WPI

CS 534 Artificial Intelligence
Assignment 3

Group 10

March 30, 2018

Group Member

Yixuan	Jiao	yjiao@wpi.edu
Yinkai	Ma	yma7@wpi.edu
Jiaming	Nie	jnie@wpi.edu
Pinyi	Xiao	pxiao@wpi.edu

1 Procedure of Expectation and Maximization

The EM algorithm will initialize the related parameters θ as Gaussian distribution parameters, then perform the expectation and maximization process.

At the beginning, we randomly pick K number of points as the initial state, and then generate clusters by using reasonable variances. Then, we use E-step to calculate the probabilities of each point for being in the K different clusters, which is also known as the expected probability distribution. After having the probability distribution, we use M-step to recalculate the center points (mean) according to the probability distribution of E-step and find the new variance of each new cluster. Finally, repeat the E-step and M-step until the clusters convergent into stable clusters.

1.1 Expectation Procedure

- Calculate probability of data point x_j belongs to cluster i $P_{ij} = P(C = i|x_j)$
- According to Bayes' theorem, $P_{ij} = \alpha P(x_j|C = i)P(C = i)$
- $P(x_j|C = i)$ is the probability of point x_j belongs to cluster i, $P(C = i)$ is the weight of the probability.
- Calculate effective number $n_i = \sum P_{ij}$ of the data point is in cluster i.

1.2 Maximization Procedure

In the Maximization procedure, the mean μ and variance σ^2 will be updated according to the result obtained in Expectation procedure.

$$\mu_i = \sum P_{ij}/n_i \tag{1}$$

$$\sum_i = \sum_i P_{ij}(x_j - \mu_i)(x_j - \mu_i)^T/n_i \tag{2}$$

$$w_i = n_i/N \tag{3}$$

The expectation procedure is the process of calculation of expected value P_{ij} with the hidden variable Z_{ij}. In the maximization procedure, the log-likelihood will be updated and the parameters of different Gaussian models.

2 Random Starts and Number Determination

2.1 Cluster Number Determination

In the Extended EM part, the Bayesian Interference Criteria combined with the log-likelihood will be used to assess the clustering accuracy.

The log-likelihood is defined as the following:

$$L(\theta) = \sum_i P(X_i, C_j|\theta_j) \tag{4}$$

The equation of BIC is defined on the following:

$$BIC(k) = -2 \times log_likelihood + k \times ln(N) \tag{5}$$

In the iterations of log-likelihood, for different iterations, the log-likelihood and BIC values will vary to a large extent sometimes, which means that variance of the BIC will be large. The phenomenon will show up under some occasions that the cluster number is not very accurate.

In this assignment, the variance Var is given and the program will restart if the change exceeds some degree.

The calculation procedure is on the following:

- For certain cluster number K, calculate BIC
- Random restart for cluster number K, obtain the result array of the BIC
- Calculate the variance var of the BIC list, which will is used to assess the algorithm accuracy

2.2 Restart Criteria

The restart for certain cluster number K will be performed for certain numbers.

3 Initialization of Cluster Centers

Given the cluster number K, random choose K points from the data array as the cluster number.

We randomly pick K number of points from the raw data as the initial cluster centers, and then initialize the clusters according to these centers. When this project initializes the variance of the center, the variance is randomly getting between the 1 times total data variance and 3 times total data variance. Such as, when $K = 3$, given a set of points, this project picks 3 different points of the data as the 3 clusters center. And then, calculate the variance of all the data, and get the variance of x and y, which is a, b, and then, set the variance of the cluster is random from a to 3a, b to 3b.

4 Termination Criteria on EM Algorithm

After each M-step, we have generated new centers of K clusters, then we compute the distance between the former center locations and the current center locations.

By adding up these distances together and dividing the sum with number K, we calculate a result measuring the progress of iteration. If the sum is larger than 0.1, then the iterations continue; if not, process complete and terminate it.

5 Data File Generation

5.1 Generated Data Description

The data that we created includes 3 Gaussian distribution, the means of them are [0,0], [10,10] and [20,20], the variance of them are [[100,1], [1,100]], [[10,1], [1,10]] and [[50,1], [1,50]]. When using extended EM, the result is like that:

5.2 Performance of EM Algorithm

According to the part 2, the criteria is the variance of the random restart process of the calculation of BIC value.

