Distributed Mixture of Gaussian Clustering

Milestone Report

Hongzhe Cheng, Yuzhou Wang

1. Summary of Work

In the past few weeks, we started to implement the project from scratch. We achieved mainly 2 goals in different directions: dataset sampling and sequential algorithm implementation.

For the whole project to work, we first need to create scripts that automatically generate sample dataset and clusters that are easy to scale (for the number cluster and data points), and easy to visualize. We created the scripts in Python to write into a data file which would then be read by the C++ algorithm. We also created the 2-d visualization for the sample dataset so that we can see how they are distributed, and in the same graph we will be able to visualize the clusters with its trained parameters (mean, covariance matrix).

For the sequential algorithm, we first created many linear algebra functions including matrix inverse, multiplication, transpose and so on. Based on these functionalities, we are able to both get the probability of a data point according to a multi-dimensional gaussian distribution, and then complete the sequential version of the algorithm.

2. Progress, Plan and Goals

We meet the plan perfectly on finishing the sequential algorithm and basic data visualization. We will be able to get all planned work done. For the “Nice to Have”, we will try before the poster session to see if the idea is actually doable and if time allows.

Goals:

- Create the correct baseline GMM EM algorithm for clustering
- Parallelize the algorithm with MPI to separate nodes/threads.
- Benchmark the performance, and create data visualizations
- Try a real world data set and see the performance in realistic settings (workload and accuracy)

Here is our future plan:

- **4/19 - 4/22:** Finalize data visualization for clearer graphs, create framework for MPI distributed algorithm (Hongzhe)
- **4/23 - 4/25:** Further implement the distributed algorithm in detail (Yuzhou)
- **4/26 - 4/29:** Finalize distributed algorithm, test for correctness, try real world dataset (Hongzhe)
- **4/30 - 5/5:** Performance testing, report writing (Hongzhe), extra goal algorithm if time allows (Yuzhou)

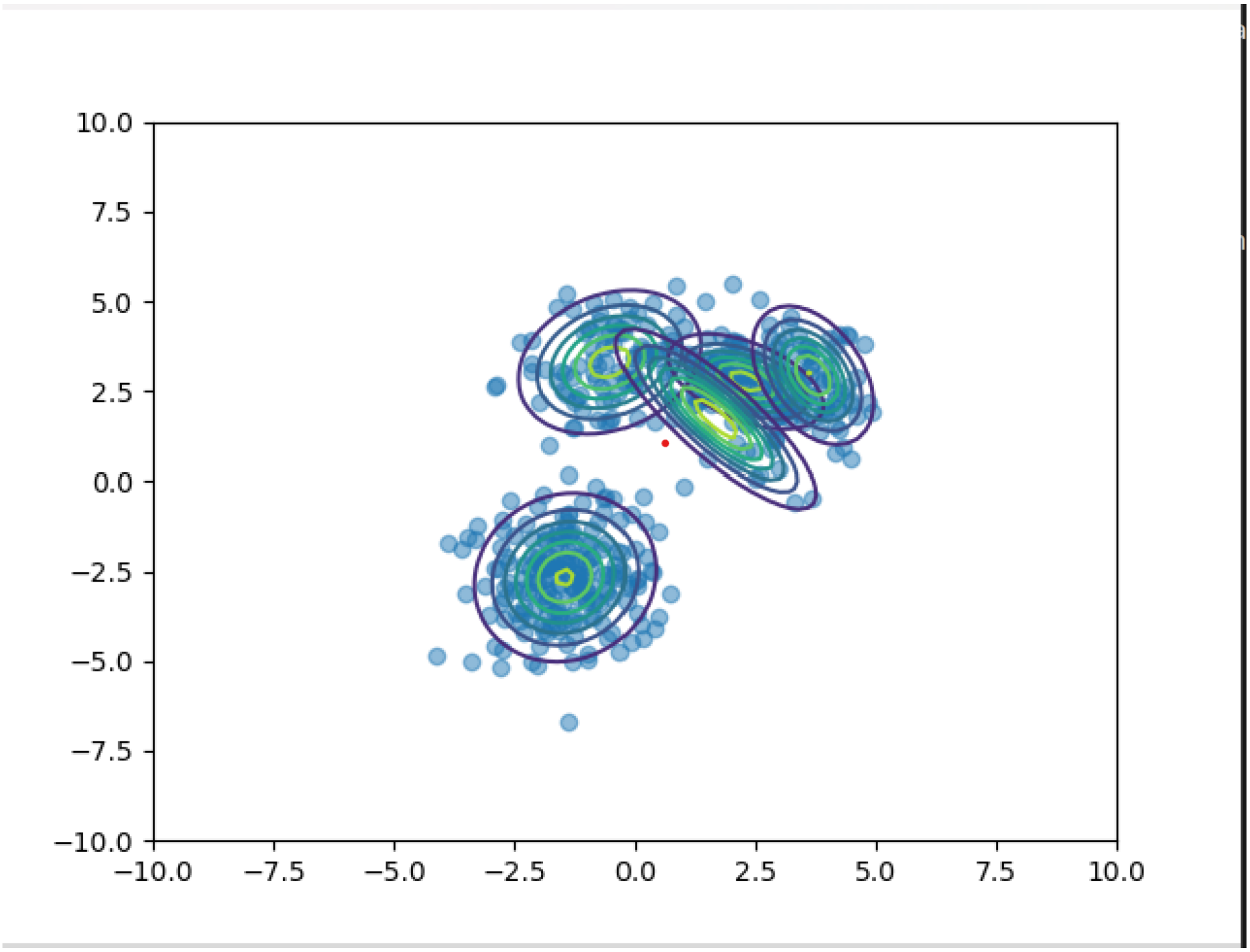
3. Deliverable at Final Presentation

- The demo would contain the image illustration of how our algorithm performs and successfully cluster the dataset

- The demo would presents the speedup graphs we have been creating for the labs in this course
- The demo would present the real world dataset we obtained, and how it classifies and clusters it

4. Preliminary Results

Here we will show the manual dataset sampling process result visualization and the result of visualizing the algorithm results (cluster means and covariance) as 2-d eclipse to showcase the correctness of the algorithm.



5. Concerns

We are concerned about the following things:

1. How can we convert the real world dataset into a more sensible way for the algorithm to cluster? Should we discard all non-numerical features or use techniques like Bag of Words etc to convert them into vectors?

2. How hard would the sequential portion of the algorithm hurt the system? It would be less significant if the portion is too big and we achieve sublinear speedup.

3. Having the edge case of compatibility for C++ linear algebra we need to fix, where in rare cases we will get NAN as results