CMPE 544: Pattern Recognition (Fall 2020)

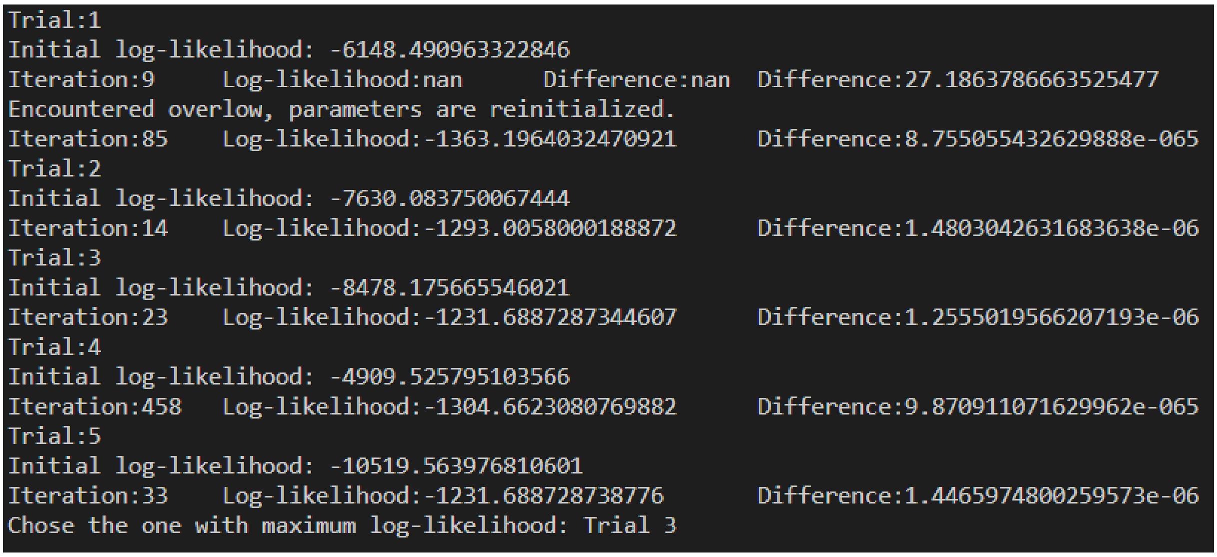
Homework2 Report – Yaşar Alim Türkmen – 2019700123

Implementation Details

The program has a main function called *run().* The function takes a number n as a trial number and runs EM algorithm for n times and takes the solution with the maximum log-likelihood as the optimal one. Based on that solution, it also assigns clusters to the points and plots them. To make the program more readable, a class called GM (Gaussian Mixture) is used. This object both stores the data and means and covariance matrices for Gaussian distributions. EM algorithm is also implemented under this class.

In each trial, *run()* function calls *EM()* function which follows all of EM steps one by one. First, it creates an instance of GM object which initializes random means (between minimum and maximum values of features) and $\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$ covariance matrices and equal mixing coefficient values for each 3 clusters. After initialization, E-step is processed (*E_step()*). Conditional probabilities of distributions are computed for the given data. These probabilities are stored in GM instance as *expectation_matrix*. Then, in M-step, mixing coefficient values, means and covariance matrices of clusters are updated respectively(*update_coeffs(), update_means(), update_covars()*). Finally, log-likelihood is calculated for iterated parameters in *is_converged()* function. If the difference of two successive iterations' log-likelihood values is less than the threshold, which is set to 5.10^{-5}, EM algorithm stops and returns GM object and last log-likelihood value. Sometimes, it is possible to encounter overflow or singularity problem for covariance matrix. Therefore, when it occurs, the parameters of GM object are reinitialized.

Based on returned last log-likelihood values, optimal solution is chosen:

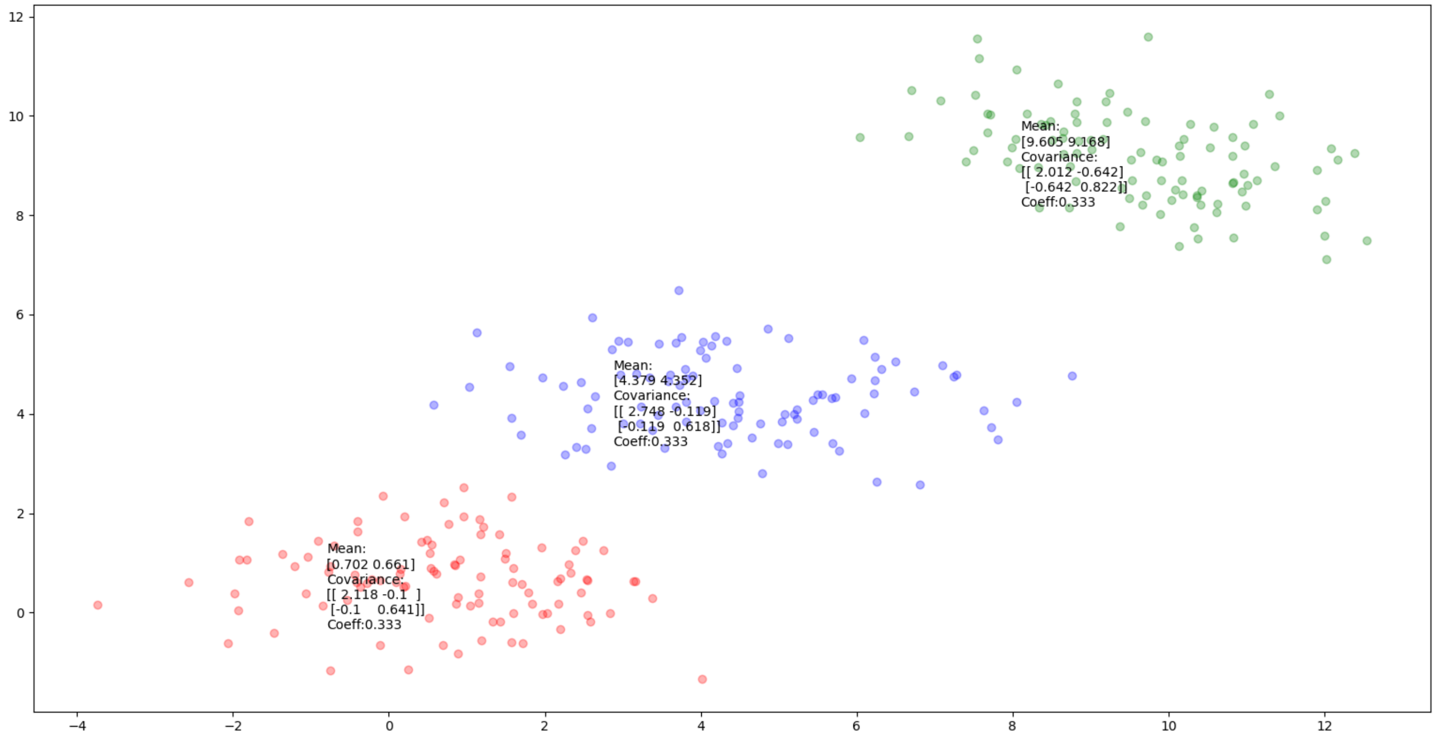


```
Trial:1
Initial log-likelihood: -6148.490963322846
Iteration:9     Log-likelihood:nan      Difference:nan  Difference:27.1863786663525477
Encountered overlow, parameters are reinitialized.
Iteration:85    Log-likelihood:-1363.1964032470921      Difference:8.755055432629888e-065
Trial:2
Initial log-likelihood: -7630.083750067444
Iteration:14    Log-likelihood:-1293.0058000188872      Difference:1.4803042631683638e-06
Trial:3
Initial log-likelihood: -8478.175665546021
Iteration:23    Log-likelihood:-1231.6887287344607      Difference:1.2555019566207193e-06
Trial:4
Initial log-likelihood: -4909.525795103566
Iteration:458   Log-likelihood:-1304.6623080769882      Difference:9.870911071629962e-065
Trial:5
Initial log-likelihood: -10519.563976810601
Iteration:33    Log-likelihood:-1231.688728738776       Difference:1.4465974800259573e-06
Chose the one with maximum log-likelihood: Trial 3
```

The maximum log-likelihood reached by the algorithm is **-1231.68883**.

The parameters estimated by EM algorithm and the plot are given below:

	Mean	Covariance Matrix	Mixing Coefficient
Cluster1 (Red)	$\begin{bmatrix}0.702\\0.661\end{bmatrix}$	$\begin{bmatrix}2.118 & -0.1\\-0.1 & 0.641\end{bmatrix}$	0.333
Cluster2 (Blue)	$\begin{bmatrix}4.379\\2.352\end{bmatrix}$	$\begin{bmatrix}2.748 & -0.119\\-0.119 & 0.618\end{bmatrix}$	0.333
Cluster3 (Green)	$\begin{bmatrix}9.605\\9.168\end{bmatrix}$	$\begin{bmatrix}2.012 & -0.642\\-0.642 & 0.822\end{bmatrix}$	0.333


Mean:
[0.702 0.661]
Covariance:
[[2.118 -0.1]
[-0.1 0.641]]
Coeff:0.333
Mean:
[4.379 4.352]
Covariance:
[[2.748 -0.119]
[-0.119 0.618]]
Coeff:0.333
Mean:
[9.605 9.168]
Covariance:
[[2.012 -0.642]
[-0.642 0.822]]
Coeff:0.333