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 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 finding 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.(Abstract)**

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

I. Introduction

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 by using Gaussian mixtures, this is where a set of data points can be grouped into clusters. Gaussian mixtures also tend to be very resistant when talking about outliers, meaning if at the end some data points are not present in the clusters, the results will still be very accurate. Therefore this is a very useful method to cluster data, and an efficient method as well when saving time.

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.

II. Update of the Gaussian mixture model

III. Simplification of the Gaussian mixture model

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

The expectation-maximization method is a very powerful that is commonly used in order to estimate the parameters of the Gaussian Mixture Model. Expectation can be named E and maximization can be named M. When looking at GMM each component is represented by a Gaussian parameter and in order to find and identify these parameters we use Expectation.On the other hand, Maximization is responsible for identifying which new points can be included in the model and which cannot.

$$\mathcal{N}(x;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Fig. 1. Normal Distribution Equation
[?]

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 algorithm starts by assuming the random components distributed around the origin, and then each point is assigned a probability of being generated by each component. After this the alteration of the parameters is done in order to maximize the chance of finding the data. This process which can also be called and iterative process keeps repeating itself until some sort of convergence is achieved.

V. 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.

A. Gaussian Distributions

The function below is knows 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.

When looking at this equation we can see that there are two paramaters 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.

B. Problems associated with clustering

Clustering can take alot of time as large number of data is dealt with along with a number of dimensions and this could sometimes create problems due to time complexity.

Results that are produced can be interepreted in alot of ways

VI. How GMM differs from other Models

In the Gaussian Mixture Model, a clustering approach is used to create a data point from a mixture of Gaussian distributions that have unknown parameters. 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 centroid.

In cases where time is of the essence, K-means clustering can be used as these models have faster converging time, and faster converging time directly impacts the run time.

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.

Gaussian Mixture Model are also known to be more flexible in general as they can take different type of cluster shapes whereas the K-clustering can only have spherical clusters.

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.[4]

VII. Scenarios which use Gaussian Mixture Models

The Gaussian Mixture Model is quite useful in scenarios where there us ambiguity about the right number of clusters present as well as the varieties of shapes that these clusters come in when data is generated by a combination of Gaussian distributions, therefore in such cases Gaussian Mixtures can be fully utilized as they help with the accuracy of the results. Moreover, when there is a doubt about the number of clusters present the chances of errors increase, and Gaussian Mixtures can prevent that from occuring. In addition to this usually when data is produced by a mix of Gaussian distributions it can get quite difficult to identify patterns and therefore this model can assist in finding those underlying patters in the data set.

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.

VIII. 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.

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, such as cancer.

B. Predicting Stock Prices

Utilizing GMM in Finance has helped develop the stock market by making it more advanced. It has become more advanced in a way that by uncovering turning points in stock prices that are usually difficult to spot.

IX. Explaining how the Gaussian Mixture Model works

Acknowledgment

References

References

1 Sarker, I. H. (2021, March 22). Machine learning: Algorithms, real-world applications and research directions - SN computer science. SpringerLink. Retrieved April 6, 2022, from https://link.springer.com/article/10.1007/s42979-021-00592-xSec15

2 Douglasnbsp;Reynolds, amp; Douglasnbsp;Reynolds1. (1970, January 1). Gaussian mixture models. SpringerLink. Retrieved April 7, 2022, from https://link.springer.com/referenceworkentry/10.1007/978-0-387-73003-5196

3 Kubara, K. (2020, October 8). Gaussian mixture models vs K-means. which one to choose? Medium. Retrieved May 8, 2022, from https://towardsdatascience.com/gaussian-mixture-models-vs-k-means-which-one-to-choose-62f2736025f0

4 Kumar, A. (2022, April 14). Gaussian mixture models: What are they amp; when to use? Data Analytics. Retrieved May 12, 2022, from https://vitalflux.com/gaussian-mixture-models-what-are-they-when-to-use/What$_i s_t he_e xpectation - maximization_E M_m ethod_i n_r elation_t o_G MM$