Theory of Clustering Using EM Algorithm

EM Algorithm

The Expectation-Maximization (EM) algorithm is a probabilistic method used to find the parameters of statistical models when data is incomplete or has latent (hidden) variables — like the unknown cluster a data point belongs to.

In clustering, EM is often used for Gaussian Mixture Models (GMMs), where the data is assumed to be generated from a mixture of several Gaussian distributions.

Gaussian Mixture Model (GMM)

A GMM is a model of the form:

$$P(x) = \sum_{k=1}^{K} \pi_k \cdot \mathcal{N}(x|\mu_k, \sigma_k^2)$$

Where:

- K: number of clusters
- π_k: the **mixing coefficient** (prior probability of cluster k), where $\sum \pi_k = 1$
- μ_k, σ_k: the **mean** and **variance** of cluster k
- $\mathcal{N}(x|\mu_k, \sigma_k^2)$: the Gaussian probability density function

Steps of the EM Algorithm

1. **Initialize** the parameters μ_k, σ_k, π_k
2. **E-step (Expectation)**
   Compute the **responsibilities** — the probability that each data point belongs to each cluster:
$$\gamma(z_{ik}) = \frac{\pi_k \cdot \mathcal{N}(x_i|\mu_k, \sigma_k)}{\sum_{j=1}^{K} \pi_j \cdot \mathcal{N}(x_i|\mu_j, \sigma_j)}$$
3. **M-step (Maximization)**
   Update the parameters using the new responsibilities:
$$\pi_k = \frac{1}{N}\sum_{i=1}^{N} \gamma(z_{ik}), \quad \mu_k = \frac{\sum_{i=1}^{N} \gamma(z_{ik}) x_i}{\sum_{i=1}^{N} \gamma(z_{ik})}$$
$$\sigma_k^2 = \frac{\sum_{i=1}^{N} \gamma(z_{ik})(x_i - \mu_k)^2}{\sum_{i=1}^{N} \gamma(z_{ik})}$$
4. **Repeat steps 2 and 3** until convergence.

For more details, you can refer to the below link. Maybe you find some notational differences.

https://medium.com/@gallettilance/gmm-clustering-from-scratch-a8d06a47c77d

Numerical Example

Consider one dimensional data for easiness. Two-dimensional data requires computation of covariance matrix in the bivariate Gaussian distribution.

I request the students to practice one dimensional and two dimensional as well. I will give both the examples.

One dimensional data:

X = [1.0, 2.0, 5.0]

We'll assume two clusters (K = 2).

Initialization

- Means: $\mu_1 = 1.0$, $\mu_2 = 5.0$
- Variances: $\sigma_1^2 = \sigma_2^2 = 1.0$
- Mixing coefficients: $\pi_1 = \pi_2 = 0.5$

Step 1: E-step

We compute the **responsibility** $\gamma(z_{ik})$ for each point using the Gaussian PDF:

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

Point x=1

- For cluster 1 (μ = 1.0):

$$\mathcal{N}(1|1, 1) = \frac{1}{\sqrt{2\pi}} e^{0} = 0.3989$$

- For cluster 2 (μ = 5.0):

$$\mathcal{N}(1|5, 1) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(4)^2}{2}} = 0.3989 \cdot e^{-8} \approx 0.3989 \cdot 0.0003355 \approx 0.000134$$

Responsibilities:

$$\gamma(z_{1,1}) = \frac{0.5 \cdot 0.3989}{0.5 \cdot (0.3989 + 0.000134)} \approx \frac{0.19945}{0.19952} \approx 0.99933$$

$$\gamma(z_{1,2}) = 1 - 0.99933 = 0.00067$$

Point x=2

- For cluster 1 (μ = 1.0):

$$\mathcal{N}(2|1,1) = 0.3989 \cdot e^{-0.5} \approx 0.3989 \cdot 0.6065 \approx 0.24197$$

- For cluster 2 (μ = 5.0):

$$\mathcal{N}(2|5,1) = 0.3989 \cdot e^{-4.5} \approx 0.3989 \cdot 0.0111 \approx 0.00443$$