The figure on the following:

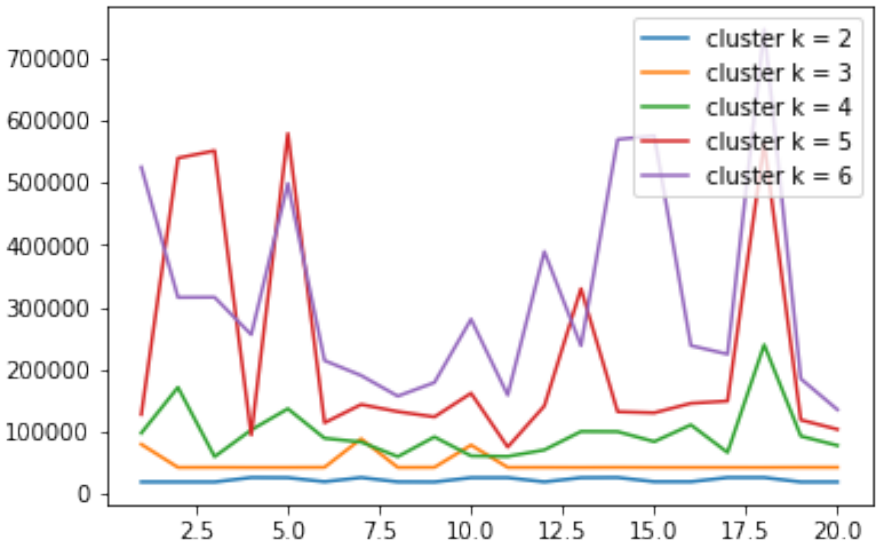

Figure 1: Generated Data File

Through the calculation of the variance of the normalized BIC value, when cluster number $K = 3$ gives the second best performance which is near to the $K = 2$ clustering result. Other clustering results give the 2 or 3 orders of magnitude.

5.3 Results o f the Extended EM Algorithm

If given the correct number of the clusters, in this project, I give the k=3, and run the basic EM, could get the figure of the log-likelihood vs iteration is like the figure below 2:

