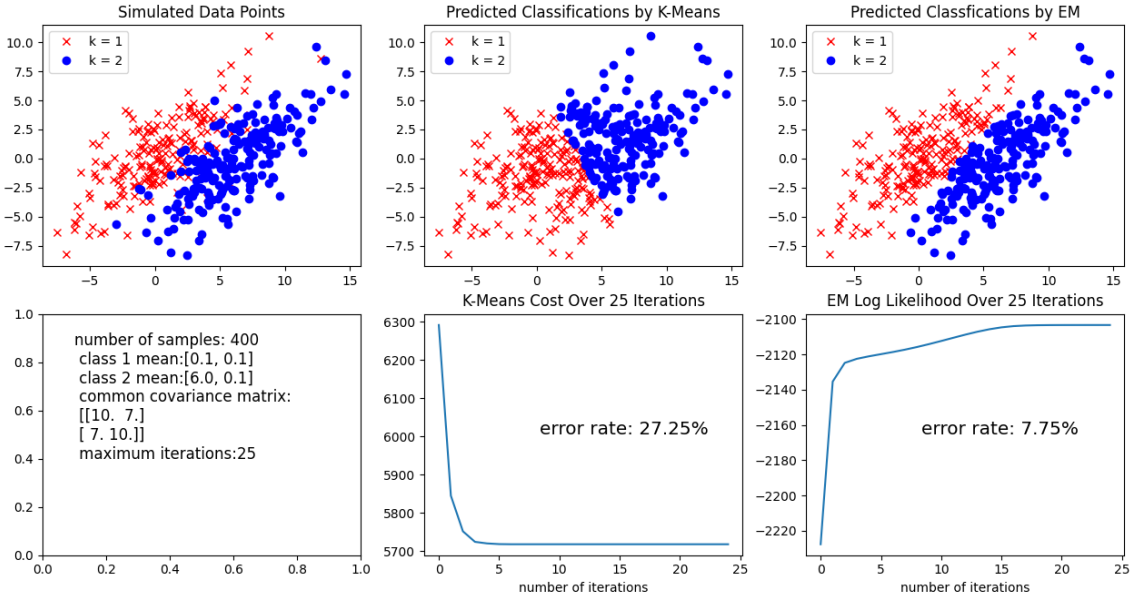
1. Generate 5 datasets from a multi-normal distribution. For each dataset, label half of the samples as “class 1” and the remaining as “class 2”. Shuffle the generated datasets.
2. Apply both algorithms to the 5 datasets, plot the predicted clusters and error rates.
3. Using cost/loglikelihood functions to visualize the convergence rate of both algorithms
4. Compare the results and draw conclusions.

