ECE – 20875
Homework – 10

-Abhimanyu Agarwal

Question: Run gmm-em.py on data.txt for k = 2,3,4,5,6 using a tolerance = 1. Write out the fitted mixture model formulas px(x) for each k. What do you observe about the log-likelihood ?

Answer:

K = 2:

$P_x(X)$ = 0.4705795783666359 * N(x | 3.1887189775230573 , 1.6993391855585842) + 0.529420421633364 * N(x | 11.00216752523677 , 9.562726217189121)

Log-likelihood: -1091.8565606736677

K = 3:

$P_x(X)$ = 0.24912166694205726 * N(x | 2.027654629380135 , 0.34397885444855747) + 0.2448077902430767 * N(x | 4.4675208509050135 , 0.08024393023611451) + 0.5060705428148662 * N(x | 11.315619284187953 , 7.741562183985081)

Log-likelihood: -999.7635118345282

K = 4:

$P_x(X)$ = 0.2501560667245576 * N(x | 2.02957992952585 , 0.3447708276434597) + 0.24984393326038062 * N(x | 4.470202987829373 , 0.08109582116454865) + 0.24962164945931076 * N(x | 8.892912170212226 , 0.3533543949093422) + 0.2503783505557511 * N(x | 13.902519769332546 , 1.5193385690769485)

Log-likelihood: -910.9246936882649

K = 5:

$P_x(X)$ = 0.2501863499084379 * N(x | 2.0298245032134132 , 0.3452264790042515) + 0.24981365008085846 * N(x | 4.470253909148835 , 0.08108098701022916) + 0.05729870974902365 * N(x | 8.727043845173746 , 0.30182166408564526) + 0.19233320327688483 * N(x | 8.942374156556664 ,

0.3580926774219373) + 0.2503680869847951 * N(x | 13.902688621600076 , 1.5187012730705338)

Log-likelihood: -910.8393241677736

K = 6:

0.22350958756466713 * N(x | 1.993645470310374 , 0.3328443914755654) + 0.031037572978978797 * N(x | 2.609159016167419 , 0.7764528446628084) + 0.24545284113408494 * N(x | 4.473299220137845 , 0.08044073864497918)+ 0.24948077589397516 * N(x | 8.892473164201851 , 0.35312495359798496)+ 0.012871024675185498 * N(x | 13.291403563763962 , 1.4886719360347807)+ 0.23764819775310847 * N(x | 13.933109118810618 , 1.5096453334216047)

Log-likelihood: -912.4968389442097

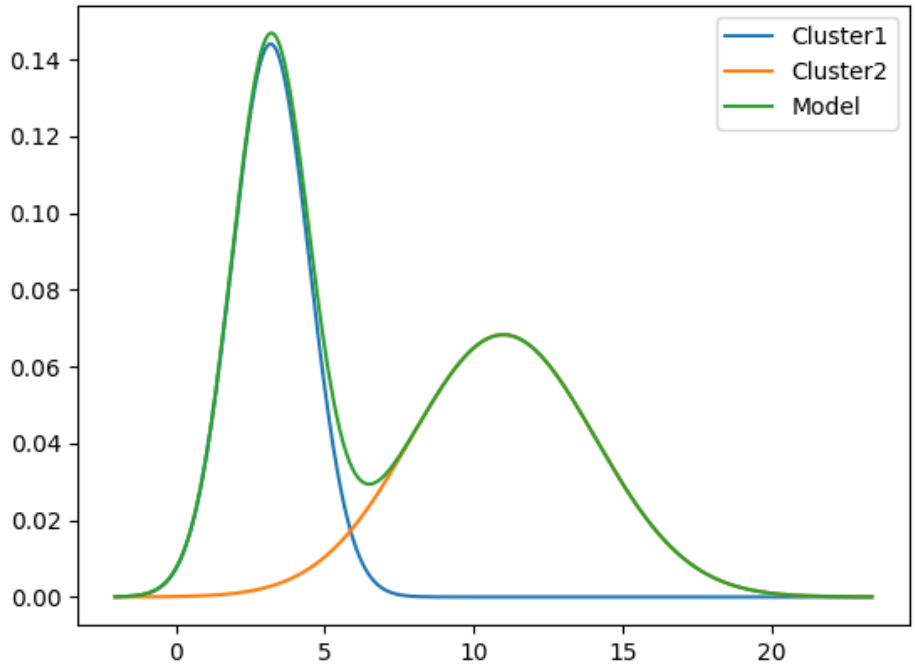
Observation:

In general, a trend in the value of the log-likelihood is observed as K value increases. It is observed from K = 2 to 5, the log-likelihood increases. However, at K = 6, a slight decrease is observed in the value. However, it is not significant enough.

