- **Gaussian Distribution**
- **Gaussian mixture model**


Jagendra Singh

# GAUSSIAN DISTRIBUTION

- The Gaussian distribution is the healthy-studied probability distribution. It is for nonstop-valued random variables.
- It is as well stated as the normal distribution.
- In real life, many datasets can be modeled by Gaussian Distribution (Univariate or Multivariate).
- So it is quite natural and intuitive to assume that the clusters come from different Gaussian Distributions.

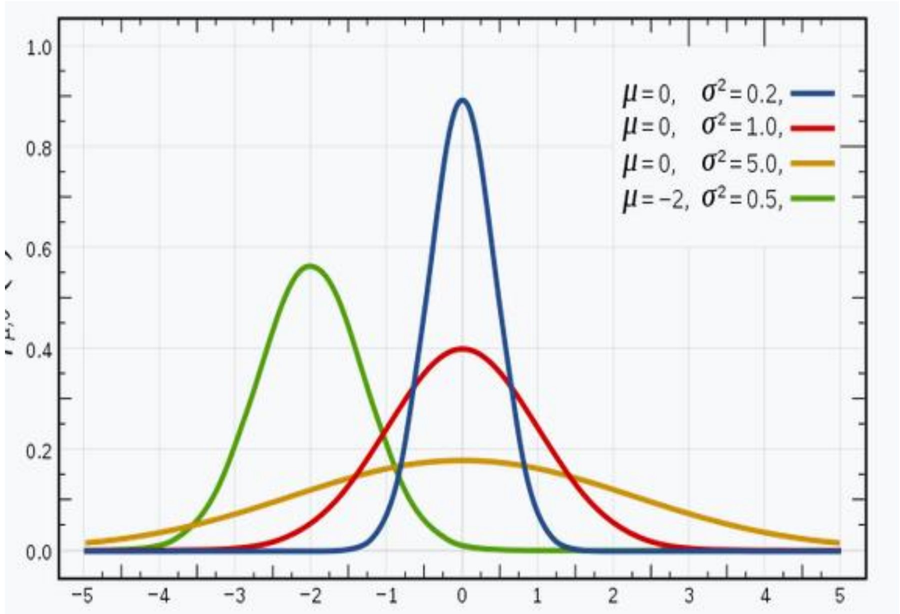
# GAUSSIAN DISTRIBUTION

- Or in other words, it is tried to model the dataset as a mixture of several Gaussian Distributions.
- In one dimension the probability density function of a Gaussian Distribution is given by

$$G(X|\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$$

where μ and σ ^ 2 are respectively mean and variance of the distribution.

# GAUSSIAN DISTRIBUTION




For Multivariate ( let us say d-variate) Gaussian Distribution, the probability density function is given by

$$G(X|\mu,\Sigma) = \frac{1}{\sqrt{(2\pi)|\mathbf{\Sigma}|}} \exp\left(-\frac{1}{2}(X-\mu)^T \mathbf{\Sigma}^{-1}(X-\mu)\right)$$

Here, μ is a d dimensional vector denoting the mean of the distribution and **Σ** is the d X d covariance matrix.

# Parameters of Gaussian Distribution

## Mean

The mean (also know as average), is obtained by dividing the sum of observed values by the number of observations, n. Although data points fall above, below, or on the mean, it can be considered a good estimate for predicting subsequent data points.

## Standard deviation

The standard deviation gives an idea of how close the entire set of data is to the average value.

Data sets with a small standard deviation have tightly grouped, precise data. Data sets with large standard deviations have data spread out over a wide range of values.

$$\sigma = \sqrt{\frac{\sum(x_i - \mu)^2}{N}}$$

## CI

**A confidence interval**, in statistics, refers to the probability that a population parameter will fall between a set of values for a certain proportion of times.

