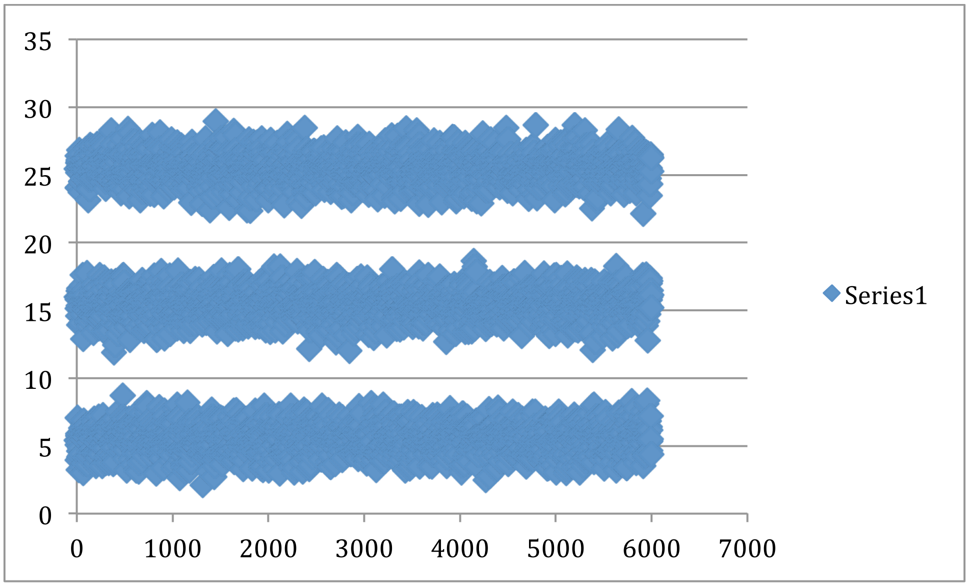
Project Report – EM clustering
Tanvi Patil

Number of clusters: Instead of running my algorithm several times, trying to figure out the number of cluster, I simple found that out by visualizing the data. The input data provided, is segregated into **three clusters**. This is easily deciphered from the scatter plot of the data provided. The scatter plot is shown below:



Initialization Strategy used: Initialization of mean to **random value and** variance to a multiple of the actual variance**.**
Reason : When the variance was initialized to random value, instead of multiple of actual variance over entire data set, the clusters converged at local maxima.

Sample Run Outputs:

```
1.
Initial GMM Parameters:
Mean[1] : 26.02             Mean[2] : 26.98             Mean[3] : 6.91
Variance[1] : 1             Variance[2] : 1             Variance[3] : 1
Final GMM Parameters :
Mean[1] : 26.06             Mean[2] : 25.34             Mean[3] : 10.65
Variance[1] : 0.95          Variance[2] : 0.78          Variance[3] : 27.68

2.
Initial GMM Parameters :
```

```
Mean[1] : 24.95             Mean[2] : 25.93             Mean[3] : 17.91
Variance[1] : 607.63        Variance[2] : 67.51         Variance[3] : 337.57
Final GMM Parameters :
Mean[1] : 15.45             Mean[2] : 25.49             Mean[3] : 5.51
Variance[1] : 0.97          Variance[2] : 1             Variance[3] : 1.03

3.
Initial GMM Parameters :
Mean[1] : 5.08              Mean[2] : 25.29             Mean[3] : 25.33
Variance[1] : 472.6         Variance[2] : 270.06        Variance[3] : 0
Did not converge

4.
Initial GMM Parameters :
Mean[1] : 26.02             Mean[2] : 6.53              Mean[3] : 7.54
Variance[1] : 0             Variance[2] : 405.09        Variance[3] : 202.54
Did not converge

5.
Initial GMM Parameters :
Mean[1] : 24.87             Mean[2] : 6.09              Mean[3] : 6.33
Variance[1] : 607.63        Variance[2] : 337.57        Variance[3] : 135.03
Final GMM Parameters :
Mean[1] : 25.49             Mean[2] : 15.45             Mean[3] : 5.51
Variance[1] : 1             Variance[2] : 0.97          Variance[3] : 1.03

6.
Initial GMM Parameters :
Mean[1] : 16.71             Mean[2] : 15.45             Mean[3] : 16.01
Variance[1] : 202.54        Variance[2] : 135.03        Variance[3] : 270.06
Final GMM Parameters :
Mean[1] : 25.49             Mean[2] : 5.51              Mean[3] : 15.45
Variance[1] : 1             Variance[2] : 1.03          Variance[3] : 0.97

7.
Initial GMM Parameters :
Mean[1] : 5.27              Mean[2] : 16.46             Mean[3] : 24.58
Variance[1] : 202.54        Variance[2] : 270.06        Variance[3] : 270.06
Final GMM Parameters :
Mean[1] : 5.51              Mean[2] : 15.45             Mean[3] : 25.49
Variance[1] : 1.03          Variance[2] : 0.97          Variance[3] : 1

8.
Initial GMM Parameters :
Mean[1] : 3.66              Mean[2] : 5.6               Mean[3] : 4.72
Variance[1] : 405.09        Variance[2] : 337.57        Variance[3] : 607.63
Final GMM Parameters :
Mean[1] : 15.45             Mean[2] : 5.51              Mean[3] : 25.49
Variance[1] : 0.97          Variance[2] : 1.03          Variance[3] : 1

9.
```

```
Initial GMM Parameters :
Mean[1] : 15.7                Mean[2] : 5.92                Mean[3] : 3.88
Variance[1] : 337.57          Variance[2] : 405.09          Variance[3] : 540.12
Final GMM Parameters :
Mean[1] : 25.49               Mean[2] : 5.51                Mean[3] : 15.45
Variance[1] : 1               Variance[2] : 1.03            Variance[3] : 0.97

10.
Initial GMM Parameters :
Mean[1] : 16.96               Mean[2] : 6.56                Mean[3] : 25.61
Variance[1] : 540.12          Variance[2] : 337.57          Variance[3] : 202.54
Final GMM Parameters :
Mean[1] : 15.45               Mean[2] : 5.51                Mean[3] : 25.49
Variance[1] : 0.97            Variance[2] : 1.03            Variance[3] : 1
```

Sensitivity to initialization: The cluster did not converge when any of the variance was initialized to zero.
Also, when all three means are same, or wrongly initialized (very far from actual mean), it takes times to converge.

