Homework 3
IST 597
Physics-Informed Machine Learning

Subarna Pudasaini (sfp5828@psu.edu)

Question 1

Implement a K-means algorithm from scratch (so do not use any clustering algorithms) and apply it to a multi-dimensional data set of multiple Gaussian blobs that are sufficiently separated (you can generate these blobs yourself). Check for the convergence times of K-means (remember I need statistics here so run multiple trials for different initial positions of cluster centroids) with varying dimensionality of the data (so start with a 2D blob and then make it 3D, 4D, etc).

Ans:

Code: hw3.ipynb

The simulation results for other combinations of the number of data points, clusters and trials are in the Jupyter notebook.

- No of Datapoints: 1000
- No of Clusters: 4
- No of Trials: 10

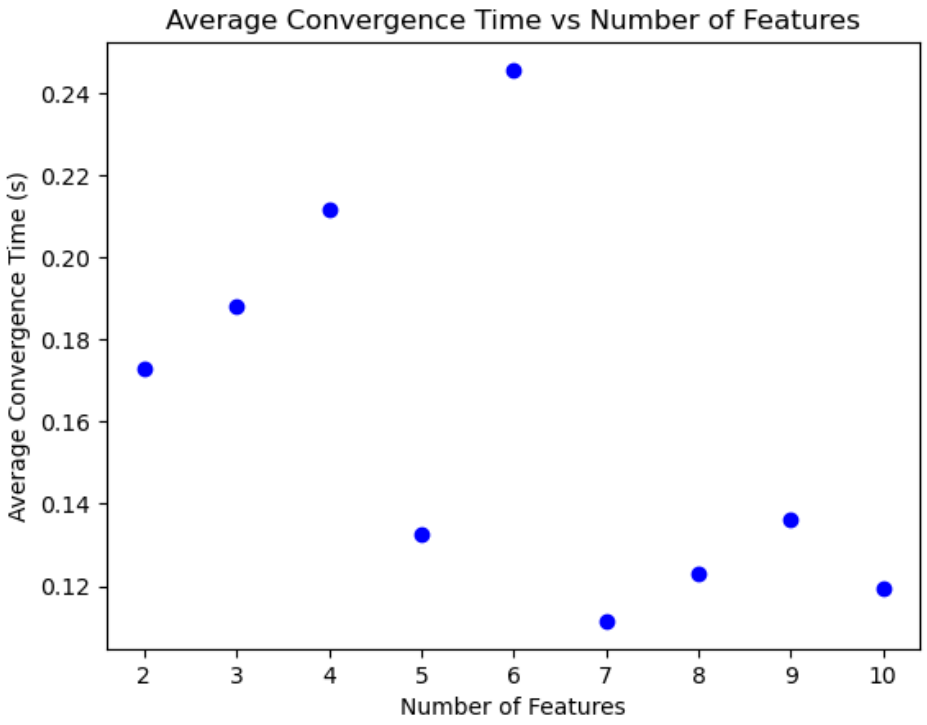

Figure 1: Average Convergence Time vs Number of Features (KMeans Clustering)

Question 2

Implement K-means++ for initializing cluster centroids more effectively and re-run the analysis above.

Ans:

Code: hw3.ipynb

The simulation results for other combinations of the number of data points, clusters and trials are in the Jupyter notebook.