A range of values so defined that there is a specified probability that the value of a parameter lies within it.

$$CI = \bar{x} \pm z\frac{s}{\sqrt{n}}$$

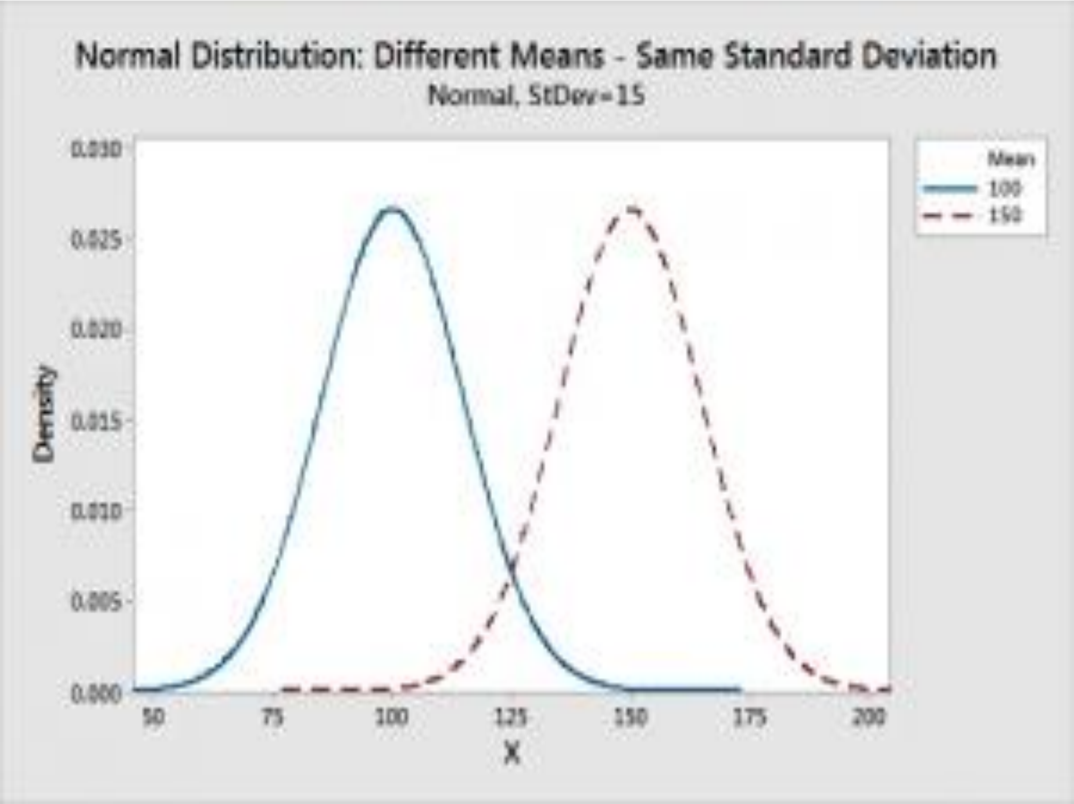
# Mean

•Scientists used the mean or average value as a measure of dominant trend.

•It may be used to define the distribution of variables that are measured as ratios or intervals.