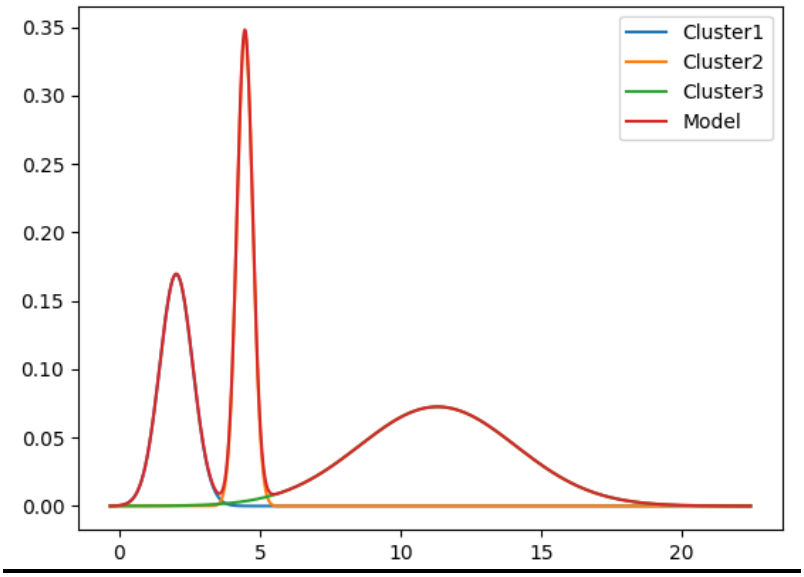
Question 4: Input your results from gmm-em.py for each value of k into gmm-visualize.py. Save your plots turn them in as a part of your writeup. How many clusters does this dataset have ? Explain.

Plots

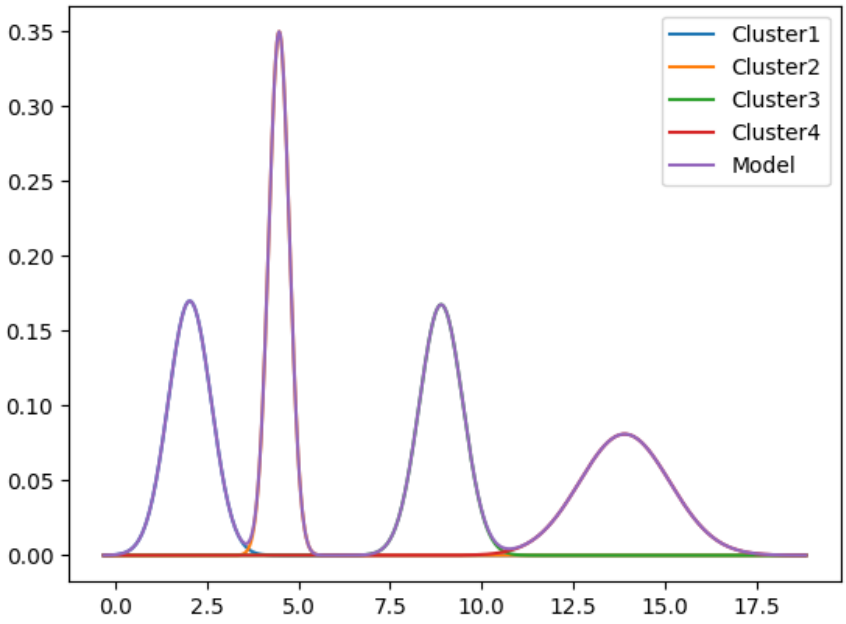
K = 2:



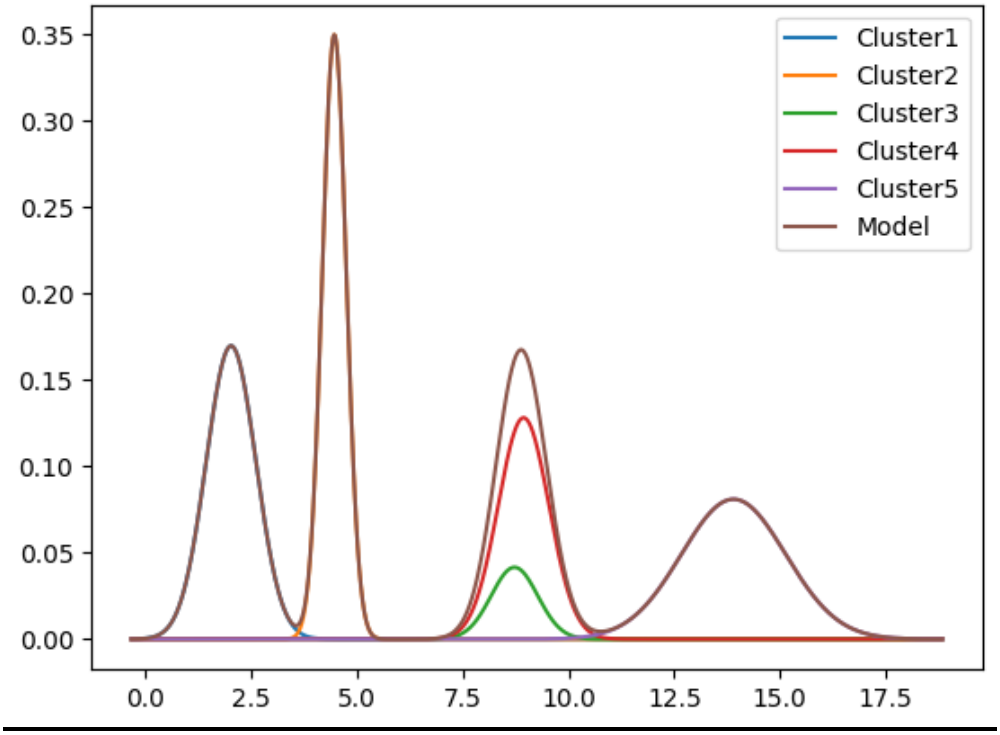
K = 3:



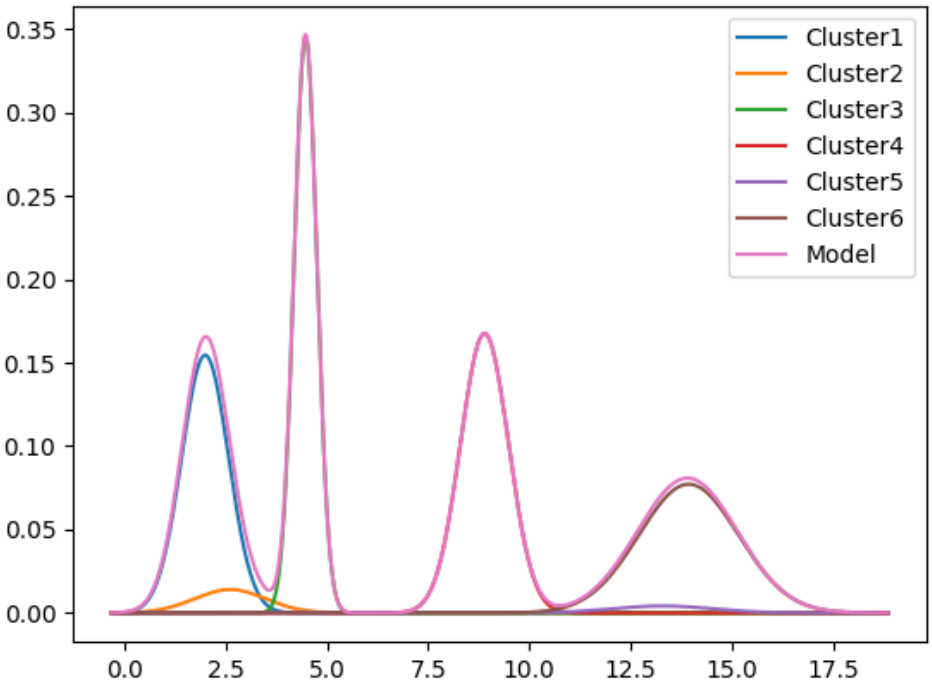
K =4:


Cluster1
Cluster2
Cluster3
Cluster4
Model
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5

K=5:


Cluster1
Cluster2
Cluster3
Cluster4
Cluster5
Model
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5

K = 6:



Observation:

From K = 4 to 5, it is observed that the change in log-likelihoods was less than the tolerance (which in this case is equal to 1). In other words, it is also seen that there is a significant overlap for K = 5 which implies that some Gaussians are invalid. So, it can be inferred that the convergence occurs at K = 4. Therefore, this data set has 4 clusters.