UNIVERSITAT DE BARCELONA

Final Project

FUNDAMENTAL PRINCIPLES OF DATA SCIENCE MASTER

Bayesian Statistics and Probabilistic Programming

Gaussian Mixture Models applied to Machine Learning

Project by Johnny Nuñez, Marcos Plaza, Qijun Jin

Professor: Josep Fortiana

Realized at: Department of Mathematics and Computer Science

Date: 27 June 2022

Table of Contents

1. Introduction; Motivation and work scope

Machine learning has become an essential topic in recent years. In fact, today there are more and more applications that rely on these methods. Therefore, at this point, we are probably wondering how Bayesian statistics is related to machine learning solutions. The first key point to note is that virtually all machine learning is based on the Bayesian notion of probability. Or put another way, machine learning always implicitly assumes that it can assign probabilities to events that are not repeatable just to express its degree of belief that this event is happening. In this project, we will try to focus on the methods of unsupervised learning. Also we will study one particular case by using *Gaussian Mixture Models* (you can find a notebook with the *GMM* use case in the folder called *code*).

To put it in context, in unsupervised machine learning, we are just provided with the input variable data without any labeled data, so its goal is to find some structure in the data without being explicitly given this thing that we are trying to predict. Therefore, unsupervised learning is closely related to clustering, where Bayesian methods help machine learning algorithms to extract crucial information from data sets so we could predict to which group future observations belong.

Gaussian Mixture Models (*GMMs*) are a type of machine learning algorithm. **They are used to classify data into different categories based on the probability distribution.**

2. Gaussian Mixture Model

Definition

Gaussian mixture models (GMM) are a probabilistic concept used to model real-world data sets. According to the ***Central Limit Theorem***, when there is a relatively large set of samples, whatever the distribution of the sample mean, it will follow approximately a normal distribution (also known as Gaussian Distributions). Thus, there is a wide variety of real-world data following this kind of distributions.

The Gaussian mixture model is a probabilistic model that assumes all the data points are generated from a mix of Gaussian distributions with unknown parameters (unsupervised learning). It can be used for clustering, which is the task of grouping a set of data points into clusters. *GMMs* can be used to find clusters in data sets where the clusters may not be clearly defined. Additionally, *GMMs* can be used to estimate the probability that a new data point belongs to each cluster. This method is also relatively robust to outliers, meaning that they can still yield accurate results even if there are some data points that do not fit neatly into any of the clusters. For this reason the *GMM*s are a flexible and powerful tool for clustering data. It can be understood as a probabilistic model where Gaussian distributions are assumed for each group and they have means and covariances which define it's parameters. GMM consists of two parts – mean vectors (μ) & covariance matrices (Σ). A Gaussian distribution is defined as a continuous probability distribution that takes on a bell-shaped curve, like the ones that appear in the *Figure 1*.

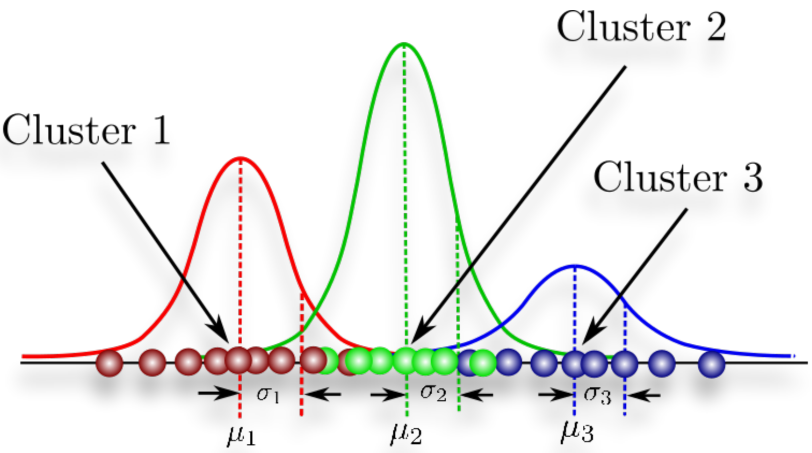


Figure 1. Graphical representation of Gaussian Mixture Models (in this case with 3 different clusters).

