Gaussian Mixtures

Muhammad Subhan Khan
Department of Electronic Engineering
Hochshule Hamm Lippstadt
Lippstadt, Germany
muhammad-subhan.khan@stud.hshl.de

***Abstract*—When talking about Gaussian mixtures, we can say that it is a type of machine learning algorithm. With the help of Probability, distribution data is classified into different categories. These mixtures are implemented in a lot of areas such as Finance, marketing, image processing, etc. Using Gaussian mixtures we are able to identify and find the underlying groups of data. In general, we assume the Gaussian distributions of each group by looking at the means and covariances of these groups and then define their parameters. This paper will discuss and go into detail with the real-world examples where Gaussian mixtures are used, as well as the key steps of using Gaussian mixture models. Moreover, this paper will also focus on how discovering clusters in data sets have been made easier by using Gaussian mixtures. This work will also reveal how well Gaussian mixtures work with non-linear distributions.**

***Index Terms*—Probability, covariances,**

I. Introduction

The information in the chapter is based on [1].This type of model is a probabilistic model in which the all the given data points are created from just a finite number of Gaussian distributions with unknown parameters. Clustering can be achieved and solved by using Gaussian mixtures, this is when we group a set of data, which show extremely similar characteristics and are identical are brought together. Gaussian mixtures also tend to be very resistant when talking about outliers points that are different in nature, meaning if at the end some data points are not present in the clusters, the results will still yield very high accuracy. Therefore this is a very useful and a reliable method to cluster data, and an efficient method as well when saving time.[1]

In the Gaussian Mixture Model each cluster is treated as a separate Gaussian distribution. On the basis of probability it can identify which distribution was responsible for which data point.[1]

II. Gaussian Distributions

The function shown in figure 1 is known as the probability density function. Looking at this function in the graph, the area below the graph will show the interval in which the variable will fall in. The whole area under the graph will equal to the probability of a discrete random variable occuring.[2]

When looking at the equation in Figure 1 we can see that there are two parameters present. These 2 parameters we have the mean and standard deviation, this standard deviation can sometimes be squared and when this happens we have the variance. The width of the bell is controlled by the standard deviation, and due to this fact the width is directly proportional to the standard deviation which means if the standard deviation increases so does the width of the curve. Moreover, a higher standard deviation would also indicate that the data is more dispersed.[2]

$$y = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

Fig. 1. probability density function [4]

As far as 1 dimensional space is concerned this equation fits.

This article [4] discusses how Gaussian Mixtures can also be perceived as a Gaussian distributions which are basically probability distributions. And by the use of this method the overall population that is being dealt with can be arranged into clusters. This distribution can be used in data science as well. Since the majority of the given points that are being dealt with are quite near to the value of mean which is where the majority of the data is sort of centered around, the Gaussian distribution is going to be even, specifically when its close to the mean value.The centre of the curve that we are dealing with which is the bell curve, the centre of the curve consists of the majority of the data points. A component of the curve which is known as the probability density function can be utilized to determine the likelihood of that specific point belonging to the curve.[4]

III. What is a Gaussian Mixture Model

The information in this chapter is based on [4].Since we are dealing with unsupervised data, problems emerge since the data points that have to be worked with are widely dispersed which makes it a challenge. This eventually leads to

$$N(\mu, \sum) = \frac{1}{(2\pi)^{\frac{d}{2}}\sqrt{|\sum|}} exp(-\frac{1}{2}(x-\mu)^T \sum^{-1}(x-\mu))$$

Fig. 2. probability density function of GMM [4]

complications when clusters have to be formed and managed from this unsupervised data. Moreover in the Gaussian distribution , attempt to fit various points under the bell curve is done and experimented with and the result of that which is a peak value which we name as the mean of all the data points that are situated under the curve.The points found beneath the curve are strongly interconnected to one another forming a cluster of data points.The data set which follows this distribution will be exceptionally well maintained because of the distribution have a continuous nature. Machine learning can benefit from such data.[4]

