DEPARTMENT OF ELECTRICAL AND COMPUTER ENGINEERING

Homework Assignment No. 04:

HW No. 04: Gaussian Mixture Distribution Parameter Estimation

submitted to

Professor Joseph Picone
ECE 8527: Introduction to Pattern Recognition and Machine Learning
Temple University
College of Engineering
1947 North 12th Street
Philadelphia, Pennsylvania 19122

February 27, 2024

prepared by:

Leo Berman
Email: leo.berman@temple.edu

A. GENERATE DATA AND PLOT A SINGLE COMPONENT GAUSSIAN MIXTURE MODEL

In terms of the nature of the fit seen below, it's clear that a single component Gaussian distribution is insufficient for representing this dataset. There are clearly 3 peaks which represent the three means. The reason the middle peak is largest is due to the variance of the other sets overlap.

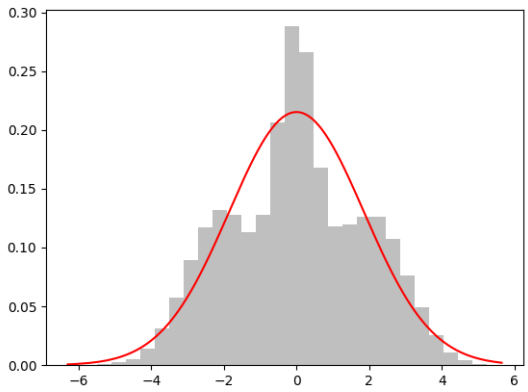

Figure 1: 1 component Gaussian mixture model

B. PLOT A 1,2, AND 3 COMPONENT GAUSSIAN MIXTURE MODEL

As can be seen below, a two component mixture model clearly does a significantly better job than a single component mixture model, but the nature of the fit really only begins to be well represented when we have a component for each mean.

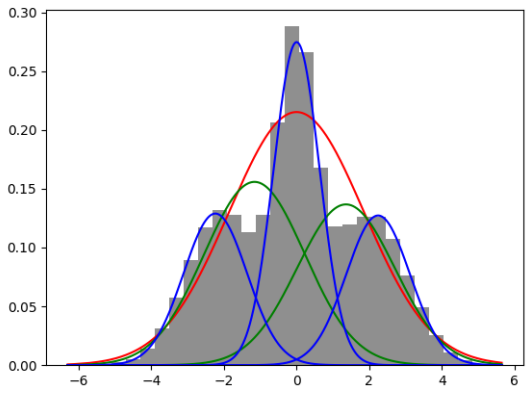

Figure 2: 1, 2, and 3 component Gaussian mixture model

C. PLOT A 1,2, AND 4 COMPONENT GAUSSIAN MIXTURE MODEL

When comparing to the figure from the last section, we can see that having a fourth mixture component maintains the representation of a three component mixture model, but the change is small and the peak that is second from the left seems to just be subcomponent of the middle peak shown in a three component mixture model.

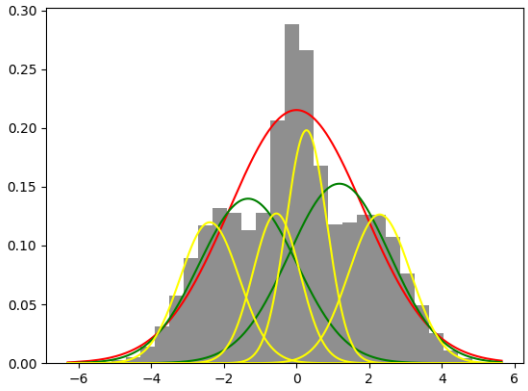

Figure 3: 1, 2, and 4 component Gaussian mixture model

D. PLOT THE LOG PROBABILITY OF THE DATA BELONG TO N COMPONENT GAUSSIAN MIXTURE MODELS

As can be seen, the log likelihood of the data set peaks when the Gaussian mixture model has three components, but levels out afterwards. Since this dataset has 3 peaks, this is logical and we can see the rapid rise of log likelihood betweek one componenet and three components.

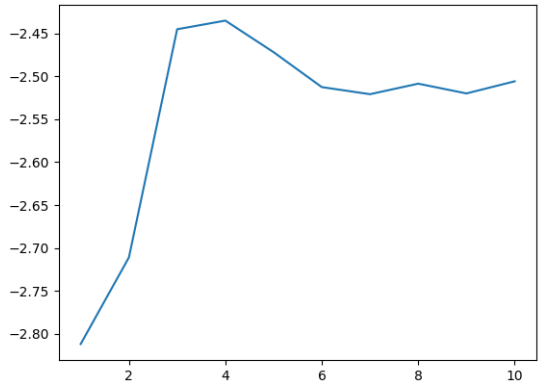

Figure 4: Log likelihood as a function of Gaussian mixture model components

E. PLOT THE LOG PROBABILITY OF DATA SET 13 AS FUNCTION OF N COMPONENT GAUSSIAN MIXTURE MODELS

When we plot the log probability of the two individual classes we can see that both plots have a similar shape in a different place. As expected, the more components we add to our Gaussian Mixture model, the better the performance of the model. Logically, a lack of knowledge of the dataset should mean that the more components there are, the better outliers are handled as well as the regular dataset.

