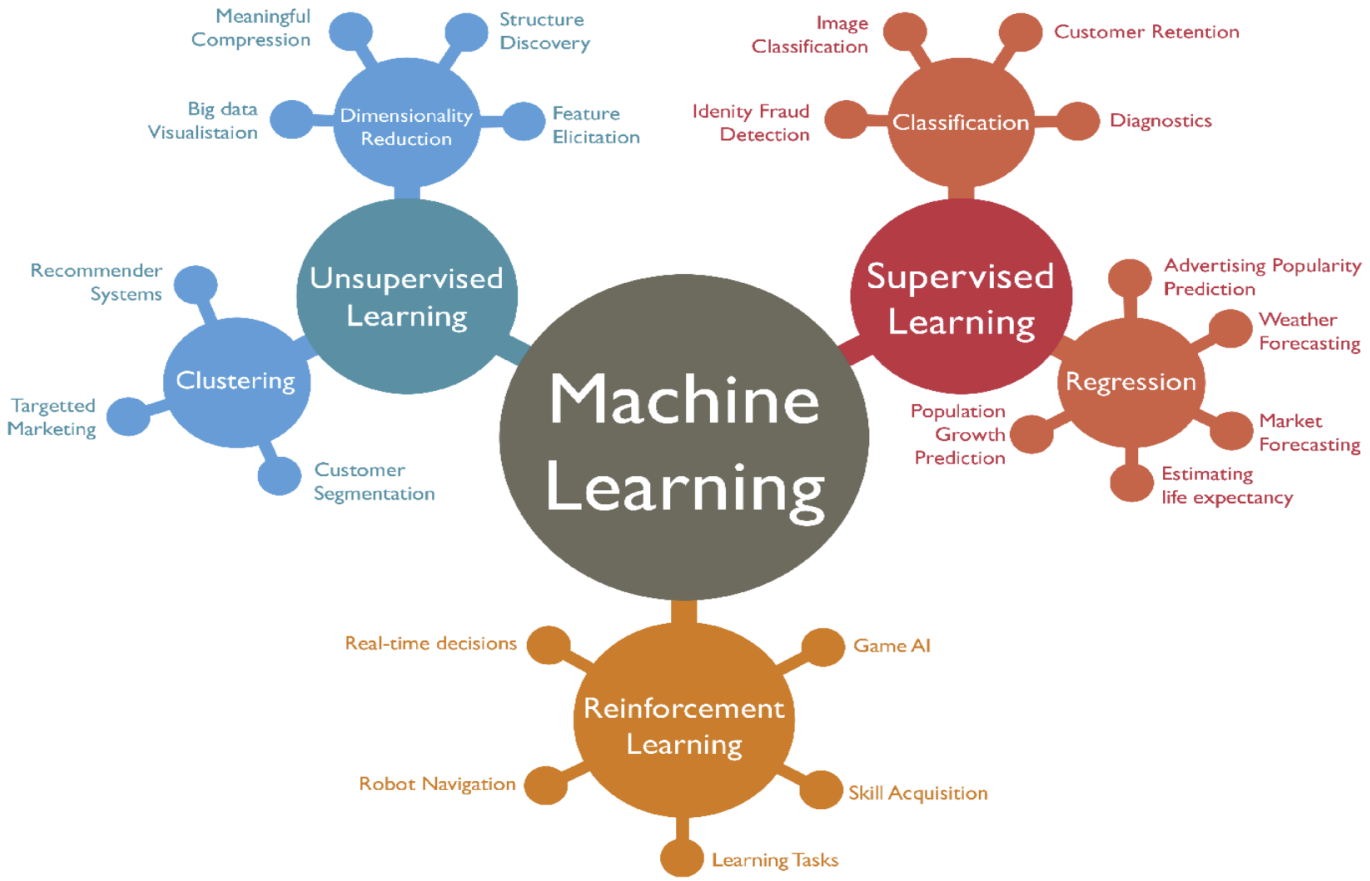
Gaussian Mixtures

By: Muhammad Subhan Khan



Machine Learning

Figure 1: Machine Learning [5]