A lot of times in the Gaussian distribution, peaks are produced, and the reason for this to happen is due to the fact that data that is being handled or dealt with comprises of a number of groups that are similar to Gaussian distribution. Filtering and taking out the data from these groups is achieved by Gaussian Mixtures. Moreover, Gaussian Mixtures will consist of more than one number of Gaussian distributions and when this happens we use the probability density function displayed in figure 2. A Gaussian distribution is sufficient when only one peak in the data is present, however in the presence of multiple peaks in the data which will usually be the case then we will get more than one number of peaks and therefore when more number of peaks are present then a mixture of distributions will be Gaussian Mixtures, hence the word mixtures is used. In figure 2 on the left hand side of the equation we have a scaling term which is there to make sure that integral will always equal to 1. In figure 2 also a covariance variable is introduced which allows to rotate the Gaussian distributions in different directions, and also the exponential exp term squared gives us the shape of the bell curve.[4]

IV. Clustering in Gaussian Mixtures

In order to use these models it is very important to decide on a covariance matrix that describes the relationship between the Gaussians. It is important that the Gaussians should be alike as this would result in closer means. Moreover, if one has to find out how many clusters there are, this is achieved by identifying the number of Gaussians in each group.[1]

A. Covariance

Covariance is an imporant aspect especially in Gaussian Mixtures, however it does not display the effect each variable has on the other however it does show that how the 2 variables progress and change together.[3]

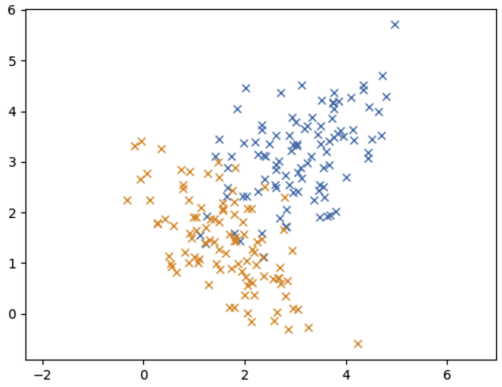

Fig. 3. Example of data for clustering[2]

B. Expectation and Maximization(EM) in Gaussian Mixture Models

The information in this subsection is based on [1].Approximating the parameters in the GMM model is an important process and to assist with this process The expectation-maximization method is utilized and frequently employed. When looking at GMM each component is represented by a Gaussian parameter, for us to locate and recognize these parameters, expectation can be made use of. On the other hand, Maximization is responsible for identification of the data points,for instance those new data points which are fit will be included otherwise they will be excluded.[1]

The use of the EM method is extremely useful because when dealing with unlabeled data it can be extremely difficult to identify where the points came from. And by the use of the EM algorithm this problem can be solved. The EM method is one where it operates in steps in this case 2 steps, and these two steps are the expectation and maximization terms which are mentioned above. And the process goes back and forth between the expectation and maximization until the process finishes. The EM approach changes between completing an expectation step and the maximization step. The goal of the process is to begin by initializing the parameters, this is to start the process and when this is done then going back and forth it can improve these approximations. And the maximization step takes care of reconditioning of the mean values. The alteration of the data points keeps happening in order to locate the data until we reach convergence. Moreover, the expectation part that is mentioned gives a probability of using the present parameter approximations and then expand these for the generation of a new updated Gaussian[1]

C. Clustering when data is overlapping

The information in this subsection is primarily based on [2]. As shown in Figure 3, there is the overlapping of data, for instance if data was not overlapping then K-means clustering could have easily been implemented here however as this

$$p(x) = \sum_{j=1}^{k} \phi_j \mathcal{N}(x; \mu_j, \Sigma_j)$$

$$\sum_{j=1}^{k} \phi_j = 1$$

Fig. 4. Generative modeling equation [2]

is not the case Gaussian Mixtures is the appropiate method to use here. And The data above also seems to follow the normal distribution and this is on the basis of the central Limit theorem which states that when enough random samples are present from any given distribution it will resemble that of Normal distribution.Moreover when K-means is used, the covariance of the data present is never taken into account, as shown in figure 3 there is a link between X and Y among the blue dots, for example the larger the value of X will be here it will yield a larger value of Y. Further more for instance if K means clustering was used here, and 2 points in figure 2 were at an equal distance from the centre of the cluster, however one of the points was following the trend, then for K-means both would be the same, but it appears that the point that will follow the trend would be a far more better fit than the point that doesn't seem to fit the trend.[2]