### Results:

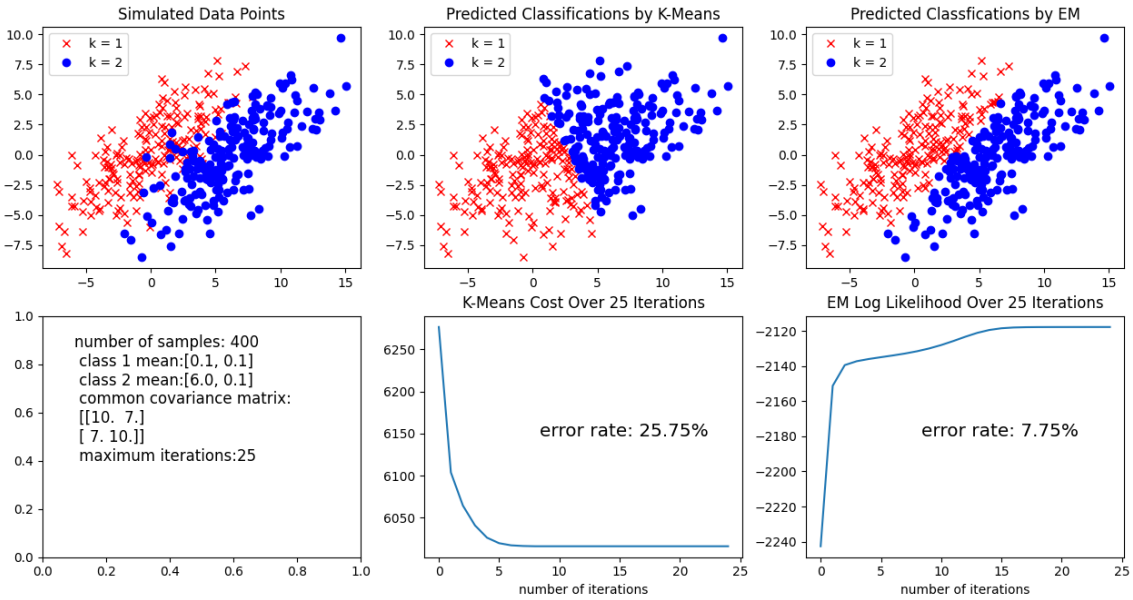


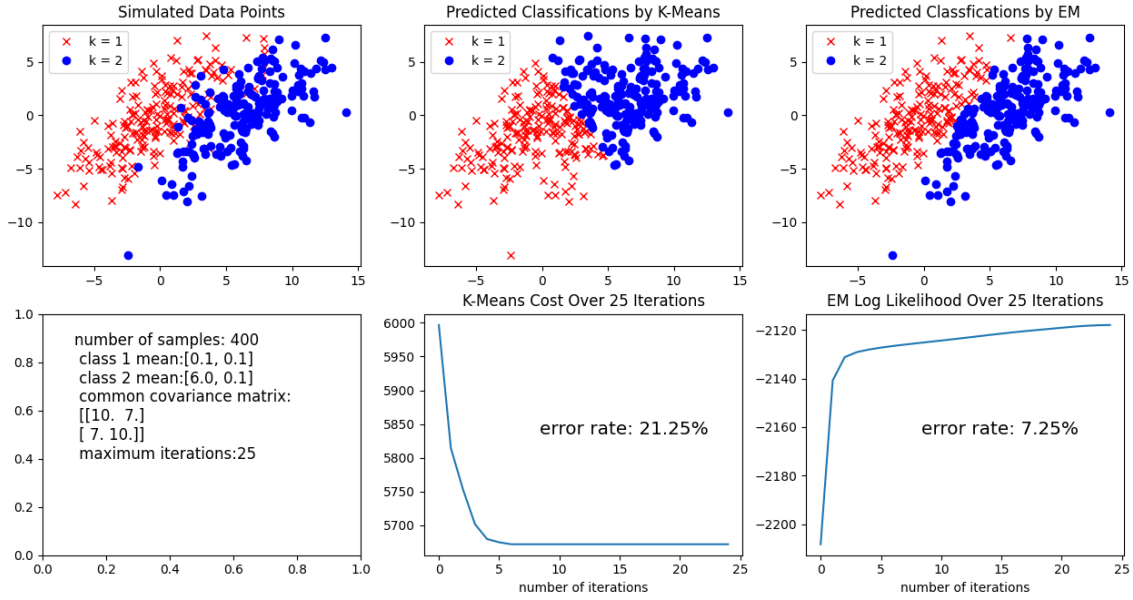
Graphs for dataset 3




Graphs for dataset 4




Graphs for dataset 5



**Conclusion:**

The K-Means cost graph and EM loglikelihood graph are indicators of convergence rates of the corresponding algorithms. The algorithm converges when the cost/loglikelihood curve is stable, that is, no rapid increase or decrease.

For all 5 datasets, EM algorithm has a higher accuracy (20%-5% higher than K-Means algorithm), however, EM algorithm has a lower convergence rate. It takes about 15 iterations for EM algorithm to converge, whereas K-Means algorithm only needs 5 iterations. Hence EM algorithm has a higher computational cost.

**Written by Liman Wei and Posted on https://github.com/weiliman/Course-Assignments**