Introduction to Gaussian Mixture Models and Expectation Maximization Algorithm (DRAFT)

Jun Cheng
Dept. of Intelligent Information Eng. and Sci., Doshisha University
Kyoto 610-0321 Japan
jcheng@ieee.org

***Abstract*—This manuscript introduces the basic concept of Gaussian Mixture models and the expectation maximization algorithm for clustering.**

I. Gaussian Mixture Models (1-Dimension)

A. Data Points from K Independent Sources

Let $X^{(k)}$, $k = 1, 2, \ldots, K$, be the (latent) random variable (RV) associated with the k-th source (or cluster) $\mathcal{C}_k$. The RV $X^{(k)}$ is Gaussian distributed with the probability density function (PDF)

$$p_X(x|\mathcal{C}_k) = \frac{1}{\sqrt{2\pi}\sigma_k} e^{-\frac{(x-\mu_k)^2}{2\sigma_k^2}} \triangleq \mathcal{N}(x; \mu_k, \sigma_k^2).$$

The probability of the source being active is $\pi_k \triangleq p(\mathcal{C}_k)$, where $\sum_{k=1}^K p(\mathcal{C}_k) = 1$. Let $x_1^{(k)}, x_2^{(k)}, \ldots, x_{n_k}^{(k)}$ be n_k samples from the k-th source $\mathcal{C}_k$. For K sources, we observe $N = \sum_{k=1}^K n_k$ data points in set

$$\mathcal{X} = \{x_1^{(k)}, x_2^{(k)}, \ldots, x_{n_k}^{(k)} \mid k = 1, 2, \ldots, K\}.$$

B. Mixture Models

Often times we don't observe the sources $\mathcal{C}_k$ ($k = 1, 2, \ldots, K$). This is why the $X^{(k)}$ is sometimes called latent RV. We observe only the data points in set $\mathcal{X}$ which is rewritten as

$$\mathcal{X} = \{x, x_2, \ldots, x_n, \ldots, x_N\}.$$

From the law of total probability, we know that the marginal probability of x_n is

$$p(x_n) = \sum_{k=1}^K p(x_n, \mathcal{C}_k) = \sum_{k=1}^K p(\mathcal{C}_k) p(x_n|\mathcal{C}_k) = \sum_{k=1}^K \pi_k p(x_n|\mathcal{C}_k)$$

which is the mixture model of the observation. Here $\pi_k = p(\mathcal{C}_k)$ are called mixture proportions or mixture weights. We call $p(x_n|\mathcal{C}_k)$ the mixture component, and it represents the distribution of x_n assuming it came from component $\mathcal{C}_k$. The mixture components in this note are Gaussian distributions.

If we observe N independent samples $x_1, x_2, \ldots, x_N$ from the mixture, the likelihood function is

$$p(x_1, x_2, \ldots, x_N) = \prod_{n=1}^N p(x_n) = \prod_{n=1}^N \sum_{k=1}^K \pi_k p(x_n|\mathcal{C}_k).$$

Taking the logarithm yields the following log-likelihood function (LLF)

$$\log p(x_1, x_2, \ldots, x_N) = \sum_{n=1}^N \log \sum_{k=1}^K \pi_k p(x_n|\mathcal{C}_k). \tag{1}$$

Now assume we are in the Gaussian mixture model setting where the k-th component $p(x_n|\mathcal{C}_k)$ is Gaussian $\mathcal{N}(x_n; \mu_k, \sigma_k^2)$ and the mixture proportions are π_k. A natural next question to ask is how to estimate the parameters $\{\mu_k, \sigma_k^2, \pi_k\}$ from our observations $x_1, x_2, \ldots, x_N$. In other words, how to label (or classify) each observation to its cluster $\mathcal{C}_k$, $k = 1, 2, \ldots, K$.

II. EM algorithm (1-Dimension)

Given N observations $x_1, x_2, \ldots, x_N$, we will estimate the parameters $\{\mu_k, \sigma_k^2, \pi_k\}$, $k = 1, 2, \ldots, K$. Furthermore, we label (or classify) each observation to its cluster $\mathcal{C}_k$, $k = 1, 2, \ldots, K$.