Introduction to Gaussian Mixtures

Gaussian Mixtures is a Machine learning algorithm

With the help of this model, data can be classified into various categories

Data is divided into different categories on the basis of probability distribution

Gaussian Distributions

- Gaussian is basically a distribution
- Distribution is a listing of results in an experiment and the probability associated with each result produced
- Gaussian distribution is defined by the mean and the standard deviation
- Probability of a data point being part of the curve can be measured using the probability density function

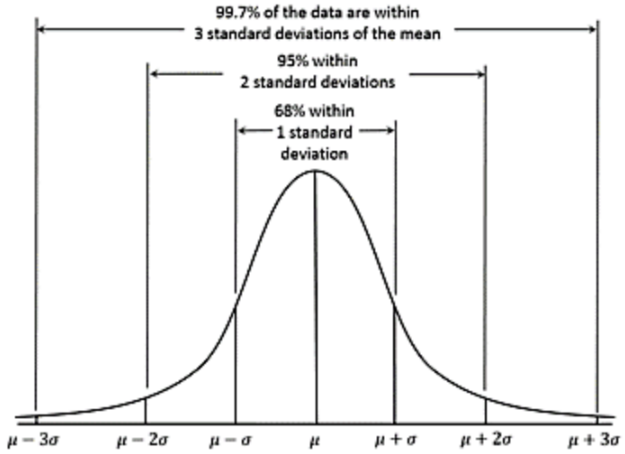


Figure 1: Example of a Gaussian distribution [4]

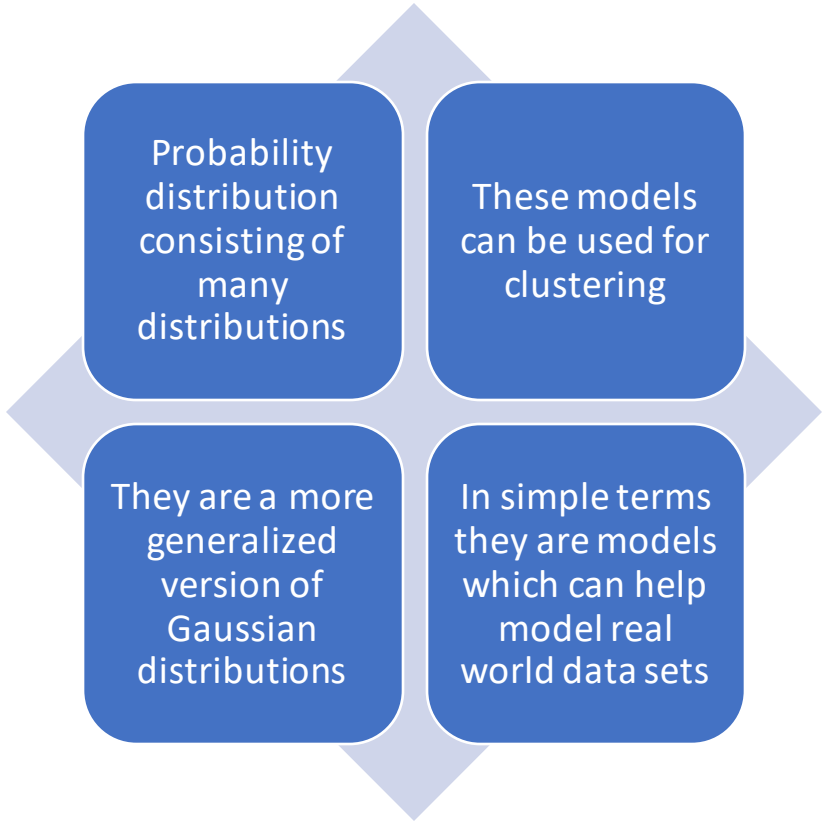
$$y = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

where μ= Mean

σ=Standard Deviation

Figure 2: Probability density function [4]

Gaussian Mixture Models



The probability distribution function of GMM can be defined as:

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

Where
- μ= Mean
- ∑= Covariance Matrix of the Gaussian
- d= The numbers of features in our dataset
- x=the number of datapoints

Figure 3: Probability density function of GMM [4]