- No of Datapoints: 1000
- No of Clusters: 4
- No of Trials: 10

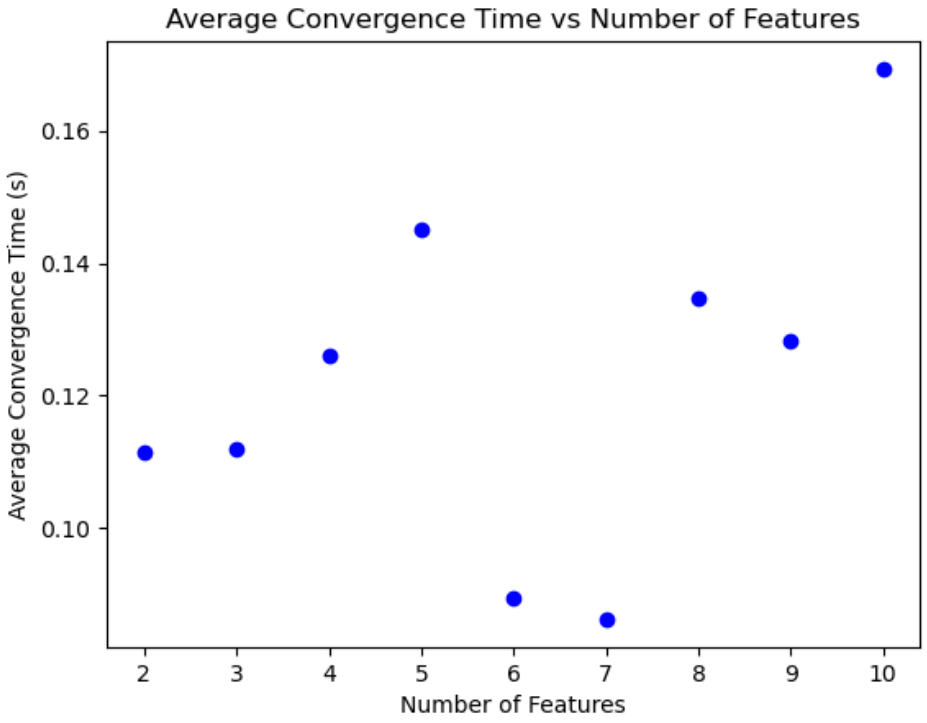

Figure 2: Average Convergence Time vs Number of Features (KMeans++ Clustering)

Question 3

Implement a Gaussian mixture model for clustering this data set and re-run the analysis above.

Ans:

Code: hw3.ipynb

The simulation results for other combinations of the number of data points, clusters and trials are in the Jupyter notebook.

- No of Datapoints: 1000

- No of Clusters: 4
- No of Trials: 10

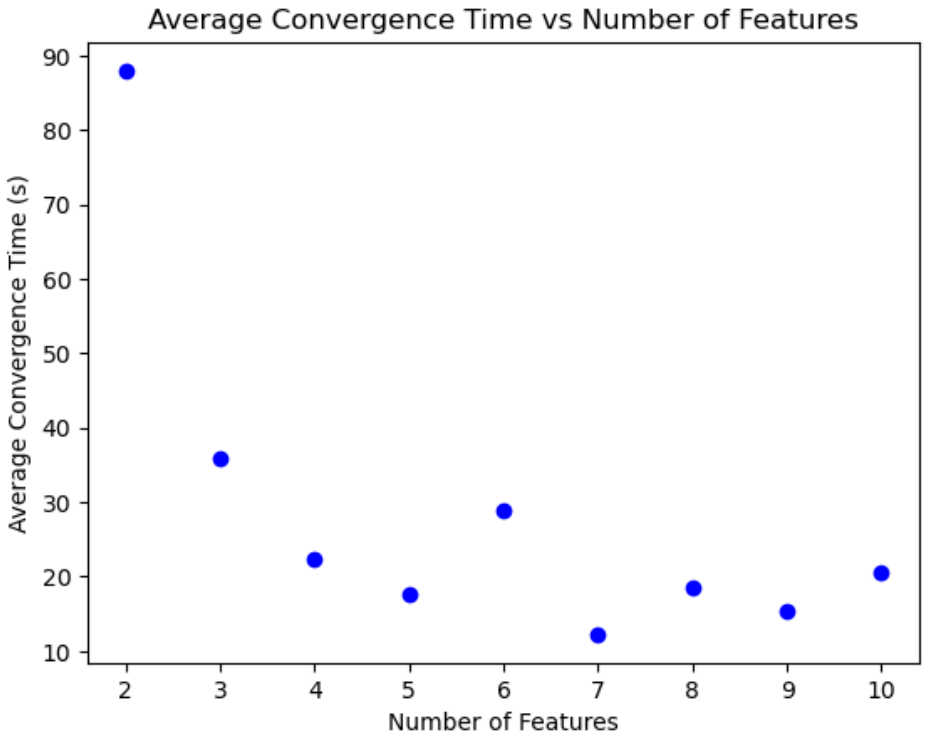

Figure 3: Average Convergence Time vs Number of Features (Gaussian Mixture Models)