Solutions to Assignment 2

18214613 陈承勃

Dec 31st, 2018

Abstract

In this report, I first derive the mathematical solutions for Gaussian Mixtures using EM algorithm from scratch in **Section 1**, and give a general framework of EM algorithm for GMM. In **Section 2**, I display my solutions to **Assignment 2** based on the theoretical results I derived. The results of my experiments show that EM algorithm can be well applied to Gaussian mixtures, and is quite sensitive to the initialization of parameters, that is, under different initializations, not only the speed of convergence variates, but also the training get down to different convergence. The experiments are implemented using MATLAB.

1 Framework of EM Algorithm for GMM model

1.1 Notations

In this section, I will derive the mathematical solution for Gaussian Mixed Model using EM algorithm, that is, to give a general framework of how EM algorithm serves GMM model.

Suppose we have observations $X = \{x^{(1)}, x^{(2)}, \ldots, x^{(N)}\}$, where $x^{(n)} = (x_1^{(n)}, x_2^{(n)}, \ldots, x_d^{(n)})^T$. Then, the Gaussian mixture distribution for each observation $x^{(n)}$ can be written as

$$p(x^{(n)}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x^{(n)}|\mu_k, \Sigma_k) \tag{1}$$

Let $Z = (z^{(1)}, z^{(2)}, \ldots, z^{(N)})$. The corresponding latent variable $z^{(n)}$ is a K-dimensional binary random variable which satisfies $z_k^{(n)} \in \{0, 1\}$ and $\sum_k z_k^{(n)} = 1$, where K denotes the number of Gaussians. The corresponding mixing coefficients π_k specifies the marginal distribution over $z^{(n)}$, such that

$$p(z_k^{(n)} = 1) = \pi_k \tag{2}$$

where the parameters $\{\pi_k\}$ must satisfy

$$0 \leq \pi_k \leq 1 \tag{3}$$

together with

$$\sum_{k=1}^{K} \pi_k = 1 \tag{4}$$

Because $z^{(n)}$ uses a 1-of-K representation, we can also write the distribution of $z^{(n)}$ as

$$p(z^{(n)}|\pi) = \prod_{k=1}^{K} \pi_k^{z_k^{(n)}} \tag{5}$$

1.2 Training Objective

Therefore, the **in-complete likelihood** is given by

$$p(X|\pi,\mu,\Sigma) = \prod_{n=1}^{N}\sum_{k=1}^{K}\pi_k\mathcal{N}(x^{(n)}|\mu_k,\Sigma_k) \tag{6}$$

And the corresponding **in-complete log-likelihood** is

$$\ln p(X|\pi,\mu,\Sigma) = \sum_{n=1}^{N}\sum_{k=1}^{K}\pi_k\mathcal{N}(x^{(n)}|\mu_k,\Sigma_k) \tag{7}$$

Note that a summation over k is inside the logarithm term which makes it hard to derive the solution. Thus we choose to maximize the much more simple but identical **complete likelihood**

$$p(X,Z|\pi,\mu,\Sigma) = \prod_{n=1}^{N}\prod_{k=1}^{K}[\pi_k\mathcal{N}(x^{(n)}|\mu_k,\Sigma_k)]^{z_k^{(n)}} \tag{8}$$

That is, to maximize the **complete log-likelihood**

$$\ln p(X,Z|\pi,\mu,\Sigma) = \sum_{n=1}^{N}\sum_{k=1}^{K}z_k^{(n)}[\ln\pi_k + \ln\mathcal{N}(x^{(n)}|\mu_k,\Sigma_k)] \tag{9}$$

According to the analysis above, **the objective of the EM algorithm is to maximize the complete log-likelihood**.