Practical applications

As you would expect, GMM has many applications, such as density estimation, clustering, and image segmentation. For density estimation, GMM can be used to estimate the probability density function of a set of data points. For clustering, GMM can be used to group together data points that come from the same Gaussian distribution. And for image segmentation, GMM can be used to partition an image into different regions.

Gaussian mixture models can be used for a variety of use cases, including identifying customer segments, detecting fraudulent activity, and clustering images. In each of these examples, the Gaussian mixture model can identify clusters in the data that may not be immediately obvious. As a result, **Gaussian mixture models are a powerful tool for data analysis and should be considered for any clustering task.**

3. Expectation-Maximization Algorithm

In Gaussian mixture models, an expectation-maximization method is an algorithm for estimating the parameters of a Gaussian mixture model (GMM). The expectation is termed E and maximization is termed M. Expectation is used to find the Gaussian parameters which are used to represent each component of gaussian mixture models. Maximization is involved in determining whether new data points can be added or not.

The expectation-maximization method is a two-step iterative algorithm that alternates between performing an expectation step, in which we compute expectations for each data point using current parameter estimates and then maximize these to produce a new gaussian, followed by a maximization step where we update our gaussian means based on the maximum likelihood estimate. The EM method works by first initializing the parameters of the GMM, then iteratively improving these estimates. At each iteration, the expectation step calculates the expectation of the log-likelihood function with respect to the current parameters. This expectation is then used to maximize the likelihood in the maximization step. The process is then repeated until convergence. Here is a picture representing the two-step iterative aspect of the algorithm:

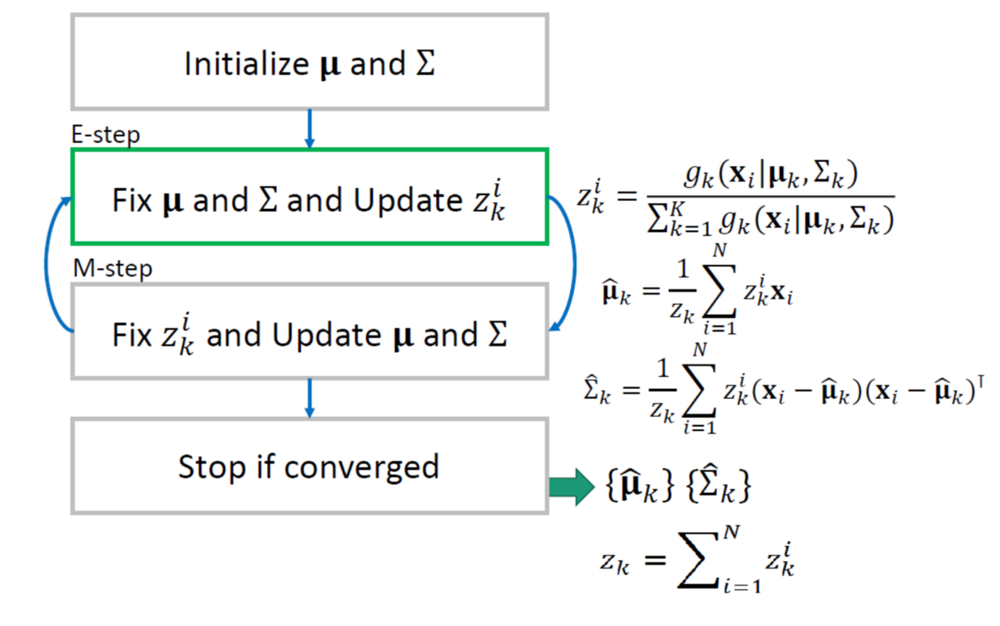

Initialize $\boldsymbol{\mu}$ and Σ
E-step
Fix $\boldsymbol{\mu}$ and Σ and Update z_k^i
M-step
Fix z_k^i and Update $\boldsymbol{\mu}$ and Σ
Stop if converged
$z_k^i = \frac{g_k(\mathbf{x}_i|\boldsymbol{\mu}_k, \Sigma_k)}{\sum_{k=1}^{K} g_k(\mathbf{x}_i|\boldsymbol{\mu}_k, \Sigma_k)}$
$\hat{\boldsymbol{\mu}}_k = \frac{1}{z_k}\sum_{i=1}^{N} z_k^i \mathbf{x}_i$
$\hat{\Sigma}_k = \frac{1}{z_k}\sum_{i=1}^{N} z_k^i (\mathbf{x}_i - \hat{\boldsymbol{\mu}}_k)(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_k)^{\mathrm{T}}$
$\{\hat{\boldsymbol{\mu}}_k\}\ \{\hat{\Sigma}_k\}$
$z_k = \sum_{i=1}^{N} z_k^i$

Figure 2. Graphical representation of Expectation-Maximization Algorithm.

4. Gaussian Mixture Model for clustering

The clustering done by the Gaussian Mixture Model method is a smoother clustering in the sense that it looks at what is the probability that a data point can belong to one cluster center compared to the other cluster. This makes us more flexible in the sense that each data point can be in any group but if we have calculated the probabilities of belonging to each group, we will assign it to the group in which it has the highest probability of belonging.

To calculate this probability, we not only define a single cluster center as a single data point, but we define a distribution for each of the clusters.

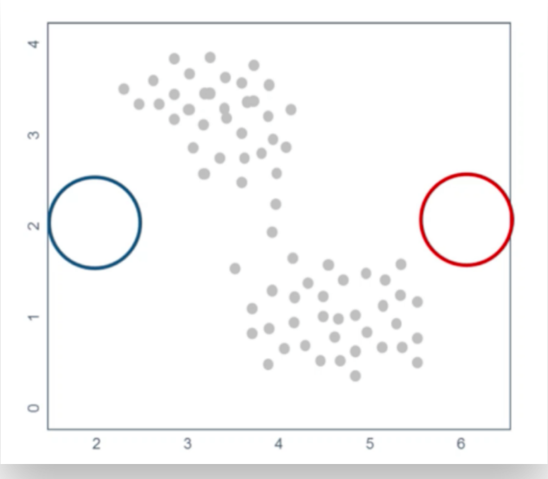
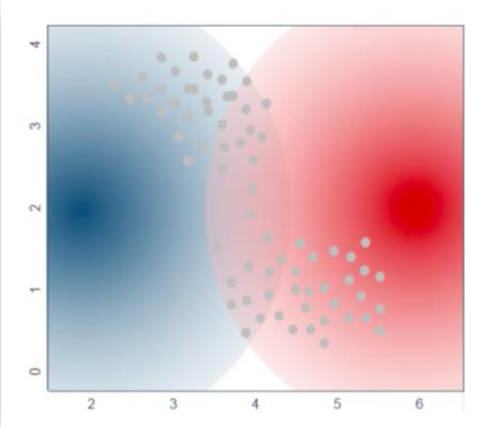


Figure 3. Graphical representation of the initial centers.

It starts with a random mean, where the center of this distribution will be, and a variance that indicates how the data should be dispersed in each group. In addition, we assign each group the a priori probability, where our best guess at the beginning, when we talk about two groups, would be to assign each group an a priori probability of 50%.

Figure 4. Graphical representation of the initial probabilities.



Once we have initialized means and random variances and the a priori probability of belonging to each group, we can calculate the probability of belonging to each group for all the data. As we have seen above this can be calculated using Bayes' Theorem:

$$P\left(\text{ data }_i \in \text{ group }_j \mid \text{ values }\right)$$

In this case, for each data i we would have calculated the probability of belonging to group 1 and the probability of belonging to group 2. Then it will recalculate the mean of the distributions, variances, and priors. And the new updated prior will be the average probability over all data points that a data point belongs to cluster 1 or 2. For the means, the updated value would be the weighted average value based on the probabilities of cluster 1 or 2, and a similar procedure is used to re-estimate the variance.