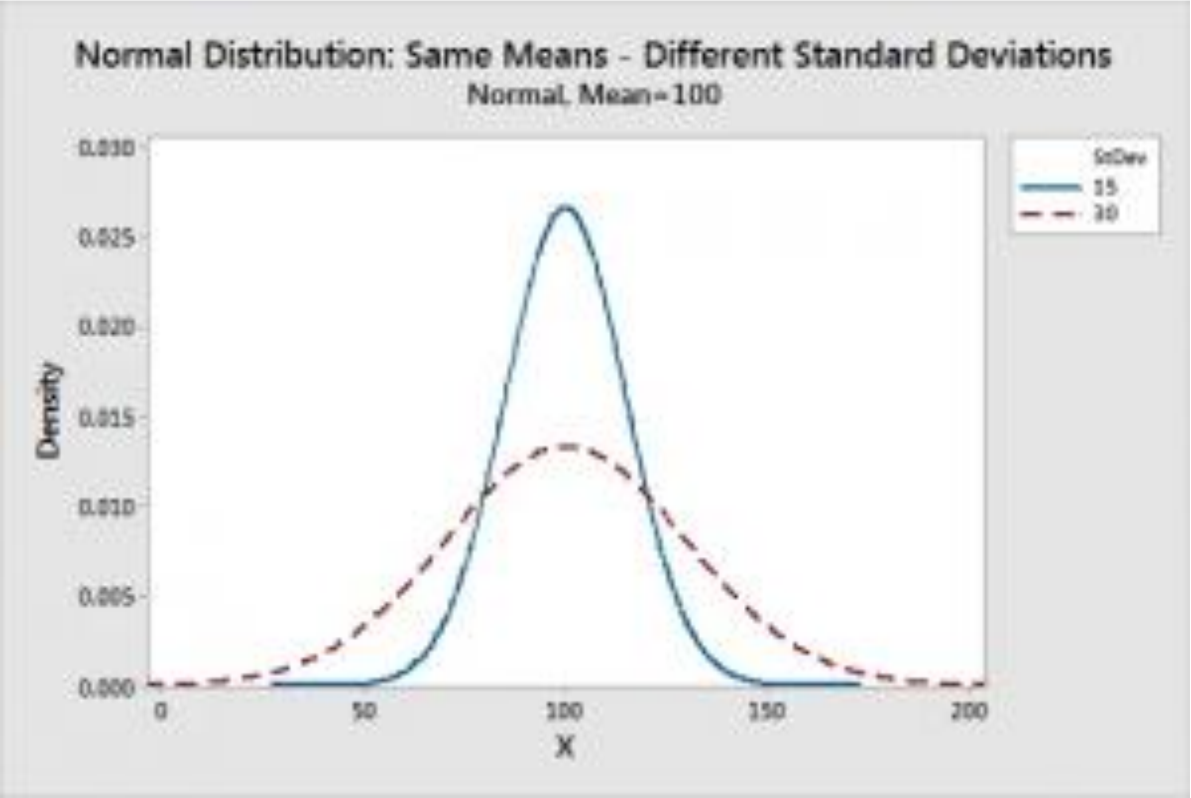
•The mean decides the site of the peak. The many of the data points are gathered around the mean in a normal distribution graph.

•By changing the value of the mean we can move the curve of Gaussian Distribution moreover to the left or right along the X-axis.



# Standard Deviation

- It acts how the data points are circulated comparative to the mean.
- It fixes how distant the data points are missing from the mean.
- It characterizes the distance between the mean and the data points.
- It describes the width of the graph. Consequently, varying the value of it make tighter or enlarges the width of the distribution along the x-axis.
- A lesser standard deviation with respect to the mean results in a steep curve and a greater standard deviation marks in a flatter curve.