Knowing that the Data is Gaussian in figure 4, we can conclude the fact that the data is Gaussian. And with keeping this fact in mind each point produced will be a mixture of Gaussians. And in order to compute the probability of that we can show this using the equation in figure 3. The first equation in figure 4 describes that x comes from the linear combination of K Gaussians.Each Gaussian will be given ϕ_j, which reflects the strength of the Gaussian. The second equation shown in figure 4 is a weight constraint and all should be equal to 1.[2]

V. How GMM differs from other Models

The information in this chapter is primarily based on [1] which compares the Gaussian Mixture Model primarily with K-means clustering. The Gaussian Mixture Model, a clustering approach is used to create a data point from a mixture of Gaussian distributions that have unknown parameters. Both of these methods mentioned have the same unsupervised approach. The ultimate purpose is to calculate the distribution's parameters and also calculate the fraction of data points that are coming from each distribution. In comparison to this model we could take the example of the K-clustering model where there is no assumption made about the underlying distribution of the given data points. And due to this fact, this model simply divides the data points into K-clusters where each data point is defined by its centre.[1]

In cases where time is of the essence, K-means clustering can be utilized as the Gaussian mixture models can be a bit difficult to train when dealing with a crucial factor such as time, and as K means is quicker in converging it can be used here .[1]

The size of the data sets can help identify which method would be the best to use. For instance, K-clustering is beneficial when dealing with a large set of data, whereas Gaussian Mixture Model can be used when the data set is smaller and the clusters are not very well separated.[1]

Differentiating both the methods by looking at their flexibility can also help us decide which method to use, and as Gaussian mixture model is workable around a wide variety of clusters it is deemed more flexible in comparison to the K means clustering, which can only occupy certain shapes such as those that have a spherical nature .[1]

Looking at both methods K-means clustering is by far the faster method due to the lower number of iterations required in order to converge whereas a large number of iterations is required in the Gaussian Mixture Model in order to end up with a convergence.[1]

VI. Scenarios which use Gaussian Mixture Models

The information in this chapter is primarily based on [1].In cases where data is produced by a combination of Gaussian distributions, there are always situations where the amount of clusters that are present are not known, or the clusters can consist of a variety of shapes, and when this is the case. The important role of aiding in the accuracy of our final solutions or results can be done by Gaussian Mixtures. This model also greatly lessens the likelihood of mistakes as it assists in identifying patterns and trends in the data, these patterns are created by Gaussian distributions.[1]

Due to its effectiveness and reliability, it is often used in places to pick out clusters in the data which might not have been easy to find. Some examples can include, suspicious activities and the clustering of images. Suspicious or fraud activities could be easily identified with the help of this method by detecting anomalies, a data point which is quite different from a large set of data points, and by identifying these points, this activity can be detected.[1]

VII. Application of Gaussian Mixtures

Gaussian Mixture models have been shown to be very useful when it comes to application, especially when dealing with large data sets is required, and as the data set gets larger and can be difficult to find clusters and Gaussian Mixture Models are the perfect solution for this.[1]

A. Medical Datasets

In the area of medicine the GMM has been proven to be quite handy as this model has the ability to categorize photos

based on the content in the photos given, as well as find out trends in the medical data set. When trends are found in medical sets this leads to advancement in diagnosing diseases as a cluster of patients might have similar symptoms to one specific disease.[1]

B. Predicting Stock Prices

Utilizing GMM in Finance has helped grow and develop the stock market by making it more advanced. It has become more advanced in a way, that by uncovering tough to see changes in stock prices that are usually difficult to spot.[1]

VIII. Initialization Methods

This following information in this section is based on [5], we can use 4 initialization methods in order to produce an initial centre.

