Gaussian Mixtures

Abdullah Zafar
DEEP LEARNING 1
Hochschule Hamm-Lippstadt
Lippstadt, Germany
abdullah.zafar@stud.hshl.de

***Abstract*—This paper focuses on Gaussian Mixtures, which is an unsupervised learning method. Mathematical basics including the set of parameters for the algorithm are also mentioned in a good detail. Furthermore, Estimation algorithms such as Expectation-Maximization and Variational Inference are also presented and their importance is highlighted as well. K means is another unsupervised clustering method that is commonly used. Drawbacks of K means Clustering furthermore emphasis on the importance of Gaussian Mixture Models and how they fulfill the void, by implementing soft clustering. Gaussian mixture can also be used for anomaly detection.**

Mathematical equations and graphical images are used repeatedly throughout the paper for better explanation and understanding of the topic, this is largely due to the fact that Gaussian Mixtures rely heavily on graphical representation. Lastly, an implemented example of Gaussian Mixtures is presented using python framework.

I. Introduction

Machine Learning includes two methods of learning, supervised and unsupervised. The main distinguishing factor between the both includes the method it uses to deal with data. Clustering is a unsupervised method where clusters area is made around the data set that share the same characteristics. Furthermore,Clustering could be regarded as grouping of similar data points together, based on the features and the attributes that they share. K-means is a popular method of clustering but a major drawback associated to it, is that it does not specify on how much a data point is associated with a specific cluster. This is commonly known as hard Clustering. Gaussian Mixture Models empowers us to implement soft clustering and get a better representation of the data available.

Gaussian Mixture Models presume that a specific number of Gaussian distributions exist and each one them represents a cluster. As a result, the data points belonging to a single distribution tend to be grouped together in a Gaussian Mixture Model.

A Gaussian mixture model is a probabilistic model that assumes all the given data points are generated from a mixture of a finite number of Gaussian distributions with unknown parameters and tends to group the given points to a single distribution that it belongs to.All the instances that are produced from the same Gaussian Distribution form a cluster that usually resembles an ellipsoid. Clusters vary in many different aspects from one another, few of those variations include shape, density, orientation, and size. [1]

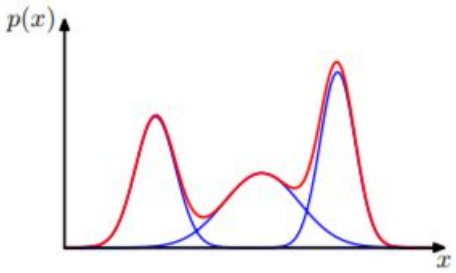


Fig. 1. Gaussian mixture distribution with three Gaussians represented in red and their sum in Blue [1]

The Gaussian mixtures implement the expectation-maximization algorithm for fitting mixture-of-Gaussian models. Gaussian Mixtures further compute Bayesian Information Criterion to asses the number of clusters in a given data. Gaussian Mixture fit method enables the learning of a Gaussian Mixture Model from a given train data. Each sample is then assigned to a Gaussian using Gaussian Mixture predict method. [2]

II. Gaussian Mixtures mathematical basics

Inability of Gaussian Distribution to model real data sets overshadow its important analytical properties. Super position of several Gaussian always gives a better representation of the data available. Gaussian mixture models enables us to form super positions and can be formulated as probabilistic models known as mixture distributions.

By utilizing an adequate number of Gaussians, and by changing their methods and covariances just as the coefficients in the linear combination, practically any continuous density can be approximated to arbitrary precision. We in this manner consider a superposition of K Gaussian densities of the form, which are known as Gaussian Mixtures.As mentioned earlier, Gaussian mixture model is basically a sum of K component Gaussian densities and it is formulated by the following equation: [1]

- μ that defines the center.
- covariance that is represented using $\sum$.
- A mixing probability π that defines how big or small the Gaussian function will be.

$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

Fig. 2. Equation 1 [1]

Each Gaussian density $N(x|\mu_k, \sum_k)$ is more commonly referred as the component of the mixture and it has its own covariance $\sum_k$ and mean μ_k [1].