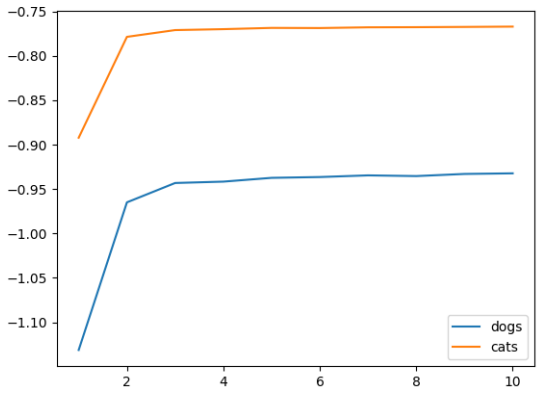

Figure 5: Log likelihood as a function of Gaussian mixture model components

F. PLOT DATA SET 13 USING HISTOGRAM TECHNIQUES AND 3 DIMENSIONAL GAUSSIAN MIXTURE MODELS

In the first plot, we can see a pretty well defined peak in the center of the data. There also seems to be some more local peaks surrounding the center, but this model could be relatively well modeled by a single component Gaussian mixture model. In the single component Gaussian mixture model, we can see that the looks similar to the histogram technique, but converges slightly more steeply. We can see this happening due to a loss of resolution with only a single component. For the two and four component models, we begin to see more of the shape of the histogram. The two component Gaussian mixture model builds peaks aroudn the gap in the center of the data, and while this is clearly a better representation, the four component mixture model begins to build a better representation by building a sort of well around the gap in the data.

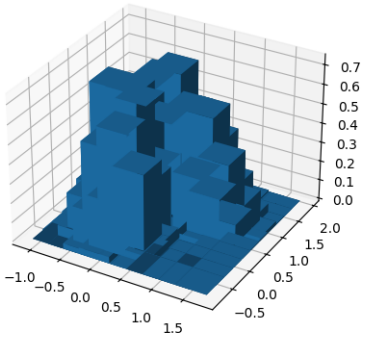

Figure 6: Represent the data with a histogram technique.

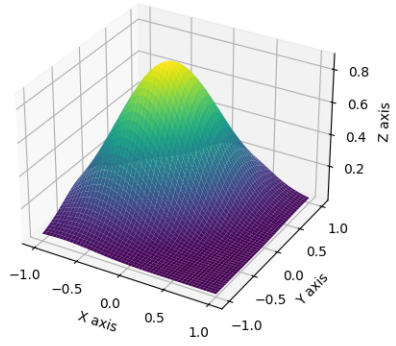

Figure 7: Represent the data as a single component Gaussian mixture model

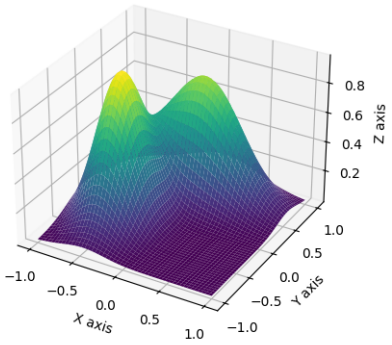

Figure 8: Represent the data as a two component Gaussian mixture model

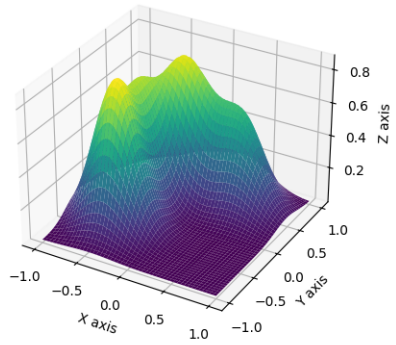

Figure 9: Represent the data as a four component Gaussian mixture model

G. PLOT DATA SET 13 USING HISTOGRAM TECHNIQUES AND 3 DIMENSIONAL GAUSSIAN MIXTURE MODELS

What we can see in the table, is that for the evaluation data, since our data seems to be fit to a Gaussian, the Gaussian mixture model does a better job. The statistical significance of that can be called to question. More importantly, we can see that the Gaussian mixture model almost consistently gets better with more components on the training model, but for the evaluation model it does better being trained on less components. This highlights the effects of over training because on a one or two component Gaussian, since the model is more generalized it does better on the evaluation set. What this essentially demonstrates is the fact that the more components you add to a Gaussian mixture model, the more specified to the training dataset it becomes, and the less applicable it is to a slightly different dataset.

Algorithm	Data	Train	Eval
IMLD - QDA	Set No. 13	11.39	20.49
Python - QDA	Set No. 13	11.12	20.83
GMM (N = 01)	Set No. 13	11.40	19.91
GMM (N = 02)	Set No. 13	11.29	18.89
GMM (N = 03)	Set No. 13	11.11	19.88
GMM (N = 04)	Set No. 13	11.08	20.19
GMM (N = 05)	Set No. 13	10.98	20.17
GMM (N = 06)	Set No. 13	11.12	19.76
GMM (N = 07)	Set No. 13	11.09	19.98
GMM (N = 08)	Set No. 13	11.08	19.80
GMM (N = 09)	Set No. 13	11.02	20.36
GMM (N = 10)	Set No. 13	10.91	20.11

H. SUMMARY

Overall, we can see that Gaussian mixture models can be fit very accurately to data by adding more and more components. However, when they are overfitted to the data their ability to generalize other datasets lessens. In application, this means that on a single dataset, the more components the better, but if you are trying to use it as a algorithm for multiple datasets, it's important to get the balance of the components correct. This helps improve the model's ability to generalize data while maintaing accuracy for a single dataset.