1. Initialization: Given K, randomly set $\{\mu_k^0, (\sigma_k^0)^2, \pi_k^0\}$, $k = 1, 2, \ldots, K$

2. Expectation-Step (E-Step): at the ℓ-th iteration,

$$\begin{aligned}
\underbrace{p(\mathcal{C}_k^\ell|x_n)}_{c_{k,n}^\ell} &= \frac{p(\mathcal{C}_k^{\ell-1}, x_n)}{\sum_{k'=1}^K p(\mathcal{C}_{k'}^{\ell-1}, x_n)} \\
&= \frac{p(\mathcal{C}_k^{\ell-1}) p(x_n|\mathcal{C}_k^{\ell-1})}{\sum_{k'=1}^K p(\mathcal{C}_{k'}^{\ell-1}) p(x_n|\mathcal{C}_{k'}^{\ell-1})} \\
&= \frac{\pi_k^{\ell-1} p(x_n|\mathcal{C}_k^{\ell-1})}{\sum_{k'=1}^K \pi_{k'}^{\ell-1} p(x_n|\mathcal{C}_{k'}^{\ell-1})} \\
&= \frac{\pi_k^{\ell-1} \mathcal{N}(x_n; \mu_k^{\ell-1}, \sigma_k^{\ell-1})^2)}{\sum_{k'=1}^K \pi_{k'}^{\ell-1} \mathcal{N}(x_n; \mu_{k'}^{\ell-1}, \sigma_{k'}^{\ell-1})^2)}.
\end{aligned} \tag{2}$$

3. Maximization-Step (M-Step):

$$\begin{aligned}\pi_k^\ell &\triangleq p(\mathcal{C}_k^\ell) = \sum_{n=1}^{N} p(\mathcal{C}_k^\ell, x_n) \\ &= \sum_{n=1}^{N} p(x_n) p(\mathcal{C}_k^\ell|x_n) = \frac{1}{N}\sum_{n=1}^{N} \underbrace{p(\mathcal{C}_k^\ell|x_n)}_{c_{k,n}^\ell}. \end{aligned} \tag{3}$$

Since

$$\begin{aligned} p(x_n|\mathcal{C}_k^\ell) &= \frac{p(\mathcal{C}_k^\ell, x_n)}{\sum_{n'=1}^{N} p(\mathcal{C}_k^\ell, x_{n'})} = \frac{p(x_n)p(\mathcal{C}_k^\ell|x_n)}{\sum_{n'=1}^{N} p(x_{n'})p(\mathcal{C}_k^\ell|x_{n'})} \\ &= \frac{(1/N)p(\mathcal{C}_k^\ell|x_n)}{\sum_{n'=1}^{N}(1/N)p(\mathcal{C}_k^\ell|x_{n'})} = \frac{p(\mathcal{C}_k^\ell|x_n)}{\sum_{n'=1}^{N} p(\mathcal{C}_k^\ell|x_{n'})} \\ &= \frac{p(\mathcal{C}_k^\ell|x_n)}{N\pi_k^\ell} \end{aligned}$$

we have

$$\mu_k^\ell = \sum_{n=1}^{N} p(x_n|\mathcal{C}_k^\ell) x_n = \sum_{n=1}^{N} \frac{\overbrace{p(\mathcal{C}_k^\ell|x_n)}^{c_{k,n}^\ell}}{N\pi_k^\ell} x_n. \tag{4}$$

$$\begin{aligned} (\sigma_k^2)^\ell &= \sum_{n=1}^{N} p(x_n|\mathcal{C}_k^\ell)(x_n - \mu_k^\ell)^2 \\ &= \sum_{n=1}^{N} \frac{\overbrace{p(\mathcal{C}_k^\ell|x_n)}^{c_{k,n}^\ell}}{N\pi_k^\ell}(x_n - \mu_k^\ell)^2. \end{aligned} \tag{5}$$

4. Evaluation-Step (Eva-Step)

We evaluate (see (1))

$$\begin{aligned} \Gamma^\ell &\triangleq \sum_{n=1}^{N} \log \sum_{k=1}^{K} \pi_k^\ell p(x_n|\mathcal{C}_k^\ell) \\ &= \sum_{n=1}^{N} \log \left(\frac{1}{N} \sum_{k=1}^{K} \underbrace{p(\mathcal{C}_k^\ell|x_n)}_{c_{k,n}^\ell} \right) \end{aligned} \tag{6}$$

at each iteration of E-Step and M-Step, and check the convergence of the algorithm.