Scenarios where Gaussian Mixtures are used

- When there is doubt about the number of clusters present, Gaussian Mixtures minimize the rate of errors greatly
- Labeling different groups in a dataset
- Extremely useful when clusters overlap, as other algorithms will identify the whole cluster as one
- Identifying fraud

Clustering

- Clustering is a method used that divides the population into groups, and the data points which will be present in the group will be similar to each other
- Groups with similar traits are clustered together
- There are 2 types of clustering types hard and soft clustering
- Purpose of unsupervised learning is finding hidden patterns in unlabeled data

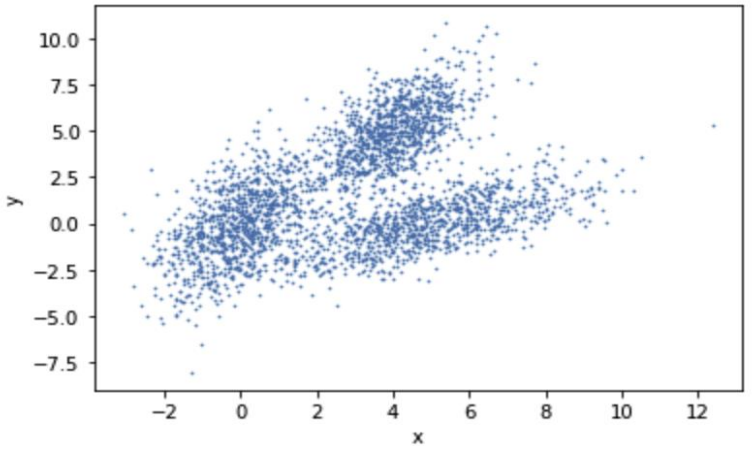


Figure 4 : Before clustering [6]

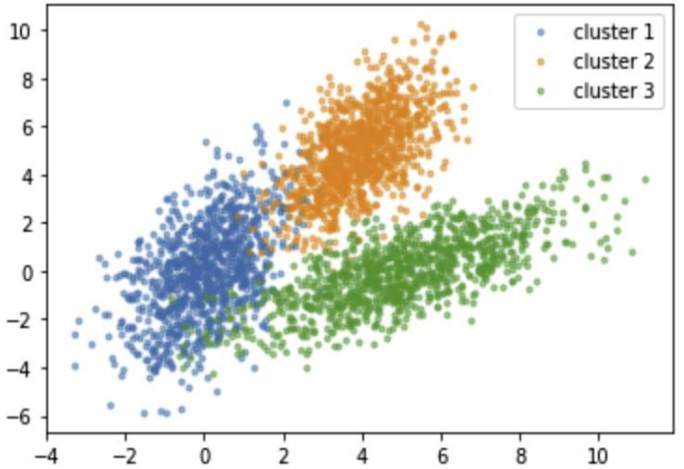


Figure 5: After clustering [6]

Clustering in Gaussian Mixtures

- Gaussian Mixtures can be used in Clustering
- Clustering is the task of grouping a set of data
- It is a soft clustering algorithm
- We can use these models to find clusters in data sets which might be difficult to find via other methods