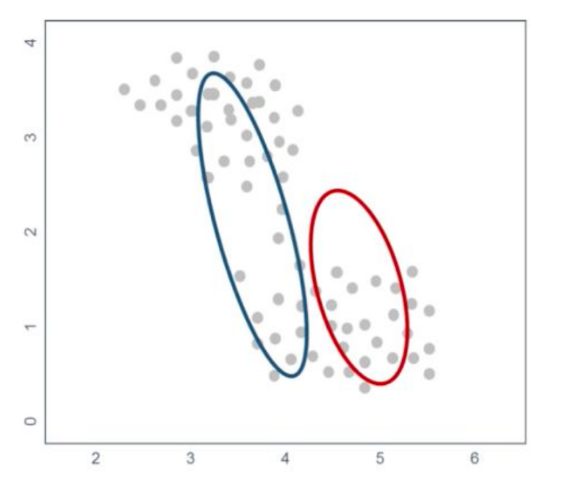


Figure 5. Recalculating centers.

As the distributions are re-estimated, we must perform the same operation on our data.

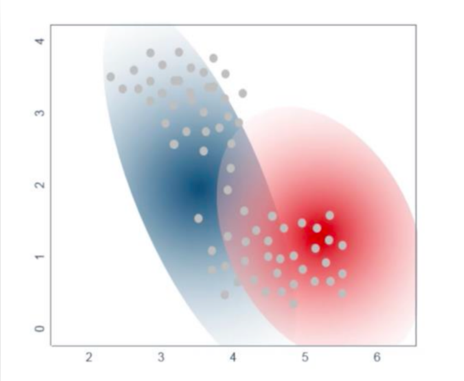


Figure 6. Recalculating probabilities.

The probability is recalculated for each piece of data that comes from each group and we update the means, variances, and priors and then move on to the next iteration. It will iterate the same procedure until it converges in the sense that the estimates that are the means, variances, and priors do not change significantly, over a long period of iterations.

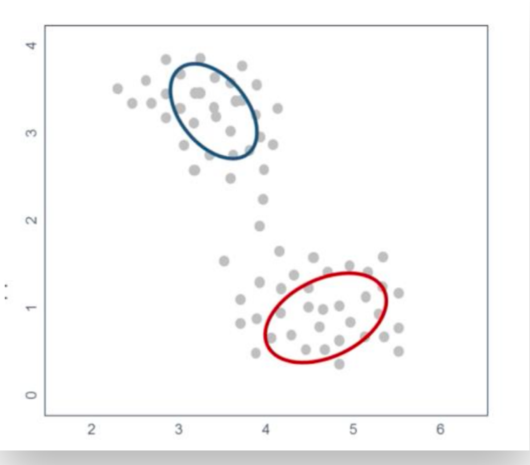
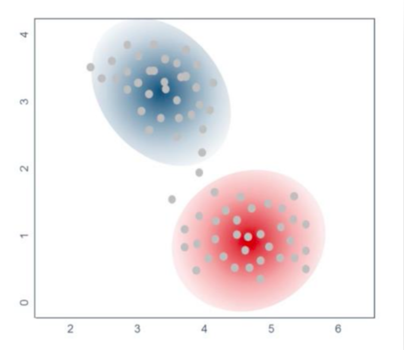


Figure 7. Centers on algorithm's convergence.

Once the convergence occurs, we can even draw the ellipses surrounding the data in each group, using the mean and variance from the last step.

The Gaussian Mixture Model applied for clustering works based on the Bayesian approach. In comparison with K-means, also a frequentist-based method for clustering, the difference is that K-means uses a hard assignment of each data point to a particular group while the GMM works with a soft assignment where this assignment is based on the probability of belonging of each data point to each cluster, given its values.

Figure 8. Probabilities on algorithm's convergence.



Moreover, we can also define the covariance matrix, which can control the degree of freedom in the shape of the clusters, as follows:

- *Diagonal* – The size of the cluster along each dimension can be set independently, with the resulting ellipse constrained to align with the axes.
- *Spherical* – It constrains the shape of the cluster such that all dimensions are equal. The resulting clustering will have similar aspects to that of k-means.
- *Tied* – All components share the same covariance matrix.
- *Full* – It allows each cluster to be modeled as an ellipse with arbitrary orientation.