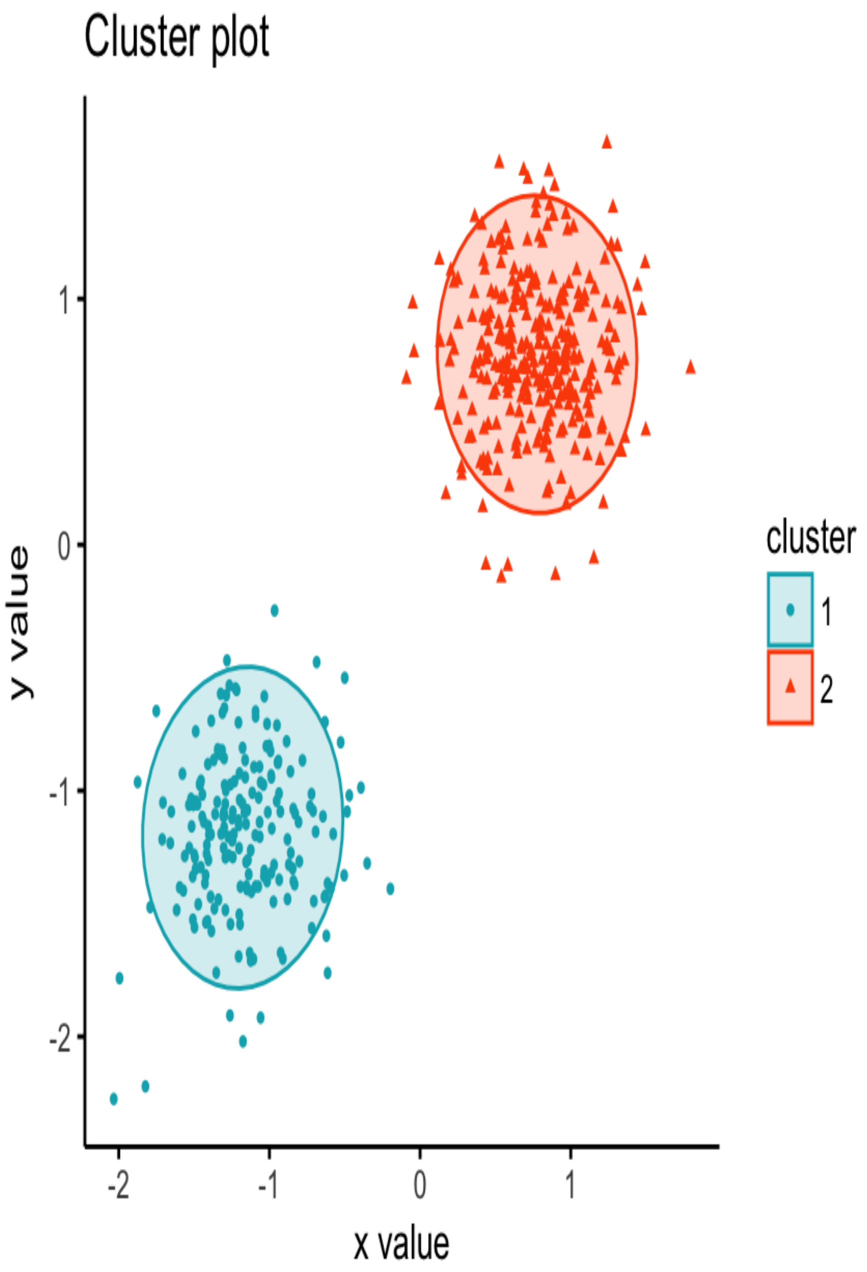


# IMPORTANCE OF GAUSSIAN DISTRIBUTION

- It is ever-present as a dataset with finite variance turns into Gaussian as long as dataset with free feature-probabilities is permitted to raise in size.
- It is the most significant probability distribution in statistics as it turns many natural phenomena such as age, height, test-scores, IQ scores, and sum of the rolls of two cubes and so on.

# IMPORTANCE OF GAUSSIAN DISTRIBUTION

- Datasets with Gaussian distributions creates valid to a diversity of methods that decrease under parametric statistics.
- The approaches for example propagation of uncertainty and least squares parameter right are related only to datasets with normal or normal-like distributions.
- Reviews and conclusions resulting from such analysis are intuitive. That also easy to explain to audiences with basic knowledge of statistics.




# CLUSTERING

- Suppose there are set of data points that need to be grouped into several parts or clusters based on their similarity. In machine learning, this is known as Clustering.

  There are several methods available for clustering:

- K Means Clustering
- Hierarchical Clustering
- Gaussian Mixture Models

# INTRODUCTION OF GAUSSIAN MIXTURE MODELS

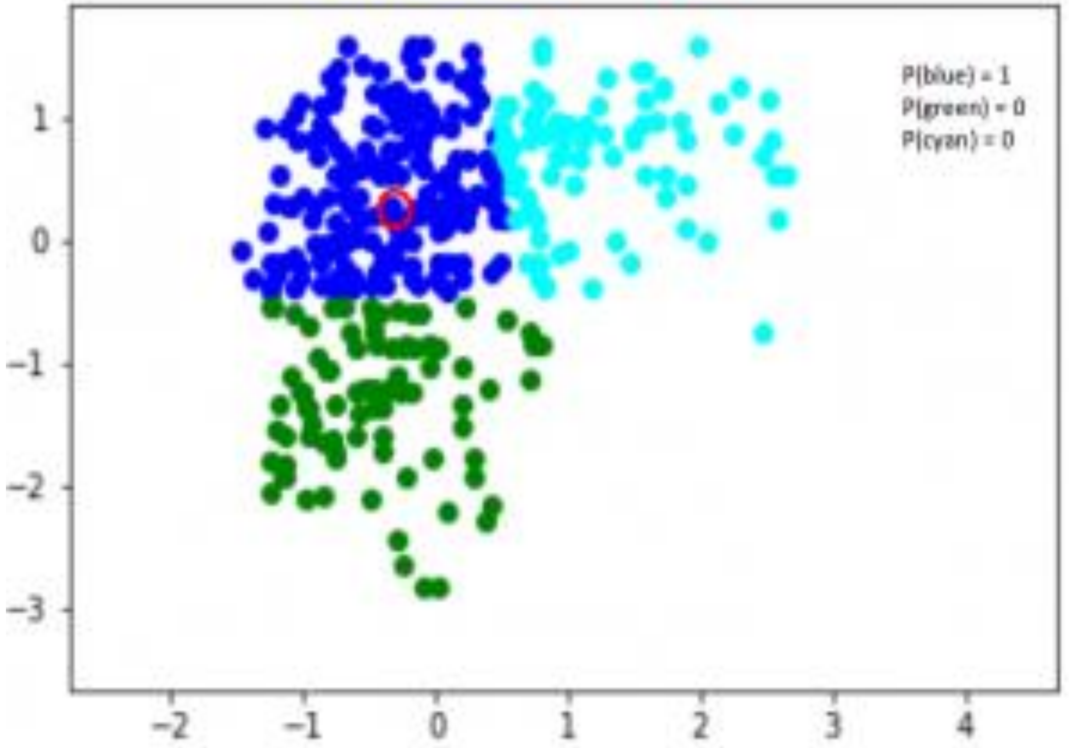
- Gaussian Mixture Models (GMMs) assume that there are a certain number of Gaussian distributions, and each of these distributions represent a cluster.
- Hence, Model tends to group the data points belonging to a single distribution together.
- Let's say we have three Gaussian distributions (more on that in the next section) - GD1, GD2, and GD3.

# INTRODUCTION OF GAUSSIAN MIXTURE MODELS

- These have a certain mean (**μ**1, **μ**2, **μ**3) and variance (**σ**1, **σ**2, **σ**3) value respectively.
- For a given set of data points, our GMM would identify the probability of each data point belonging to each of these distributions.
- Probabilistic models and use the soft clustering approach for distributing the points in different clusters.