The parameters π_k in Equation 1 are referred as mixing coefficients.If we integrate the two sides of Equation 1 with respect to x, we come to a conclusion that, both p(x) and the individual Gaussian components are normalized, hence we obtain [1]:

$$\sum_{k=1}^{K} \pi_k = 1.$$

Fig. 3. Equation 2 [1]

Fulfilling the conditions that p(x) $\geq$ 0, together with $N(x|\mu_k, \sum_k) \geq 0$, implies $\mu_k \geq 0$ for all k. Combing this with the condition that is represented in Equation 2 we obtain: $0 \leq \mu_k \leq 1$. [1]

It can be concluded that mixing coefficients satisfy the conditions to be probabilities. By applying the sum and product rules, the marginal density is given by:

$$p(\mathbf{x}) = \sum_{k=1}^{K} p(k)p(\mathbf{x}|k)$$

Fig. 4. Equation 3 [1]

which is equivalent to Equation 1 and it can be considered that $\mu_k = p(k)$ is the prior probability of picking the k^{th} component, and the density $N(x|\mu_k, \sum_k) = p(x|k)$ gives us the probability of x conditioned on k. [1]

III. GMM Graphical model interpretation

Figure 5.0 represents the structure of the conditional dependencies between random variables. Illustration 4.0 can be interpreted in the following manner [2]:

- Random variables are represented by circles.
- The squares address fixed qualities (i.e., boundaries of the model).
- The big square rectangles are called plates: they show that their substance is repeated a few times.
- Each plate's bottom right hand side has a number that denotes how many times its content is repeated.
- The categorical distribution with weights ϕ is used to generate each variable z^i. Each variable x^i is selected from a normal distribution, with its cluster z^i defining the mean and covariance matrix.
- Conditional dependencies are represented by arrows.
- The wavy arrow that points from z^i to x^i represents a switch. This means that depending on the value of z^i, the instance x^i will get sampled from a different Gaussian Distribution. [2]

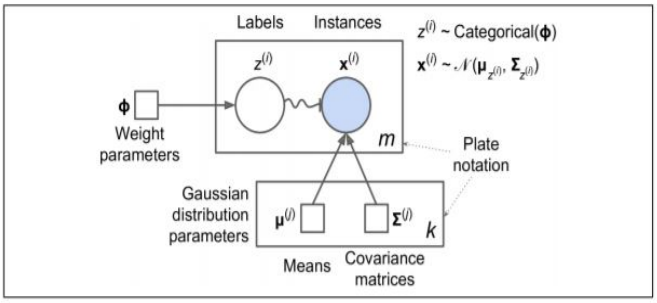


Fig. 5. A Gaussian mixture model [2]

IV. Anomaly detection using Gaussian mixture models

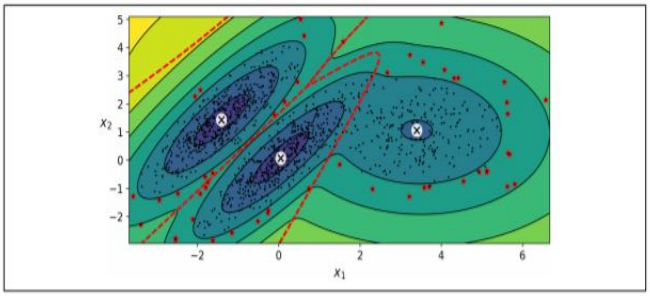


Fig. 6. Anomalies are represented as stars [2]

Anomaly Detection is the process of locating the instances that oppose the convention in a very strong manner. These instances are commonly referred as anomalies or outliers, and the remaining ones are by convention referred as inliers. Anomaly detection gets very handy when they are applied to real life scenarios. Few of the examples include, detecting fraud, detecting defective products in manufacturing or to obsolete dataset and feeding another model, which can improve the performance of the resulting models by many steps. Detecting anomalies in Gaussian mixture model is not a hefty task because any instance being located in a low density region can be referred as an anomaly. [2]
Novelty detection differs from anomaly detection in such a way that in novelty detection, the algorithm is assumed to be trained on a ''clean" dataset, free of outliers, whereas anomaly detection does not make that assumption.
Gaussian mixture models tries to accommodate all the available data, including outliers, if outliers are more in number,

the 'normality' gets effected and some outliers may be now considered as normal. If this happens, one might have to fit the model once, use it to detect and remove the most extreme outliers, then fit the model again on the cleaned up dataset. Another approach is to use robust covariance estimation methods. [2]

V. Bayesian Gaussian Mixture Models

Selecting the right number of clusters can be a long task but it is possible to use Bayesian Gaussian mixture class, which is in fact capable of giving weights equal or close to zero to the clusters those are unnecessary. Simply set the number of clusters n components to a value that you have good cause to believe is more than the optimal number of clusters, and the algorithm will automatically remove the extraneous clusters [2].