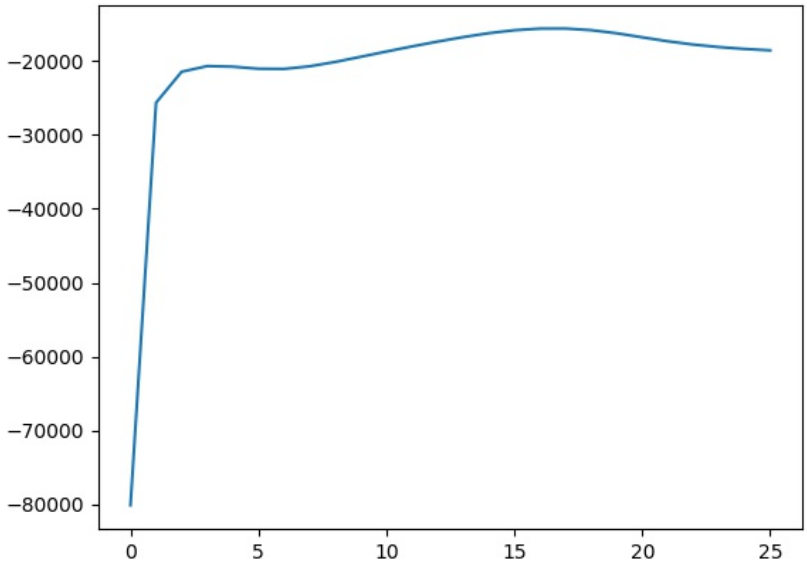

Figure 2: Log-Likelihood vs. Iteration

And get the means of three clusters are $\begin{bmatrix} 10.160927305532486 & 10.311809083724215 \\ 0.3022128376500947 & -0.028573619938553493 \end{bmatrix}$, $\begin{bmatrix} 10.160927305532486 & 10.311809083724215 \\ 0.3022128376500947 & -0.028573619938553493 \end{bmatrix}$, $\begin{bmatrix} 10.160927305532486 & 10.311809083724215 \\ 0.3022128376500947 & -0.028573619938553493 \end{bmatrix}$ and [20.672087754598735, 19.641971376001067], and the variance of them are [[12.577488828945143, 0.5694856082850692], [-0.05267943518194221, 10.673271231583815]], [[100.80295507545482, -1.0269924937393347], [-6.828986119723152, 96.44827857518334]] and , [[45.8920741435162, 1.505551578403153], [2.8330470399896046, 42.72457103704982]], the result is close to the correct, so it could find the correct means and variances of the clusters. And it does not assign points to the correct cluster, it just give the probabilities of each point assign each cluster.

6 Log-likelihood vs. Iterations

In this section, the data file is the file generated in part 5. This data file has 3 clusters generated from 3 different Gaussian Distribution models.

In figure illustrated the plot of the data, the data is illustrated on the following:

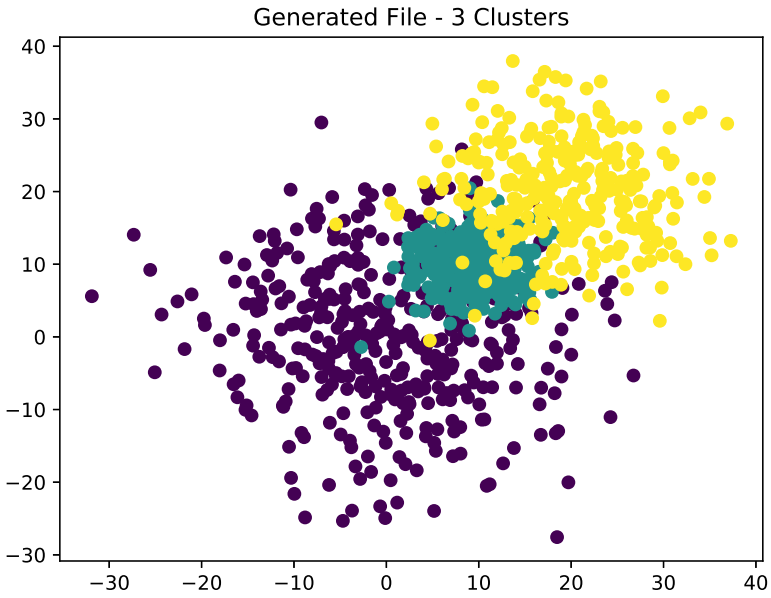

Figure 3: Generated Data File

6.1 K=2 Log-Likelihood vs. Iteration

For the cluster number $K = 2$, the graph of the log-likelihood and iteration is illustrated in figure 4.

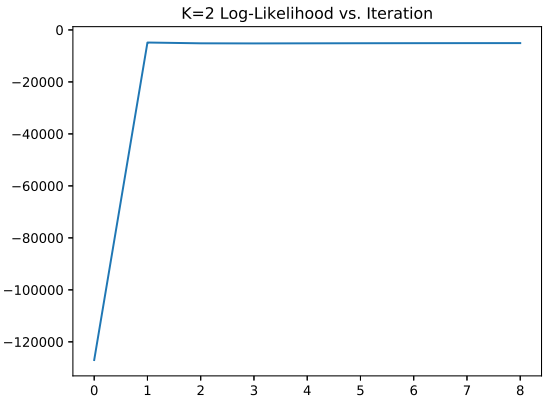

Figure 4: K=2 Log-Likelihood vs. Iteration

6.2 K=3 Log-Likelihood vs. Iteration

For the cluster number $K = 3$, the graph of the log-likelihood and iteration is illustrated in figure 5.

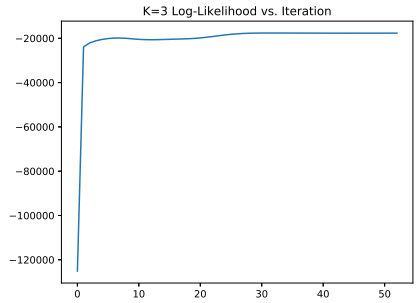

Figure 5: K=3 Log-Likelihood vs. Iteration

6.3 K=4 Log-Likelihood vs. Iteration

For the cluster number $K = 4$, the graph of the log-likelihood and iteration is illustrated in figure 6.