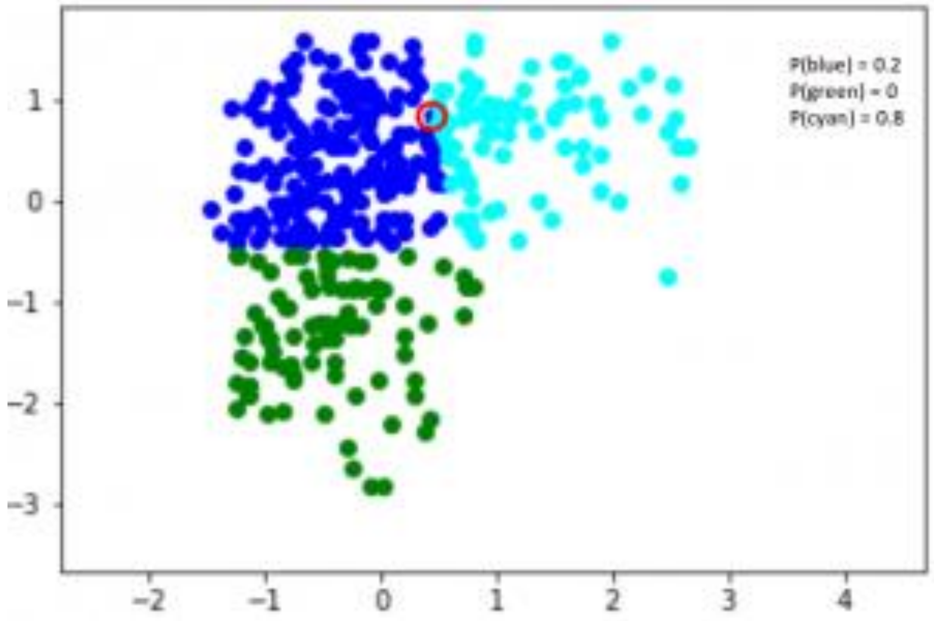
# EXAMPLE

- Here, we have three clusters that are denoted by three colors - Blue, Green, and Cyan. Let's take the data point highlighted in red.
- The probability of this point being a part of the blue cluster is 1, while the probability of it being a part of the green or cyan clusters is 0.



# EXAMPLE

- Now, consider another point – somewhere in between the blue and cyan (highlighted in the below figure).
- The probability that this point is a part of cluster green is 0, right? And the probability that this belongs to blue and cyan is 0.2 and 0.8 respectively.



# GAUSSIAN MIXTURE MODEL

- Gaussian Mixture Model or Mixture of Gaussian as it is sometimes called, it is not so much a model as it is a probability distribution.
- It is a universally used model for generative unsupervised learning or clustering.
- It is also called Expectation-Maximization Clustering or EM Clustering and is based on the optimization strategy.
- Gaussian Mixture models are used for representing Normally Distributed subpopulations within an overall population.

# GAUSSIAN MIXTURE MODEL

| Speed (Km/h) | Frequency |
|---|---|
| 1 | 4 |
| 2 | 9 |
| 3 | 6 |
| 4 | 7 |
| 5 | 3 |
| 6 | 2 |

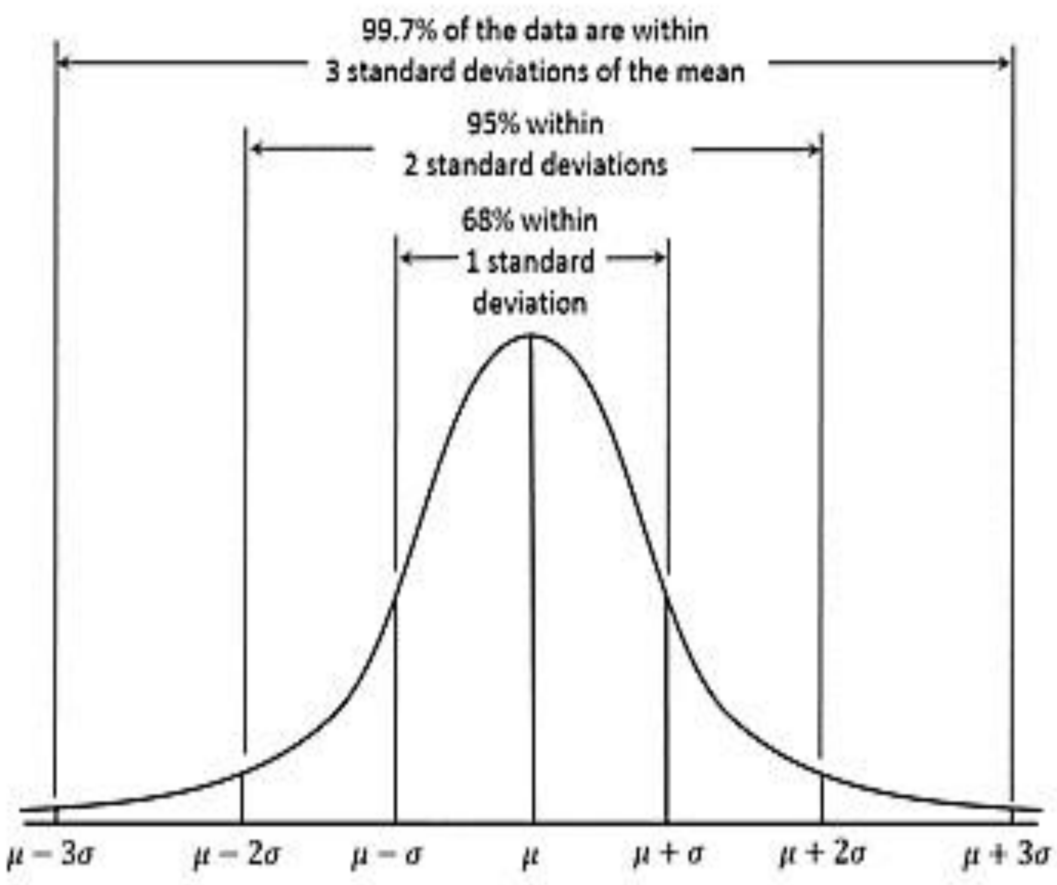
Let's take an example to understand.