K-means default: This method is quite computationally costly and uses a classic K-means clustering technique in order to initialize.[5]

k-means++:k-means++ approach for clustering is applied here in this method and will select the first center from the data at random. Using the weighted distribution of data following centres will be picked. Points which are at a distance from the weighted distribution will be favoured. Since this is the default initialization of K means it will be quicker, however this can change with the fact when dealing with larger data sets consisting of a number of components.[5]

randomfromdata: Here the starting centres will be chosen at random from the data that we are supplied with. Initialization that is taking place here is relatively quick. Possibility of Non-convergent results is quite high if the the points that are selected are too close together.[5]

random: This is quite a straightforward method because the centre is chosen a bit differently from the mean of the data, this can cause the model to take a longer time to converge.[5]

IX. How the Gaussian Mixture Model operates

The information in this chapter is based on [7]. The Gaussian Mixture model chooses a random parameter. When this initial parameter is chosen, the Gaussian Mixture algorithm will perform a series of computations to determine the optimal weights as well as the means in order to converge with the Gaussian distributions which were estimated in order to join the ones with the initial parameters. This is primarily done in order to identify in which distribution each data point will be found in.[7]

init_params is a hyperparameter that gives one the opportunity to select between random points or if not that use the default method which is k means. The one advantage of using K-means would be that convergence of data will be very quick.[7]

Another parameter worth mentioning is the covariance_parameter which allows the user to have to choose 4 type of shapes.[7]

Full: Indicates that the components have the option to take whatever position or form that suits them the best on their own.[7]

Tied: The form will be shared, however it might be of any shape[7]

Diagonal:Contour axes are in line with the coordinate axes[7]

Spherical: Will consist of circular contours[7]

X. Advantages of Gaussian Mixture Models

The information in this section is Primarily based on [1]. As discussed in [1], GMM use soft clustering, unlike K means that uses hard clustering the advantage is that data points are not assigned to the nearest cluster they will assess all the data points and give each data point a probability of it belonging to each cluster since the data point could be 80 percent related to one cluster and 20 percent related to another and therefore assigning the probability to each will yield more accurate results. It also tends to be very robust to any data that might not fit entirely into the clusters or might be different from the rest of the data. And due to its flexibility of acquiring other shapes rather than being strictly spherical makes it a useful method.[1]

XI. Disadvantages of Gaussian Mixture Models

As discussed in [1]. Gaussian Mixture Models use EM algorithms for clustering the data and although it is a very useful and effective method it can be a slow method as a lot of iterations have to be done in order to have a convergence, and this can result in alot of time being consumed so when time is limited this method might not be the best to use. Moreover [8] discusses the disadvantages of unsupervised learning methods, and since GMM is an unsupervised learning method can be a problem when accuracy of the results is needed, since there can always be an uncertainty about the results that are outputted.the article [8] also compares the supervised learning with unsupervised learning, and when compared researchers have to spend a lot more time trying to label the data as they would in supervised learning which is also a drawback. Due to clustering in unsupervised learning unique results such as understanding individual customer needs could be overlooked when dealing the area of customer segmentation.[1][8]

XII. Unsupervised Learning with Gaussian Mixtures

The information in this chapter is based on [6]. The Gaussian Mixture model is one of the types of density estimate

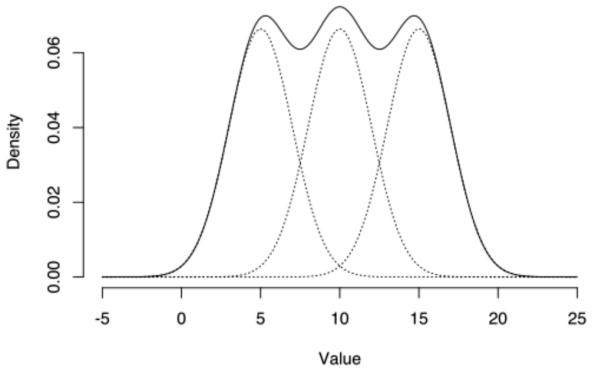

Fig. 5. Multimodal Graph [6]

$$p(x) = \pi_1 N(\mu_1, \Sigma_1) + \pi_2 N(\mu_2, \Sigma_2) + \pi_3 N(\mu_3, \Sigma_3) + \ldots$$