5. Result of the estimation

The parameters

$$\{\mu_k^{\ell\to\infty}, (\sigma_k^2)^{\ell\to\infty}, \pi_k^{\ell\to\infty}\}, k = 1, 2, \ldots, K$$

are estimated such that the LLF (see (6)) $\Gamma^{\ell\to\infty}$ converse. The probability that x_n belongs to $\mathcal{C}_k$ is

$$\Pr(x_n \in \mathcal{C}_k) \triangleq p(\mathcal{C}_k^{\ell\to\infty}|x_n) = c_{k,n}^{\ell\to\infty}, \text{(soft clustering)} \quad n = 1, 2, \ldots, N, k = 1, 2, \ldots, K.$$

The observation x_n is labeled (or clustered) to $\mathcal{C}_k$,

$$x_n \in \mathcal{C}_{k*}, \text{(hard clustering)} \quad n = 1, 2, \ldots, N$$

where

$$k^* = \underset{k=1,2,\ldots,K}{\operatorname{argmax}}\ p(\mathcal{C}_k^{\ell\to\infty}|x_n).$$

Finally, we summery the EM algorithm in Algorithm 1.

Algorithm 1: EM algorithm

Result: Labeling $x_1, \ldots, x_N$ to $\mathcal{C}_k$, $k = 1, \ldots, K$;
input: $x_1, \ldots, x_N$;
initialization: K, $\{\mu_k^0, (\sigma_k^0)^2, \pi_k^0\}$, L;
while $\underline{\ell \le L}$ **do**
  compute $c_{k,n}^\ell$ in (2) ;
  compute π_k^ℓ, μ_k^ℓ, $(\sigma_k^2)^\ell$, in (3), (4), (5), respectively;
  If $(\Gamma^\ell - \Gamma^{\ell-1})^2 < \epsilon$ **break**
end
outputs: $\Pr(x_n \in \mathcal{C}_k) = c_{k,n}^\ell$;
  $x_n \in \mathcal{C}_{k^*}$ where $k^* = \underset{k=1,2,\ldots,K}{\operatorname{argmax}}\ p(\mathcal{C}_k^\ell|x_n)$;

III. EM ALGORITHM (d-DIMENSION)

We observe N independent samples

$$\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n, \ldots, \boldsymbol{x}_N$$

from a Gaussian mixture model

$$p(\boldsymbol{x}_n) = \sum_{k=1}^{K} \pi_k p(\boldsymbol{x}_n|\mathcal{C}_k)$$

where $\boldsymbol{x}_n$ are length-d real vector, π_k are mixture weights, and mixture components are d-variable Gaussian distribution with PDF

$$\begin{aligned} & p(\boldsymbol{x}|\mathcal{C}_k) \\ &= \frac{1}{(2\pi)^{d/2}|\Sigma_k|^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_k)^{\mathrm{T}}\Sigma_k^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_k)\right) \\ &\triangleq \mathcal{N}(x; \boldsymbol{\mu}_k, \Sigma_k). \end{aligned}$$

Here the exponent part is represented as

$$(\boldsymbol{x}-\boldsymbol{\mu}_k)^{\mathrm{T}}\Sigma_k^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_k) = \sum_{j=1}^{d}\sum_{j=1}^{d}(x_{n,j}-\mu_{k,j})(\Sigma_k^{-1})_{j,i}(x_{n,i}-\mu_{k,i}),$$

and

$$\boldsymbol{x}_n = (x_{n,1}, \ldots, x_{n,d})^{\mathrm{T}}, \quad \boldsymbol{\mu}_k = (\mu_{k,1}, \ldots, \mu_{k,d})^{\mathrm{T}}.$$

Σ_k is $d \times d$ covariance matrix, Σ_k^{-1} is the inverse of Σ_k, $(\Sigma_k^{-1})_{j,i}$ is the (j,i)-th element of Σ_k^{-1}, and $|\Sigma_k|$ is the determinant of Σ_k.