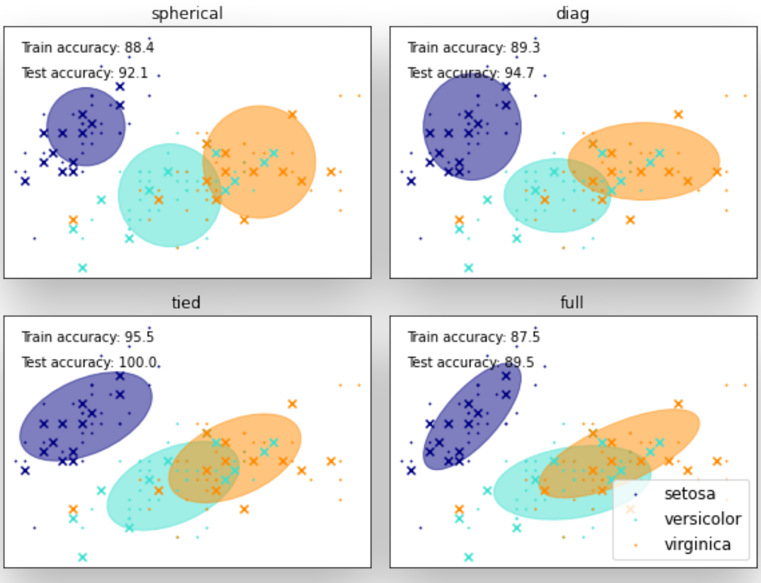


Figure 9. Variations on the cluster's shape by manipulating the covariance matrix.

As we can see in the plot, the training data is represented as dots, and the prediction data is represented as crosses. The accuracy of the prediction changes as the covariance matrices applied to the model is different. In the 2D plots, we can hardly figure out that there are several dots and crosses which are correctly classified in other subspaces. In the case of the full covariance matrix, which is the most flexible one to adjust the contour, does not obtain the best result as it is prone to overfitting and does not generalize well to test data.

5. Conclusions

After this brief but intensive study on this statistical clustering method, we can establish that it offers a number of advantages over its counterparts.

- In terms of cluster covariance, GMM is far more adaptable – Since each cluster's covariance along all dimensions approaches 0, k-means algorithm is actually a particular example of GMM. This suggests that a point will only be allotted to the cluster that is closest to it. Each cluster can have an unrestricted covariance structure when using GMM. Instead of the spherical distribution used in k-means, picture a rotated and/or elongated distribution of points in a cluster. As a result, compared to k-means, cluster assignment in GMM is significantly more flexible.

- The GMM paradigm allows for mixed membership – The fact that GMM allows for heterogeneous membership of points to clusters is another consequence of its covariance structure. While a point in a GMM belongs to each cluster to a varied extent, it only ever belongs to one cluster in k-means. The degree is determined by calculating the likelihood that a point will be created from the multivariate normal distribution of each cluster, with the cluster center serving as the distribution's mean and the cluster covariance as its covariance. Mixed membership may or may not be more appropriate for a task (news articles, for example, can belong to many topic clusters) (e.g. organisms can belong to only one species).

6. References

- https://jakevdp.github.io/PythonDataScienceHandbook/05.12-gaussian-mixtures.html
- https://brilliant.org/wiki/gaussian-mixture-model/
- https://medium.com/sfu-cspmp/distilling-gaussian-mixture-models-701fa9546d9
- https://rubialesalberto.medium.com/clustering-con-gaussian-mixture-model-en-sklearn-y-sus-parámetros-aplicado-al-marketing-74b9d8454b86
- https://towardsdatascience.com/gaussian-mixture-models-explained-6986aaf5a95
- https://towardsdatascience.com/a-simple-introduction-to-gaussian-mixture-model-gmm-f9fe501eef99
- https://economipedia.com/definiciones/teorema-central-del-limite.html