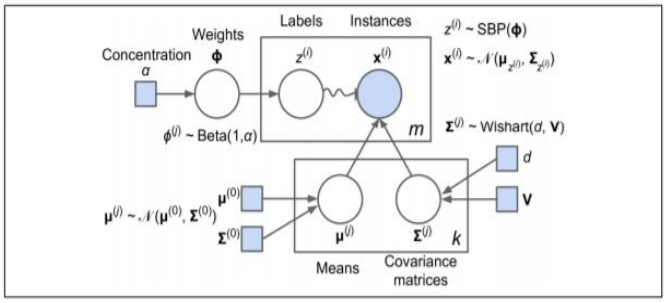


Fig. 7. Bayesian Gaussian mixture model [2]

In figure 7.0,Bayesian Gaussian mixture model, the cluster parameters (including the weights, means and covariance matrices) are not treated as fixed model parameters anymore, but as latent random variables.

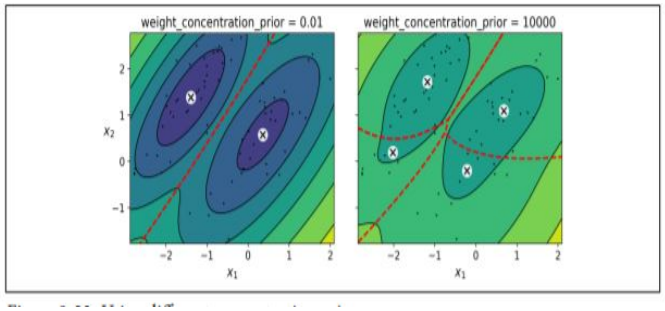


Fig. 8. Using different values of $p(z)$ [2]

Prior information of the latent variables z can be represented in the prior, which is a probability distribution p(z). For instance, We might be having an educated guess that the clusters are likely to be many(high concentration), or few(low concentration). This can be easily adjusted with the help of weightconcentrationprior hyperparameter. Figure 7 shows the results of setting the values to 0.01 and 1000 respectively, resulting into different clustering. [2]

VI. Bayes' theorem

Bayes's theorem can be used along estimated model paraments to estimate the posteriori component assignment probability. Learning the fact that an instance belongs to a specific distribution provides an important tool for learning clusters.Moreover, It assists in updating the probability distribution over the latent variables after data x has been observed. Bayes theorem computes the posterior distribution $P(z|x)$, which is the conditional probability of z given X. [2]

$$p(\mathbf{z}|\mathbf{X}) = \text{Posterior} = \frac{\text{Likelihood} \times \text{Prior}}{\text{Evidence}} = \frac{p(\mathbf{X}|\mathbf{z})\, p(\mathbf{z})}{p(\mathbf{X})}$$

Fig. 9. Bayes Theorem
[2]

VII. Estimation Algorithm : Expectation-Maximization

The main problem linked with learning Gaussian mixture models is that we are not aware of which latent component is the point coming from. Expectation-Maximization is a well developed statistical algorithm that allows us to estimate the parameters of the mixture model.
Expectation maximization is an iterative algorithmic method that tends to increase with each subsequent iteration. Generally two steps are involved in calculating Expectation maximization for Gaussian Mixture Models.

The first step involves the calculation of the expectation of the component assignments for each data point given the model parameter such as ϕ μ and σ.The first step is referred as E step. [1]

The second step is generally referred as maximization and it considers calculating the maximum expectation that was previously calculated in the first step with relation to the model parameters and this part consists of updating the values of ϕ μ and σ. As mentioned before, this method is an iterative method, so iterations take place until and unless the algorithm converges, hence giving the maximum likelihood estimate. Altering between the values that are known to be fixed or assumed, maximum likelihood successfully estimates the values those are non-fixed and they can hence be calculated in an efficient manner.

A. Steps involved in applying EM for Gaussian Mixtures

For the given Gaussian mixture model, the aim is maximizing the likelihood function with respect to its parameters . [1]

1) The initial step consists of initializing the parameters($\sum_k$,μ_k,π_k) and evaluating the initial value of the loglikelihood.
2) E Step. Utilizing current parameters for responsibilities evaluation.

$$\gamma(z_{nk}) = \frac{\pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}.$$

Fig. 10. [1]

3) M step. Finding Re-estimated values of the parameters using the current responsibilities.

$$\begin{aligned}
\boldsymbol{\mu}_k^{\text{new}} &= \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) \mathbf{x}_n \\
\boldsymbol{\Sigma}_k^{\text{new}} &= \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) \left(\mathbf{x}_n - \boldsymbol{\mu}_k^{\text{new}}\right) \left(\mathbf{x}_n - \boldsymbol{\mu}_k^{\text{new}}\right)^{\mathrm{T}} \\
\pi_k^{\text{new}} &= \frac{N_k}{N}
\end{aligned}$$

$$N_k = \sum_{n=1}^{N} \gamma(z_{nk}).$$

Fig. 11. [1]

4) The convergence of parameter or the log likelihood are to be checked, in case the convergence criterion is not fulfilled, step 2 needs to be returned to.