**We have a data table that lists a set of cyclist's speeds.** Here, we can see that a cyclist reaches the speed of 1 Km/h four times, 2Km/h nine times, 3 Km/h and so on.

We can notice how this follows, the frequency goes up and then it goes down. It looks like it follows a kind of bell curve the frequencies go up as the speed goes up and then it has a peak value and then it goes down again, and we can represent this using a bell curve otherwise known as a Gaussian distribution.

# GAUSSIAN MIXTURE MODEL

- A Gaussian distribution is a type of distribution where half of the data falls on the left of it, and the other half of the data falls on the right of it.
- It's an even distribution, and one can notice just by the thought of it intuitively that it is very mathematically convenient.



# GAUSSIAN MIXTURE MODEL

- The formula for Gaussian distribution using the mean and the standard deviation called the **Probability Density Function:**

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

where

μ= Mean

σ=Standard Deviation

# GAUSSIAN MIXTURE MODEL

- It is a probability distribution that consists of multiple probability distributions and has Multiple Gaussians.
- The probability distribution function of d-dimensions Gaussian Distribution is defined as:

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

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

# K-Means VS Gaussian Mixture Model

- It is very similar to the k-means algorithm. It uses the same optimization strategy which is the expectation maximization algorithm.
- The reason that standard deviation is added into this because in the denominator the 2 takes variation into consideration.
- when it calculates its measurement but K means only calculates conventional Euclidean distance.
- i.e K-means calculates distance and GM calculates weights.

K-means finds k to minimize $(x-\mu_k)^2$

The Gaussian Mixture Model is,

$$\frac{(x-\mu_k)^2}{\sigma^2}$$

# WHY DO WE USE THE VARIANCE-COVARIANCE MATRIX?

- The Covariance is a measure of how changes in one variable are associated with changes in a second variable.
- It's not about the independence of variation of two variables but how they change depending on each other.
- The [variance-covariance matrix]() is a measure of how these variables are related to each other.

$$V = \left[ \frac{\sum x_1^2}{N} \; \frac{\sum x_1 x_2}{N} \;\cdots\cdots\cdots\; \frac{\sum x_1 x_c}{N} \; \frac{\sum x_2 x_1}{N} \; \frac{\sum x_2^2}{N} \;\cdots\cdots\; \frac{\sum x_2 x_c}{N} \; \frac{\sum x_c x_1}{N} \;\ldots\; \frac{\sum x_c x_2}{N} \;\cdots\cdots\cdots\; \frac{\sum x_c^2}{N} \right]$$

Where, V= c x c variance-covariance matrix

N = the number of scores in each of the c datasets

xi= is a deviation score from the ith dataset

xi2/N= is the variance of element from the ith dataset

xixj/N= is the covariance for the elements from the ithand jth datasets

and the probability given in a mixture of K Gaussian where K is a number of distributions:

# ADVANTAGE

- Gaussian Mixture models are used for representing Normally Distributed subpopulations within an overall population.
- The advantage of Mixture models is that they do not require which subpopulation a data point belongs to.
- It allows the model to learn the subpopulations automatically.

# APPLICATIONS

- GMM is widely used in the field of signal processing.
- GMM provides good results in language Identification.
- Customer Churn is another example.
- GMM founds its use case in Anomaly Detection.
- GMM is also used to track the object in a video frame.
- GMM can also be used to classify songs based on genres.

# THANK YOU