1.3 E-step

Given $\pi^{old},\mu^{old},\Sigma^{old}$, **the objective of the E-step is to take expectation of the complete log-likehood with respect to the latent variables** Z, i.e.,

$$E_Z[\ln p(X,Z|\pi,\mu,\Sigma)|X,\pi^{old},\mu^{old},\Sigma^{old}] = \sum_{n=1}^{N}\sum_{k=1}^{K}E_Z[z_k^{(n)}|x^{(n)},\pi^{old},\mu^{old},\Sigma^{old}][\ln\pi_k + \ln\mathcal{N}(x^{(n)}|\mu_k,\Sigma_k)] \tag{10}$$

And we denote

$$\begin{aligned}\gamma(z_k^{(n)}) \equiv E_Z[z_k^{(n)}|x^{(n)},\pi^{old},\mu^{old},\Sigma^{old}] &= \sum_{z_k^{(n)}} p(z_k^{(n)}|x^{(n)},\pi^{old},\mu^{old},\Sigma^{old}) \\ &= \sum_{z_k^{(n)}} z_k^{(n)}\frac{p(z_k^{(n)},x^{(n)}|\pi^{old},\mu^{old},\Sigma^{old}}{\sum_{k=1}^{K}p(z_k^{(n)}|x^{(n)},\pi^{old},\mu^{old},\Sigma^{old})} \\ &= \frac{\pi_k\mathcal{N}(x^{(n)}|\mu_k,\Sigma_k)}{\sum_{k=1}^{K}\mathcal{N}(x^{(n)}|\mu_k,\Sigma_k)}\end{aligned} \tag{11}$$

1.4 M-step

The objective of the M-step is to search for new parameters π, μ, Σ, which maximize Eqn.10, that is,

$$\begin{aligned} \max_{\pi,\mu,\Sigma} \quad & \sum_{n=1}^{N}\sum_{k=1}^{K}\gamma(z_k^{(n)})[\ln \pi_k + \ln \mathcal{N}(x^{(n)}|\mu_k, \Sigma_k)] \\ \textbf{s.t.} \quad & \sum_{k=1}^{K}\pi_k = 1 \end{aligned} \tag{12}$$

The *Lagrangian* function of the optimization problem 12 is given by

$$\mathcal{L} = \sum_{n=1}^{N}\sum_{k=1}^{K}\gamma(z_k^{(n)})[\ln \pi_k + \ln \mathcal{N}(x^{(n)}|\mu_k, \Sigma_k)] + \lambda(\sum_{k=1}^{K}\pi_k - 1) \tag{13}$$

where $\lambda > 0$

1.4.1 Solution for π

$$\frac{\partial \mathcal{L}}{\partial \pi_k} = \sum_{n=1}^{N}\frac{\gamma(z_k^{(n)})}{\pi_k} + \lambda \tag{14}$$

Let $\frac{\partial \mathcal{L}}{\partial \pi_k} = 0$, obtain

$$\pi_k = -\frac{1}{\lambda}\sum_{n=1}^{N}\gamma(z_k^{(n)}) \tag{15}$$

Sum over k and make use of $\sum_{k=1}^{K}\pi_k = 1$

$$1 = \sum_{k=1}^{K}\pi_k = -\frac{1}{\lambda}\sum_{k=1}^{K}\sum_{n=1}^{N}\gamma(z_k^{(n)}) \tag{16}$$

Therefore, $\lambda = -\sum_{k=1}^{K}\sum_{n=1}^{N}\gamma(z_k^{(n)})$, substitute into Eqn.15, and obtain the final solution

$$\pi_k^{new} = \frac{\sum_{n=1}^{N}\gamma(z_k^{(n)})}{\sum_{k=1}^{K}\sum_{n=1}^{N}\gamma(z_k^{(n)})} \tag{17}$$

1.4.2 Solution for μ

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial \mu_k} &= \sum_{n=1}^{N}\gamma(z_k^{(n)})\frac{\partial \ln \mathcal{N}(x^{(n)}|\mu_k, \Sigma_k)}{\partial \mu_k} \\ &= \sum_{n=1}^{N}\gamma(z_k^{(n)})\Sigma_k^{-1}(x^{(n)} - \mu_k) \\ &= \Sigma_k^{-1}\sum_{n=1}^{N}\gamma(z_k^{(n)})(x^{(n)} - \mu_k) \end{aligned} \tag{18}$$

where $\ln \mathcal{N}(\mu_k, \Sigma_k) = -\frac{1}{2}\ln|\Sigma_k| - \frac{1}{2}(x^{(n)} - \mu_k)^T\Sigma_k^{-1}(x^{(n)} - \mu_k) + const.$