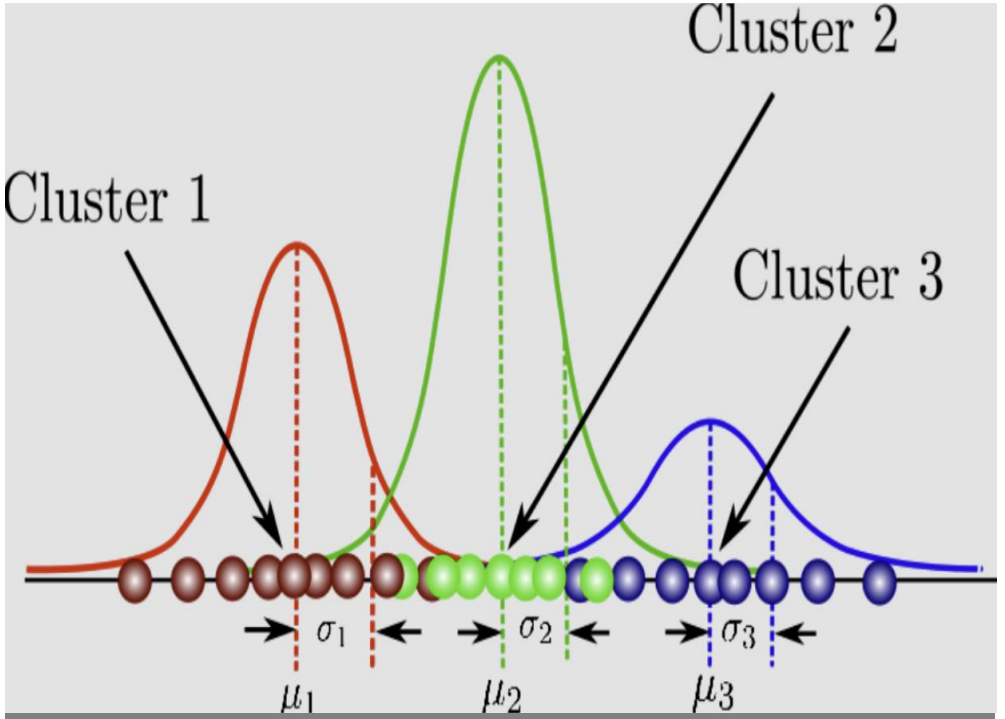


Figure 6: Clustering in GMM [1]

Expectation Maximization In Gaussian Mixtures

- This method is used for estimating the parameters (mean, variance, covariance) of the Gaussian Mixtures
- E stands for estimation and M stands for maximization
- Expectation (E) is used to identify the parameters
- Maximization (M) is used to identify the points which can be included and which are not fit to add

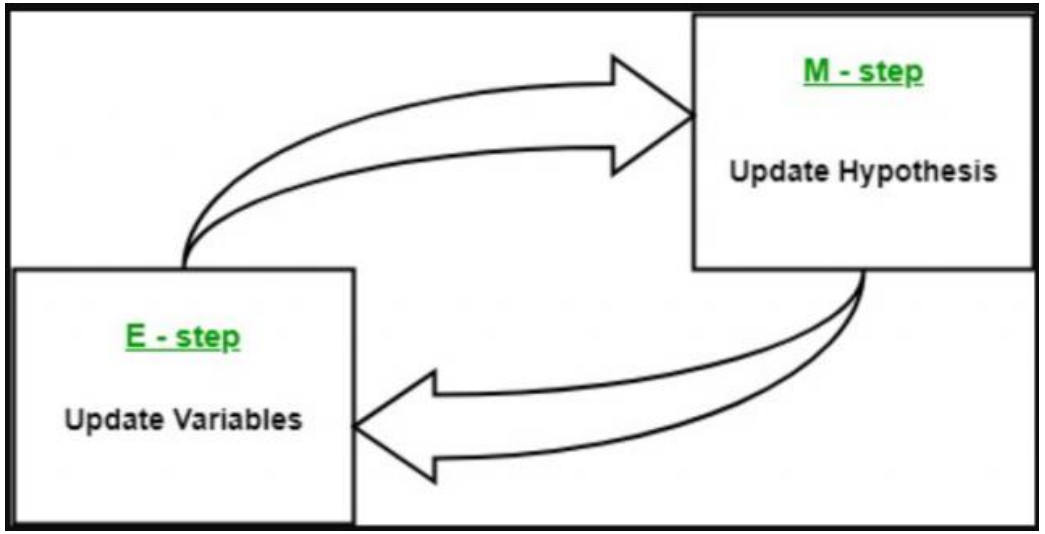


Figure 7: Expectation Maximization process [2]

Gaussian Mixture Model compared to other models

- An indication of what model to use can be decided from the size of the data set
- Gaussian Mixtures are very effective when the data set is smaller and data sets are not separated that well, whereas K-means clustering can be used vice versa
- GMM are more flexible as opposed to K-means clustering, as different type of cluster shapes can be taken

