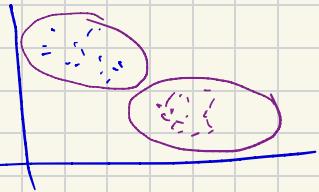


## Gaussian Mixture Model (GMM)

GMM is a probabilistic model that assumes all the data points are generated from a mixture of a finite number of Gaussian distributions with unknown parameters.



it is a type of clustering algorithm similar to K-means but much more flexible. Instead of assigning a data point to a single cluster, GMM provides a probability that a data point belongs to each of the individual clusters. This is called soft clustering.

### Components of GMM

- Number of Components ( $K$ ): The number of GD you believe exist in the data.
- Mixture weights ( $\pi_k$ ): The proportion of the total data that belongs to each cluster.

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

- Gaussian parameters for each cluster  $k$ .
  - ▷ Mean ( $\mu_k$ ) ← center/avg. of the cluster
  - ▷ Covariance ( $\Sigma_k$ ) ← the spread and shape of the cluster

$$N(x|\mu, \Sigma) = \frac{1}{\sqrt{(2\pi)^D |\Sigma|}} e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)}$$

$x$  is a data point

$\mu$  is the mean

$\Sigma$  is the covariance Matrix.  
 $D$  number of Dimensions

### GMM formula

Simplifying a weighted sum of several of these G.D.

$$p(x) = \sum_{i=1}^K \pi_k N(x|\mu_k, \Sigma_k)$$

$K$  is the number of clusters.

$\pi_k$  is the mixture weight for cluster  $k$ .  
 $N(x|\mu_k, \Sigma_k)$  is the G.D. for cluster  $k$ .

$$\boxed{\sum_{i=1}^K \pi_k = 1}$$

Expectation-Maximization (EM) algorithm is a clever iterative method that solves this problem by breaking it down into two repeating steps.

→ It starts with a random guess and refine it until it finds a good solution

## Expectation step (E-step) / fitting steps.

To calculate the responsibilities of each cluster for each data point

for a data point  $x_i$  and cluster  $K$ , the responsibility  $\gamma(z_{ik})$  is calculated as

$$\gamma(z_{ik}) = \frac{\pi_k N(x_i | \mu_k, \Sigma_k)}{\sum_{j=1}^K N(x_i | \mu_j, \Sigma_j)}$$

prob. of i<sup>th</sup> sample was gen by k<sup>th</sup> qf

prob. Sample i<sup>th</sup> belongs to k<sup>th</sup>.

Weighted sum over prob

$$\pi_k \underbrace{[0.7, 0.2, 0.1]}_{\sum_{k=1}^K \gamma(z_{ik})} = 1$$

$$= \boxed{1}$$

## The Maximization step (M-step)

To update the model's parameters using the responsibilities calculated in the E-step.

each cluster  $K$ , we update the parameters as follows

Update the mixture weights

$$\pi_k^{\text{new}} = \frac{1}{N} \sum_{i=1}^N \gamma(z_{ik})$$

Update the mean ( $\mu_k$ )

$$\mu_k^{\text{new}} = \frac{\sum_{i=1}^N \gamma(z_{ik}) x_i}{\sum_{i=1}^N \gamma(z_{ik})}$$

Update the covariance Matrices ( $\Sigma_k$ )

$$\sum_k^{\text{new}} = \frac{\sum_{i=1}^N \gamma(z_{ik}) (x_i - \mu_k^{\text{new}}) (x_i - \mu_k^{\text{new}})^T}{\sum_{i=1}^N \gamma(z_{ik})}$$

After M-step, we have a new hopefully better, set of parameters. We then repeat the E-step and M-step until the values of the parameters stop changing significantly b/w the iterations.

The EM algo. is guaranteed to find a set of parameters that are at a local maximum of log-likelihood.

Log-likelihood for a GMM

$$\ln p(x|\pi, \mu, \Sigma) = \sum_{i=1}^N \ln \left( \sum_{k=1}^K \pi_k N(x_i | \mu_k, \Sigma_k) \right)$$

### GMM vs K-Means

	K-mean	GMM
Assignment	Hard clustering	Soft clustering.
cluster shape.	Spherical	Elliptical (oval)
Mechanism	Minimizes the sum of Squared	Maximizes the likelihood of the data given model (EM Algo.)
flexibility	less flexible	Much more flexible & can model more complex data distribution.