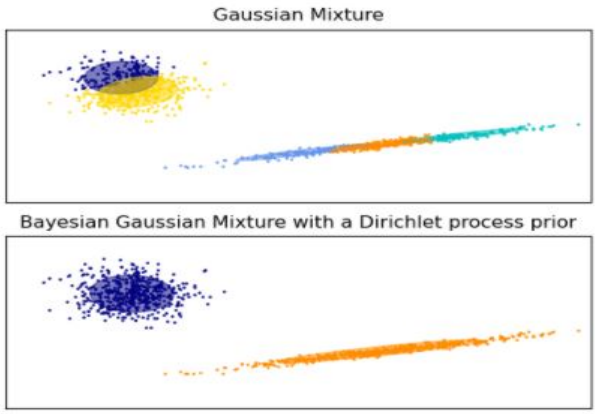
$$\ln p(\mathbf{X}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi}) = \sum_{n=1}^{N} \ln \left\{ \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right\}$$

Fig. 12. [1]

VIII. Variational Inference

Variational inference could be thought of as a continuation of expectation-maximization that tends to maximize a lower bound on model evidence rather than data likelihood. The theory behind expectation-maximization and variational inference is very similar,but variational methods tend to regularization by integrating information from prior distributions. This helps to avoid the singularities that are found in expectation-maximization solutions. Inference is however quite slow as compared to expectation maximization. [3]

Variational algorithm's Bayesian nature requires it to have more hyper-parameters as compared to expectation-maximization. weight_concentration_prior could be regarded as one of the most important concentration parameters. Setting a high concentration prior encourages a big number of components in the mixture to be active. On the other hand, setting a low concentration causes the model to place the majority of its weight on a small number of components, effectively reducing the weights of the remaining components to zero. [3]



[3]

The illustration above compares GMM with a specific number of components to the variational Gaussian mixture models with a Dirichlet process prior. It can be concluded from the given illustration that variational Gaussian mixture with a Dirichlet process prior is able to stay within only two components, whereas the Gaussian mixture fits the data involving a fixed number of components that must be set a priori by the user. The number of components in the given set was selected as 5, however it does not match the true generative distribution of the given dataset. It can easily be concluded that Gassuian mixture models with a Dirichlet process prior tends to fit only one component. [3]

IX. Relation to K-means

K-Means Clustering is a method that divides sample data into K clusters depending on the features. Clustering is accomplished by shortening the distance between the sample and the cluster's center. It is quite useful in terms of large number of samples and has a wide range of application areas which vary in many different fields.

Hard clustering is a more popular term for this. Lesser processing time, better for high-dimensional data, and it's a lot easier to install are just a few of the benefits connected with K means. Hard clustering has a number of drawbacks, including the possibility of erroneous data and the fact that, unlike Gaussian mixtures, it does not function well with complex geometrically formed data. [1]

K-Means clustering and Gaussian Mixtures have no logical relationship. The data will cluster into roughly spherical clumps centered on the means of each mixture component if it fits a spherical Gaussian mixture model effectively. However, you can use k-means clustering without making any assumptions about the data's origins. It can be used to split up data into useful and reasonably homogeneous bits in the same way that other clustering algorithms may, with

no presumption that those bits are genuine objects (eg, for market segmentation). You can prove what k-means estimates are without assuming mixed models.Gaussian mixture models can be fitted using maximum likelihood, which is a different estimator and methodology than k-means. Alternatively, Bayesian estimators and their accompanying methodologies can be used.Spherical Gaussian mixture models and k-means clustering are very comparable in several ways. They're not only distinct, but they're also different kinds of things.

X. Python Implementation(Gaussian Mixture Model Selection)

The BIC criterion can be used to select the number of components in a Gaussian Mixture in an efficient way. In theory, it recovers the true number of components only in the asymptotic regime (i.e. if much data is available and assuming that the data was actually generated i.i.d. from a mixture of Gaussian distribution). Note that using a Variations Bayesian Gaussian mixture avoids the specification of the number of components for a Gaussian mixture model.

This example shows that model selection can be performed with Gaussian Mixture Models using information-theoretic criteria (BIC). Model selection concerns both the covariance type and the number of components in the model. In that case, AIC also provides the right result, but BIC is better suited if the problem is to identify the right model. Unlike Bayesian procedures, such inferences are prior-free. [3]

In that case, the model with 2 components and full covariance (which corresponds to the true generative model) is selected.