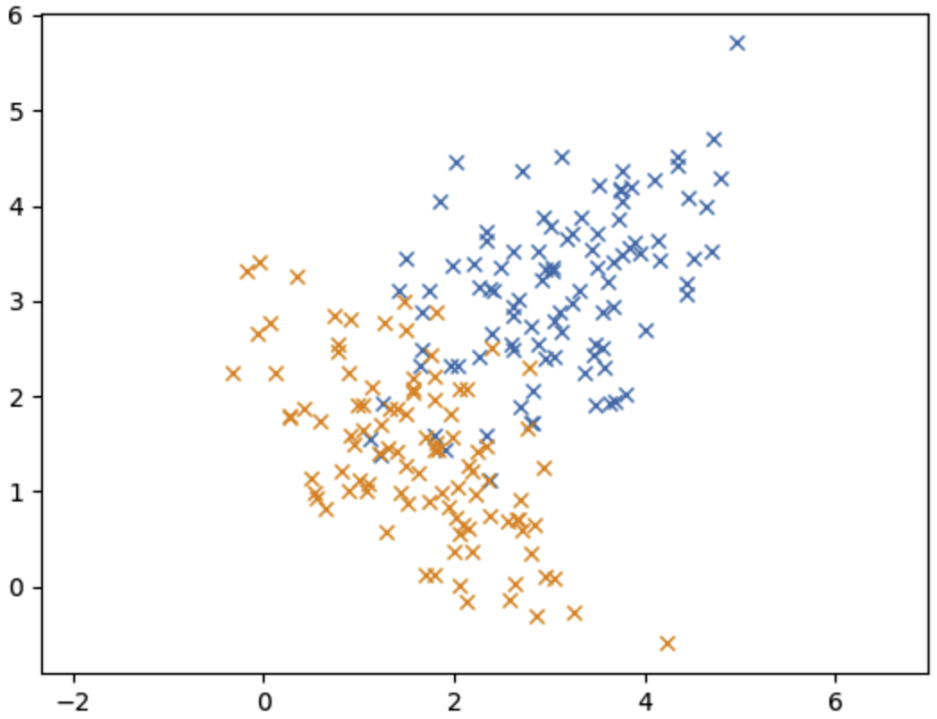


Figure 8: Example of overlapping of data [3]

Application of Gaussian Mixtures

- Customer Behavior Analysis
- Medical Datasets
- Earthquake studies
- Insurance

Advantages of Gaussian Mixture Models

01

Higher flexibility, as clusters with different shapes can be taken into account

02

Soft Classification

03

Robust to outliers

Disadvantages of Gaussian Mixtures

01

Limited number of components

02

Slow as number of iterations have too happen to achieve convergence

03

Only number of components that are accessible will be used

Conclusion

- In conclusion Gaussian Mixtures are useful where the data sets are large and patterns within the dataset have to be identified. GMM has made it very easy to label clusters. Moreover, by using this method, unsupervised data can be converted into data that is classified and clean.

References

- [1] Kumar, A. (2022, April 14). *Gaussian mixture models: What are they & when to use?* Data Analytics. Retrieved May 23, 2022, from https://vitalflux.com/gaussian-mixture-models-what-are-they-when-to-use/#What_is_the_expectation-maximization_EM_method_in_relation_to_GMM
- [2]Kumar, A. (2022, April 14). *Gaussian mixture models: What are they & when to use?* Data Analytics. Retrieved May 23, 2022, from https://vitalflux.com/gaussian-mixture-models-what-are-they-when-to-use/#What_is_the_expectation-maximization_EM_method_in_relation_to_GMM
- [3] 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/
- [4] Verma, Y. (2021b, October 25). *All You Need to Know About Gaussian Mixture Models*. Analytics India Magazine. Retrieved May 28, 2022, from https://analyticsindiamag.com/all-you-need-to-know-about-gaussian-mixture-models/
- [5] Shewan, D. (2021, December 3). *10 Companies Using Machine Learning in Cool Ways*. WordStream. Retrieved May 28, 2022, from https://www.wordstream.com/blog/ws/2017/07/28/machine-learning-applications
- [6] Alireza Khanteymoori, Anup Kumar, 2021 **Clustering in Machine Learning (Galaxy Training Materials)**. https://training.galaxyproject.org/training-material/topics/statistics/tutorials/clustering_machinelearning/tutorial.html Online; accessed Sat May 28 2022