Sample Run Outputs (Initialization of mean to random values and initialization of variance to 1.0):

```
1.
Initial GMM Parameters :
Mean[1] : 25.73               Mean[2] : 25.59               Mean[3] : 14.2
Variance[1] : 1               Variance[2] : 1               Variance[3] : 1
Final GMM Parameters :
Mean[1] : 26.06               Mean[2] : 25.34               Mean[3] : 10.65
Variance[1] : 0.95            Variance[2] : 0.78            Variance[3] : 27.68

2.
Initial GMM Parameters :
Mean[1] : 13.82               Mean[2] : 3.43                Mean[3] : 25.43
Variance[1] : 1               Variance[2] : 1               Variance[3] : 1
Final GMM Parameters :
Mean[1] : 15.45               Mean[2] : 5.51                Mean[3] : 25.49
Variance[1] : 0.97            Variance[2] : 1.03            Variance[3] : 1

3.
Initial GMM Parameters :
Mean[1] : 24.41               Mean[2] : 15.29               Mean[3] : 14.03
Variance[1] : 1               Variance[2] : 1               Variance[3] : 1
Final GMM Parameters :
Mean[1] : 25.49               Mean[2] : 15.45               Mean[3] : 5.51
Variance[1] : 1               Variance[2] : 0.97            Variance[3] : 1.03
```

4.
Initial GMM Parameters :
Mean[1] : 15.74   Mean[2] : 4.74   Mean[3] : 24.83
Variance[1] : 1   Variance[2] : 1   Variance[3] : 1
Final GMM Parameters :
Mean[1] : 15.45   Mean[2] : 5.51   Mean[3] : 25.49
Variance[1] : 0.97   Variance[2] : 1.03   Variance[3] : 1

5.
Initial GMM Parameters :
Mean[1] : 6.55   Mean[2] : 5.44   Mean[3] : 4.75
Variance[1] : 1   Variance[2] : 1   Variance[3] : 1
Final GMM Parameters :
Mean[1] : 20.31   Mean[2] : 6.13   Mean[3] : 4.87
Variance[1] : 28.1   Variance[2] : 0.48   Variance[3] : 0.66

6.
Initial GMM Parameters :
Mean[1] : 4.81   Mean[2] : 24.41   Mean[3] : 16.23
Variance[1] : 1   Variance[2] : 1   Variance[3] : 1
Final GMM Parameters :
Mean[1] : 5.51   Mean[2] : 25.49   Mean[3] : 15.45
Variance[1] : 1.03   Variance[2] : 1   Variance[3] : 0.97

7.
Initial GMM Parameters :
Mean[1] : 5.17   Mean[2] : 5.43   Mean[3] : 15.89
Variance[1] : 1   Variance[2] : 1   Variance[3] : 1
Final GMM Parameters :
Mean[1] : 4.87   Mean[2] : 6.13   Mean[3] : 20.31
Variance[1] : 0.66   Variance[2] : 0.48   Variance[3] : 28.1

8.
Initial GMM Parameters :
Mean[1] : 6.46   Mean[2] : 6.95   Mean[3] : 25.67
Variance[1] : 1   Variance[2] : 1   Variance[3] : 1
Final GMM Parameters :
Mean[1] : 5.51   Mean[2] : 15.45   Mean[3] : 25.49
Variance[1] : 1.03   Variance[2] : 0.97   Variance[3] : 1

9.
Initial GMM Parameters :
Mean[1] : 17.54   Mean[2] : 24.61   Mean[3] : 24.39
Variance[1] : 1   Variance[2] : 1   Variance[3] : 1
Final GMM Parameters :
Mean[1] : 10.65   Mean[2] : 26.06   Mean[3] : 25.34
Variance[1] : 27.68   Variance[2] : 0.95   Variance[3] : 0.78

10.
Initial GMM Parameters :
Mean[1] : 26   Mean[2] : 15.69   Mean[3] : 24.65

```
Variance[1] : 1               Variance[2] : 1               Variance[3] : 1
Final GMM Parameters :
Mean[1] : 26.06               Mean[2] : 10.65               Mean[3] : 25.34
Variance[1] : 0.95            Variance[2] : 27.68           Variance[3] : 0.78
```

Performance after initialization of variance to 1:
The frequency of clusters getting stuck at local maxima increased after initialization of all the variances to 1.
I noticed that the clusters converged only if their initialization of mean was close enough to the actual mean. When the mean were initialized to values far from actual mean , the algorithm got stuck at local maxima.
The reason for this is that the calculation of posterior value is relative to the variance and mean. Thus, assignment of variance to a value relative to the actual variance over data speeds up the convergence.

Conclusion: The clusters converged faster when initialized to multiple of actual variance.

However, the convergence got stuck at local maxima, or took time, when initialized to variance of 1.00.

Initialization to a multiple of actual variance across data ensured faster convergence and rarely got stuck at local maxima.