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 (π_K)**: The proportion of the total data that belongs to each cluster.

$$\sum_{\forall K} \pi_K = 1$$

→ **Gaussian parameters for each cluster K.**

i) Mean (μ_K) ← center/avg. of the cluster

ii) Covariance (Σ_K) ← The spread and shape of the cluster

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

x is a data point
μ is the mean
Σ is the covariance Matrix.
D number of Dimensions

GMM formula

Simply a weighted sum of several of these GD.

$$p(x) = \sum_{i=1}^{K} \pi_K N(x|\mu_K,\Sigma_K)$$

K is the number of clusters.
π_K is the mixture weight for clusters K.
$N(x|\mu_K,\Sigma_K)$ is the GD for clusters K.

$$\sum_{i=1}^{K} \pi_K = 1$$

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

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

prob. of ith sample was gen by kth gr

prob. Sample i^{th} belongs to k^{th}.

Weighted sum over prob

$$K \quad \boxed{0.7, \quad 0.2 \quad 0.1} = 1$$

$$\boxed{\sum_{\forall K} \gamma(z_{ik}) = 1}$$

The Maximization step (M-step)

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

each cluster k, we update the parameters as follo

Update the mixture weights

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

Update the mean (μ_k)

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

Update the covariance Matrices (Σ_k)

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

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

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

GMM Vs K-Means

	K-means	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 Alg
flexibility	less flexible	Much more flexible & can model more complex data distribution.