The log-likelihood function is

$$\log p(\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N) = \sum_{n=1}^{N} \log \sum_{k=1}^{K} \pi_k p(\boldsymbol{x}_n|\mathcal{C}_k). \tag{7}$$

Similar to 1-dimension, the EM algorithm is as follows:

1. Initialization: Given K, randomly set $\{\boldsymbol{\mu}_k^0, \Sigma_k^0, \pi_k^0\}$, $k = 1, 2, \ldots, K$.

2. E-Step:

$$\underbrace{p(\mathcal{C}_k^\ell|\boldsymbol{x}_n)}_{c_{k,n}^\ell} = \frac{\pi_k^{\ell-1}\mathcal{N}(\boldsymbol{x}_n; \boldsymbol{\mu}_k^{\ell-1}, \Sigma_k^{\ell-1})}{\sum_{k=1}^K \pi_k^{\ell-1}\mathcal{N}(\boldsymbol{x}_n; \mu_k^{\ell-1}, \Sigma_k^{\ell-1})}. \tag{8}$$

3. M-Step:

$$\pi_k^\ell = \frac{1}{N}\sum_{n=1}^N \underbrace{p(\mathcal{C}_k^\ell|\boldsymbol{x}_n)}_{c_{k,n}^\ell} \tag{9}$$

$$\boldsymbol{\mu}_k^\ell = \sum_{n=1}^N \frac{\overbrace{p(\mathcal{C}_k^\ell|\boldsymbol{x}_n)}^{c_{k,n}^\ell}}{N\pi_k^\ell}\boldsymbol{x}_n. \tag{10}$$

$$(\Sigma_k^\ell)_{j,i} = \sum_{n=1}^N \frac{\overbrace{p(\mathcal{C}_k^\ell|\boldsymbol{x}_n)}^{c_{k,n}^\ell}}{N\pi_k^\ell}(x_{n,j} - \mu_{k,j}^\ell)(x_{n,i} - \mu_{k,i}^\ell) \tag{11}$$

4. Eva-Step

We evaluate (see (7))

$$\begin{aligned} \Gamma^\ell &\triangleq \sum_{n=1}^N \log \sum_{k=1}^K \pi_k^\ell p(\boldsymbol{x}_n|\mathcal{C}_k^\ell) \\ &= \sum_{n=1}^N \log\left(\frac{1}{N}\sum_{k=1}^K \underbrace{p(\mathcal{C}_k^\ell|\boldsymbol{x}_n)}_{c_{k,n}^\ell}\right) \end{aligned} \tag{12}$$

at each iteration of E-Step and M-Step, and check the convergence of the algorithm.

5. Result of the estimation

The probability that $\boldsymbol{x}_n$ belongs to $\mathcal{C}_k$ is

$$\begin{aligned} \Pr(\boldsymbol{x}_n \in \mathcal{C}_k) \triangleq p(\mathcal{C}_k^{\ell\to\infty}|\boldsymbol{x}_n) = c_{k,n}^{\ell\to\infty}, &\text{(soft clustering)} \\ n = 1, 2, \ldots, N, k = 1, 2, \ldots, K. \end{aligned}$$

The observation $\boldsymbol{x}_n$ is labeled (or clustered) to $\mathcal{C}_k$,

$$\begin{aligned} \boldsymbol{x}_n \in \mathcal{C}_{k*}, &\text{(hard clustering)} \\ n = 1, 2, \ldots, N \end{aligned}$$

where

$$k^* = \underset{k=1,2,\ldots,K}{\operatorname{argmax}}\, p(\mathcal{C}_k^{\ell\to\infty}|\boldsymbol{x}_n).$$

References