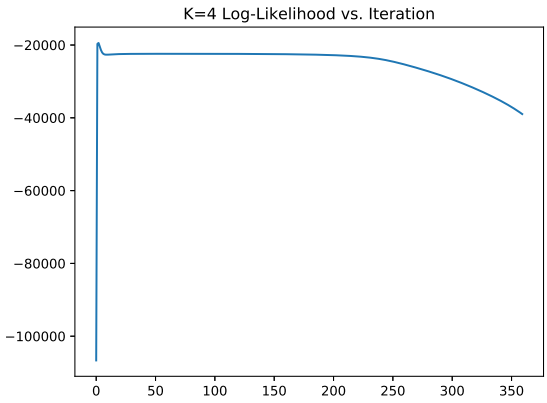

Figure 6: K=4 Log-Likelihood vs. Iteration

6.4 K=5 Log-Likelihood vs. Iteration

For the cluster number $K = 5$, the graph of the log-likelihood and iteration is illustrated in figure 7.

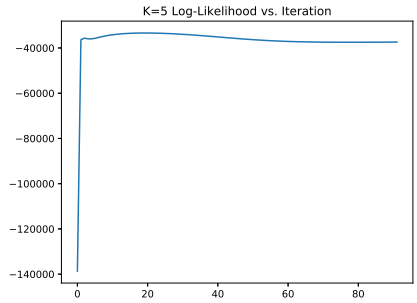

Figure 7: K=5 Log-Likelihood vs. Iteration

6.5 K=6 Log-Likelihood vs. Iteration

For the cluster number $K = 6$, the graph of the log-likelihood and iteration is illustrated in figure 8.

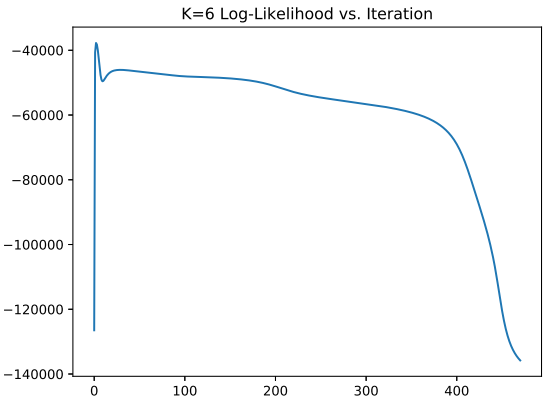

Figure 8: K=6 Log-Likelihood vs. Iteration

6.6 K=7 Log-Likelihood vs. Iteration

For the cluster number $K = 7$, the graph of the log-likelihood and iteration is illustrated in figure 9.

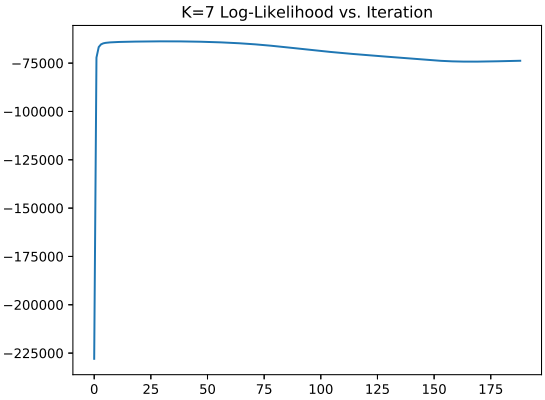

Figure 9: K=7 Log-Likelihood vs. Iteration

7 BIC Criteria and its Application

The equation of the Bayesian Interference Criteria is on the following:

$$BIC(k) = -2 \times log_likelihood + k \times ln(N) \tag{6}$$

In the equation 6, k is the number of the cluster, N is the number of all the points.

In the part 2, the extended EM part, for one test file, assume there are several kinds of cluster K, for each hypothesis K_i, the EM algorithm will test 20 iterations for the estimation of log-likelihood and BIC.

The variance of the 20 iterations of each cluster K_i will be a criteria to assess the stability of the accuracy of the clustering.

The cluster number K_i which is near to the data original clustering will have the smallest variance.