π_k = probability that x belongs to the kth Gaussian

Fig. 6. Sum of weighted Gaussian Distribution [6]

that has the ability to supply us with an approximation our data's probability density. Usually when the data is multi modal, the best way to approach it is with Gaussian Mixtures.[6]

Figure 5 displays a probability density graph, and due to 3 initial Gaussian distributions being present 3 bumps are produced in the graph. Looking at Figure 6, The distribution shown depends on the weight each one of Gaussian. The total sum of all the pi values will result in one.[6]

XIII. Conclusion

In conclusion and in the field of data science the form of machine learning technique discussed in the paper is Gaussian mixture models. They are used in a variety of situation however most commonly they are utilized in situations where the data sizes are large and due to data sizes being large it makes it difficult to locate clusters. Gaussian Mixtures makes it quite easy to name or label clusters by offering probability estimates, as opposed to K-means clustering which require more effort. When trying to find the most effective way to find patterns in data sets, Gaussian Mixtures methods is the by far the best method to choose.

XIV. Declaration of Originality

I, Muhammad Subhan Khan, herewith declare that I have composed the present paper and work by myself and without the use of any other than the cited sources and aids. Sentences or parts of sentences quoted literally are marked as such; other references with regard to the statement and scope are indicated by full details of the publications concerned. The paper and work in the same or similar form have not been submitted to any examination body and have not been published. This paper was not yet, even in part, used in another examination or as a course performance. I agree that my work may be checked by a plagiarism checker.

02.07.2022&Hamm - Muhammad Subhan Khan

Acknowledgment

XV. References

[1] Kumar, A. (2022, April 14). Gaussian mixture models: What are they amp; when to use? Data Analytics. Retrieved July 2, 2022, from https://vitalflux.com/gaussian-mixture-models-what-are-they-when-to-use/

[2] Deshpande, M. (2022, April 29). Clustering with Gaussian Mixture Models. Python Machine Learning. Retrieved May 24, 2022, from https://pythonmachinelearning.pro/clustering-with-gaussian-mixture-models/

[3] By Great Learning Team -. (2022, April 1). Covariance vs Correlation — Difference between correlation and covariance. GreatLearning Blog: Free Resources What Matters to Shape Your Career! Retrieved May 24, 2022, from https://www.mygreatlearning.com/blog/covariance-vs-correlation/

[4] Verma, Y. (2021, October 25). All You Need to Know About Gaussian Mixture Models. Analytics India Magazine. Retrieved May 24, 2022, from https://analyticsindiamag.com/all-you-need-to-know-about-gaussian-mixture-models/

[5] 2.1. gaussian mixture models. scikit. (n.d.). Retrieved July 2, 2022, from https://scikit-learn.org/stable/modules/mixture.html

[6] Alagiyawanna, I. (2021, December 14). Unsupervised Machine Learning with Gaussian Mixture Models. Medium. Retrieved May 25, 2022, from https://medium.com/@isurualagiyawanna/unsupervised-machine-learning-with-gaussian-mixture-models-ce2993e7061c

[7] Santos, G. (2022, March 11). A Simple Introduction to Gaussian Mixture Model (GMM). Medium. Retrieved May 25, 2022, from https://towardsdatascience.com/a-simple-introduction-to-gaussian-mixture-model-gmm-f9fe501eef99

[8] Pratt, M. K. (2020, July 8). What is unsupervised learning? SearchEnterpriseAI. Retrieved June 12, 2022, from https://www.techtarget.com/searchenterpriseai/definition/unsupervised-learning