GMM clustering

Daniel Amirtharaj
damirtha@buffalo.edu

1 Objective

To apply GMM clustering to the Old faithful dataset to find clusters (non-spherical) in the dataset.

2 Analysis

2.1 GMM based Clustering

The k-means clustering algorithm is a hard clustering algorithm, making hard assigns to data points. A point must lie in one cluster or the other. This does not model some data distributions well. Distributions which may have overlapping clusters, will have data points belonging partially to a number of clusters. This is modelled effectively in the GMM based clustering algorithm, also known as the EM (Estimation Maximization) algorithm which does a soft assign instead of a hard assign.

Here, the data points are assumed to be sampled from a continuous function, or in this case the weighted superposition of different gaussian distributions, with each distribution representing a cluster. A cluster assign is simply the posterior probability of the cluster given the data point. The following iterative steps are followed to GMM clusters,

1. Initialize gaussian parameters (μ, Σ, π) for each distribution i (can be done randomly)
2. Repeat until convergence (when gaussian parameters remain unchanged),
    1. Compute the posterior probability of cluster C_i given data point x_j, $P(C_i|x_j)$. This is the soft assign, assigning probabilistic membership of the data point in each cluster.
    2. Using the probabilistic cluster assignments from the previous step, recompute gaussian parameters (μ, Σ, π) for each cluster C_i, using maximum likelihood estimation.

3 Method

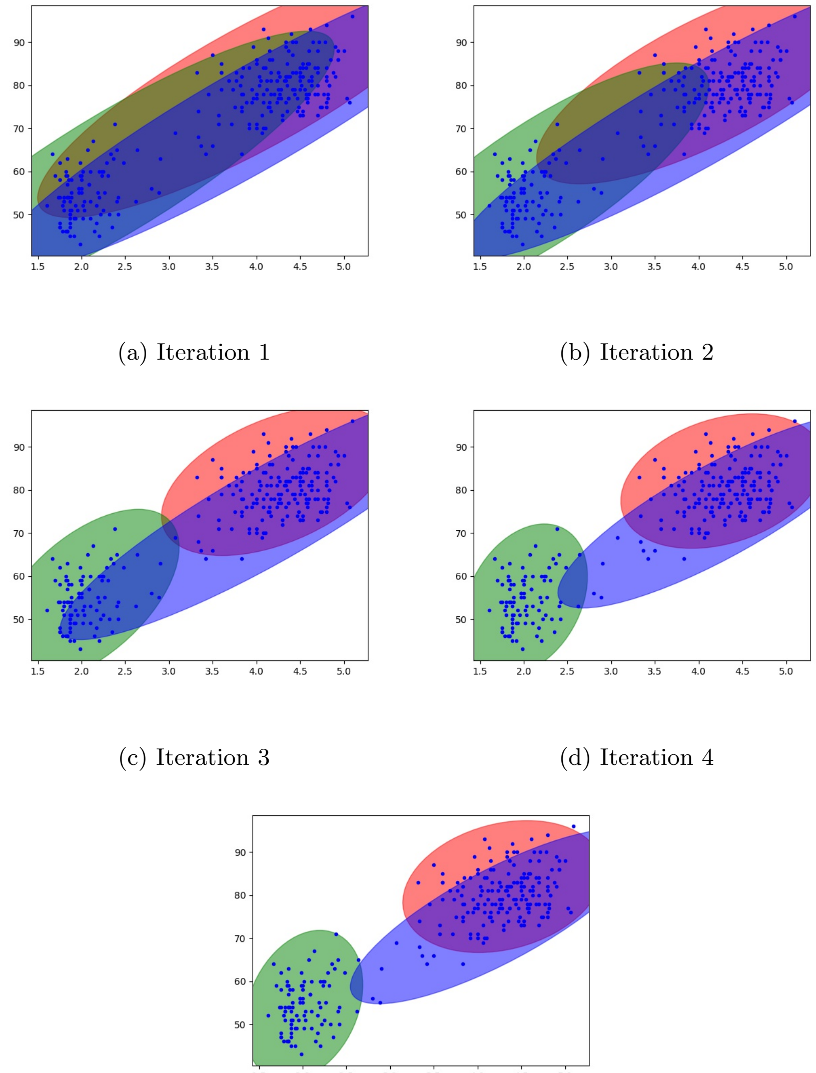
The following steps were followed to perform the GMM clustering algorithm on a given sample dataset and the old faithful dataset,

1. Model parameters such as k and gaussian parameters (μ, Σ, π) were initialized as given in the project description.
2. GMM was applied on the sample dataset and the old faithful dataset following the algorithm as described in section 3.2.1, and scatter plots at certain snapshots of the algorithms iterations were captured and saved.

4 Results

The following results were obtained at various steps of the program.

Running the EM algorithm on the old faithful dataset gave the following cluster assignments,



(a) Iteration 1

(b) Iteration 2

(c) Iteration 3

(d) Iteration 4

(e) Iteration 5

Figure 1: EM algorithm applied on the faithful dataset with k=3 and other given initial parameters for the gaussians. First 5 Iterations.