Responsibilities:

$$\gamma(z_{2,1}) = \frac{0.5 \cdot 0.24197}{0.5(0.24197 + 0.00443)} \approx \frac{0.120985}{0.1232} \approx 0.982$$

$$\gamma(z_{2,2}) = 1 - 0.982 = 0.018$$

Pont X=5

- For cluster 1 (μ = 1.0):

$$\mathcal{N}(5|1,1) = 0.3989 \cdot e^{-8} \approx 0.000134$$

- For cluster 2 (μ = 5.0):

$$\mathcal{N}(5|5,1) = 0.3989$$

Responsibilities:

$$\gamma(z_{3,1}) = \frac{0.5 \cdot 0.000134}{0.5(0.3989 + 0.000134)} \approx \frac{0.000067}{0.1995} \approx 0.000335$$

$$\gamma(z_{3,2}) = 1 - 0.000335 = 0.999665$$

Responsibility Table

Point x	$\gamma(z_1)$	$\gamma(z_2)$
1.0	0.99933	0.00067
2.0	0.982	0.018
5.0	0.00034	0.99966

Step 2: M-step

Update parameters:

- $N_k = \sum_{i=1}^{N} \gamma(z_{ik})$
- $\mu_k = \frac{\sum \gamma(z_{ik}) x_i}{N_k}$
- $\sigma_k^2 = \frac{\sum \gamma(z_{ik})(x_i - \mu_k)^2}{N_k}$
- $\pi_k = \frac{N_k}{N}$

1. Effective cluster sizes (N_k)

$$N_1 = 0.99933 + 0.982 + 0.00034 = 1.9817$$

$$N_2 = 0.00067 + 0.018 + 0.99966 = 1.0183$$

2. New Means (μ_k)

$$\mu_1 = \frac{(0.99933)(1.0) + (0.982)(2.0) + (0.00034)(5.0)}{1.9817}$$
$$= \frac{0.99933 + 1.964 + 0.0017}{1.9817} \approx \frac{2.965}{1.9817} \approx 1.496$$

$$\mu_2 = \frac{(0.00067)(1.0) + (0.018)(2.0) + (0.99966)(5.0)}{1.0183}$$
$$= \frac{0.00067 + 0.036 + 4.9983}{1.0183} \approx \frac{5.035}{1.0183} \approx 4.945$$

3. New Variances (σ^2_k)

$$\sigma_1^2 = \frac{(0.99933)(1 - 1.496)^2 + (0.982)(2 - 1.496)^2 + (0.00034)(5 - 1.496)^2}{1.9817}$$

$$= \frac{0.99933(0.246) + 0.982(0.254) + 0.00034(12.3)}{1.9817} \approx \frac{0.245 + 0.249 + 0.0042}{1.9817} \approx 0.25$$

$$\sigma_2^2 = \text{similar steps} \approx 0.25$$

4. **New Mixing Coefficients (π_k)**

$$\pi_1 = \frac{1.9817}{3} \approx 0.66, \quad \pi_2 = 1 - \pi_1 = 0.34$$

New Parameters

- $\mu_1 \approx 1.5$, $\mu_2 \approx 4.95$
- $\sigma_1^2 = \sigma_2^2 \approx 0.25$
- $\pi_1 \approx 0.66$, $\pi_2 \approx 0.34$

repeat E-step and M-step using these updated values until convergence

Final Answer

After one or two iterations, we may get

Points close to 1.5 (i.e., 1.0 and 2.0) belong to Cluster 1

Point 5.0 belongs to Cluster 2

The EM algorithm has effectively learned the clusters and the parameters of the generating Gaussians.

Two-dimensional data Problem

Problem: Cluster 2D Data Using EM

Given Data Points (2D)

We have three 2D points:

$$x_1 = \begin{bmatrix} 1 \\ 2 \end{bmatrix}, \quad x_2 = \begin{bmatrix} 1.5 \\ 1.8 \end{bmatrix}, \quad x_3 = \begin{bmatrix} 5 \\ 8 \end{bmatrix}$$

We'll cluster them into **K = 2** clusters using EM.

We initialize:

C1:

- Mean: $\mu_1 = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$
- Covariance: $\Sigma_1 = I = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$

C2:

- Mean: $\mu_2 = \begin{bmatrix} 5 \\ 8 \end{bmatrix}$
- Covariance: $\Sigma_2 = I$

Mixing coefficients:

$\pi_1 = \pi_2 = 0.5$

Step 1: E-Step (Responsibilities)

We use the **Multivariate Gaussian PDF**:

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

Since $\Sigma = I$, we simplify:

$$\mathcal{N}(x|\mu,I) = \frac{1}{2\pi} e^{-\frac{1}{2}\|x-\mu\|^2}$$

Let's compute for $x_1 = [1,2]^T$:

γ_{11}: Responsibility for cluster 1

- $\|x_1 - \mu_1\|^2 = \|[1,2]-[1,2]\|^2 = 0$
- So,

$$p(x_1|z=1) = \frac{1}{2\pi}e^0 = \frac{1}{2\pi} \approx 0.15915$$

γ_{12}: Responsibility for cluster 2

- $\|x_1 - \mu_2\|^2 = \|[1,2]-[5,8]\|^2 = (4)^2 + (6)^2 = 16 + 36 = 52$

$$p(x_1|z=2) = \frac{1}{2\pi}e^{-52/2} = \frac{1}{2\pi}e^{-26} \approx 0.15915 \cdot e^{-26} \approx 5.71 \times 10^{-13}$$

Compute responsibilities:

$$\gamma(z_{1,1}) = \frac{0.5 \cdot 0.15915}{0.5(0.15915 + 5.71 \times 10^{-13})} \approx \frac{0.079575}{0.079575} = 1$$

$$\gamma(z_{1,2}) \approx 0$$

Point $x_2 = [1.5, 1.8]$

- $\|x_2 - \mu_1\|^2 = (0.5)^2 + (-0.2)^2 = 0.25 + 0.04 = 0.29$

$$p(x_2|z=1) = \frac{1}{2\pi} e^{-0.145} \approx 0.15915 \cdot 0.865 \approx 0.1377$$

- $\|x_2 - \mu_2\|^2 = (3.5)^2 + (6.2)^2 = 12.25 + 38.44 = 50.69$

$$p(x_2|z=2) = 0.15915 \cdot e^{-25.35} \approx 1.0 \times 10^{-11}$$

Responsibilities:

$$\gamma(z_{2,1}) \approx 1, \quad \gamma(z_{2,2}) \approx 0$$

Point $x_3 = [5, 8]$

- $\|x_3 - \mu_1\|^2 = 16 + 36 = 52 \Rightarrow p = 5.71 \times 10^{-13}$
- $\|x_3 - \mu_2\|^2 = 0 \Rightarrow p = 0.15915$

So:

$$\gamma(z_{3,1}) \approx 0, \quad \gamma(z_{3,2}) \approx 1$$

Responsibility Matrix:

Point x_i	**γ(z₁)**	**γ(z₂)**
[1.0, 2.0]	1.0	0.0
[1.5, 1.8]	1.0	0.0
[5.0, 8.0]	0.0	1.0

Step 2: M-Step (Update Parameters)

$$\mu_k = \frac{\sum \gamma(z_{ik}) x_i}{\sum \gamma(z_{ik})}$$

Cluster 1:

Only x_1 and x_2 belong to cluster 1:

$$\mu_1 = \frac{[1,2] + [1.5, 1.8]}{2} = [1.25, 1.9]$$

Cluster 2:

Only x_3 belongs to cluster 2:

$$\mu_2 = [5, 8]$$

Update covariance matrices:

For cluster 1:

$$\Sigma_1 = \frac{1}{2}\left[(x_1 - \mu_1)(x_1 - \mu_1)^T + (x_2 - \mu_1)(x_2 - \mu_1)^T\right]$$

Compute:

- $x_1 - \mu_1 = [-0.25, 0.1]$
- $x_2 - \mu_1 = [0.25, -0.1]$

So:

$$\Sigma_1 = \frac{1}{2}\left(\begin{bmatrix} 0.0625 & -0.025 \\ -0.025 & 0.01 \end{bmatrix} + \begin{bmatrix} 0.0625 & -0.025 \\ -0.025 & 0.01 \end{bmatrix}\right) = \begin{bmatrix} 0.0625 & -0.025 \\ -0.025 & 0.01 \end{bmatrix}$$

For cluster 2:

Only one point ⇒ variance is undefined; we can use identity matrix or small regularized value.

Update mixing coefficients:

$$\pi_1 = \frac{2}{3}, \quad \pi_2 = \frac{1}{3}$$

After 1 Iteration

- Cluster 1:
    - Mean: $\mu_1 = [1.25, 1.9]$
    - Covariance: as above
    - $\pi_1 = 2/3$
- Cluster 2:
    - Mean: $\mu_2 = [5, 8]$
    - Covariance: regularized identity
    - $\pi_2 = 1/3$

Clustering Result

- Points [1,2] and [1.5,1.8] → Cluster 1
- Point [5,8] → Cluster 2