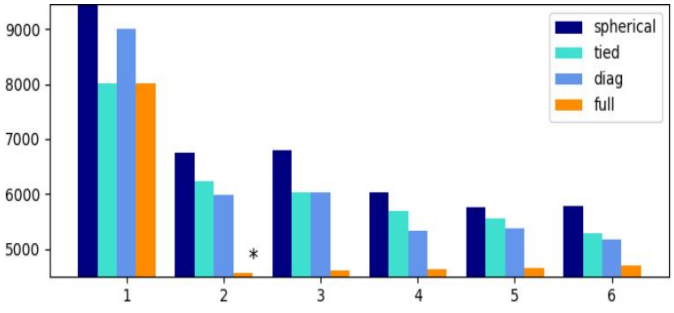


Fig. 13. BIC score per model [3]

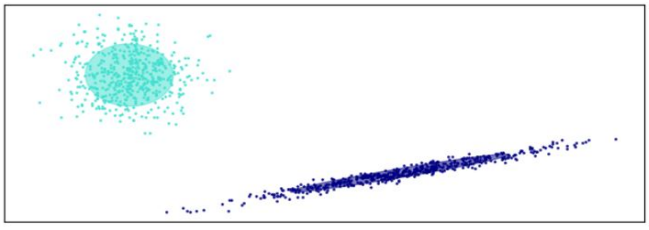


Fig. 14. Selected GMM:full model, 2 components [3]

The following is a brief description of the important sections of the code that was used to generate the illustration in figure 13 and figure 14. The code is extracted from [3]

```
import numpy as np
import itertools

from scipy import linalg
import matplotlib.pyplot as plt
import matplotlib as mpl

from sklearn import mixture
```

The above mentioned code [3] is used to import all the important libraries in order to implement this example. Numpy is fundamental package that helps in executing different computations in python. Itertools is a module that assists in solving complex interations in python. Plotting is done with the help of matplot.lib, it is library that makes plotting possible in python.

```
n_samples = 500


np.random.seed(0)
C = np.array([[0., -0.1], [1.7, .4]])
X = np.r_[np.dot(np.random.randn
(n_samples, 2), C),

 .7 * np.random.randn(n_samples, 2) +
 np.array([-6, 3])]
```

In the code [3] mentioned above, nsamples basically represents the number of samples per components, which in this example is declared as 500. np.random.seed(0) function is used to generate random samples, in our case, we are doing that for two components.

```
gmm = mixture.GaussianMixture
(n_components=n_components,
            covariance_type=cv_type)

gmm.fit(X)

bic.append(gmm.bic(X))
if bic[-1] < lowest_bic:
lowest_bic = bic[-1] best_gmm = gmm
```

Above mentioned part of code [3] helps in Fitting a Gaussian mixture with expectation maximization. fit[] is a method that helps in estimating the model parameters with the EM algorithm. bic[] is used as Bayesian information criterion for the current model on the input X. Bayesian information criterion is basically used to select a model among a finite number of models and the model with the lowest value of Bic is selected.

XI. Drawbacks of Gaussian Mixtures

Despite being one of the fastest learning mixture models out there, Gaussian mixtures have a few drawbacks associated to it. Estimating the covariance matrices becomes difficult in case there is an insufficient number of points present per mixture. In addition, Gaussian mixtures in these scenarios tend to deverge and find solutions that have infinite likelihood.This scenario can be avoided by artificially adjusting the covariances. [3]

XII. Applications of GMM

Gaussian Mixture Models being an effective clustering technique, is widely used in daily life applications. Listed below are a few examples:

- Fuzzy image segmentation
- Hand writing recognition
- Data stream clustering
- defect detection in products
- Modeling weather observations in Geo-science
- Speech recognition

XIII. Conclusion

Overcoming the drawbacks presented by K-Means, Gaussian Mixtures has proved to be a fast and effective method of clustering.Moreover, Gaussian Mixtures inability to identify which point comes from which latent component is over showed by the use of expectation maximization. Expectation maximization uses an iterative method to deduce the respective component of each point. The dominance of Gaussian mixtures can be estimated from the fact that it used in day to day applications such as fuzzy image segmentation, defect detection in products and speech recognition shows how effective.

References

[1] Christopher M Bishop. *Pattern recognition and machine learning*. springer, 2006.

[2] Aurélien Géron. *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems*. O'Reilly Media, 2019.

[3] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.