Let $\frac{\partial \mathcal{L}}{\partial \mu_k} = 0$, then $\sum_{n=1}^{N}\gamma(z_k^{(n)})(x^{(n)} - \mu_k) = 0$, thus we obtain

$$\mu_k^{new} = \frac{\sum_{n=1}^{N}\gamma(z_k^{(n)})x^{(n)}}{\sum_{n=1}^{N}\gamma(z_k^{(n)})} \tag{19}$$

1.4.3 Solutions for Σ

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial \Sigma_k} &= \sum_{n=1}^{N} \gamma(z_k^{(n)}) \frac{\partial \ln \mathcal{N}(x^{(n)}|\mu_k, \Sigma_k)}{\partial \Sigma_k} \\ &= \sum_{n=1}^{N} \gamma(z_k^{(n)})(-\frac{1}{2})[I - \Sigma_k^{-1}(x^{(n)} - \mu_k)(x^{(n)} - \mu_k)^T]\Sigma_k^{-1}\end{aligned} \tag{20}$$

Let $\frac{\partial \mathcal{L}}{\partial \Sigma_k} = 0$, then $\sum_{n=1}^{N} \gamma(z_k^{(n)})\Sigma_k^{-1} = -\sum_{n=1}^{N} \gamma(z_k^{(n)})(x^{(n)} - \mu_k)(x^{(n)} - \mu_k)^T\Sigma_k^{-1}$.

Hence, we obtain

$$\Sigma_k^{new} = \frac{\sum_{n=1}^{N} \gamma(z_k^{(n)})(x^{(n)} - \mu_k)(x^{(n)} - \mu_k)^T}{\sum_{n=1}^{N} \gamma(z_k^{(n)})} \tag{21}$$

1.5 Framework of EM for Gaussian Mixtures

According to derivation above, we now conclude the framework of EM for Gaussian Mixtures:

1. Initialize the means μ_k, covariances Σ_k and mixing coefficients π_k, evaluate the initial value of the in-complete data log-likelihood.

2. **E-step**. Evaluate the responsibilities using the current parameters $\pi^{old}, \mu^{old}, \Sigma^{old}$

$$\gamma(z_k^{(n)}) = \frac{\pi_k^{old}\mathcal{N}(x^{(n)}|\mu_k^{old}, \Sigma_k^{old})}{\sum_{k=1}^{K} \mathcal{N}(x^{(n)}|\mu_k^{old}, \Sigma_k^{old})} \tag{22}$$

3. **M-step** Re-estimate the parameters using Eqn.22

$$\pi_k^{new} = \frac{\sum_{n=1}^{N} \gamma(z_k^{(n)})}{\sum_{k=1}^{K}\sum_{n=1}^{N} \gamma(z_k^{(n)})} \tag{23}$$

$$\mu_k^{new} = \frac{\sum_{n=1}^{N} \gamma(z_k^{(n)})x^{(n)}}{\sum_{n=1}^{N} \gamma(z_k^{(n)})} \tag{24}$$

$$\Sigma_k^{new} = \frac{\sum_{n=1}^{N} \gamma(z_k^{(n)})(x^{(n)} - \mu_k^{new})(x^{(n)} - \mu_k^{new})^T}{\sum_{n=1}^{N} \gamma(z_k^{(n)})} \tag{25}$$

4. Evaluate the in-complete log-likelihood

$$\ln p(X|\pi^{new}, \mu^{new}, \Sigma^{new}) = \sum_{n=1}^{N}\sum_{k=1}^{K} \pi_k^{new}\mathcal{N}(x^{(n)}|\mu_k^{new}, \Sigma_k^{new}) \tag{26}$$

5. Check for convergence. If the convergence criterion is not satisfied, return to step 2.

2 Solutions to the Assignment

μ_k is initialized by standard Gaussian, Σ_k is initialized proportional to identity matrix, π is initialized by $Uniform(0, 1)$ and then normalized to satisfy $\sum_{k=1}^{4} = 1$ We first show the scatter plot of the dataset in Fig.1.

2.1 Solutions to Problem 1)

The in-complete data log-likelihood during training is shown in Fig.2. Note that the training converges after 35 iterations.

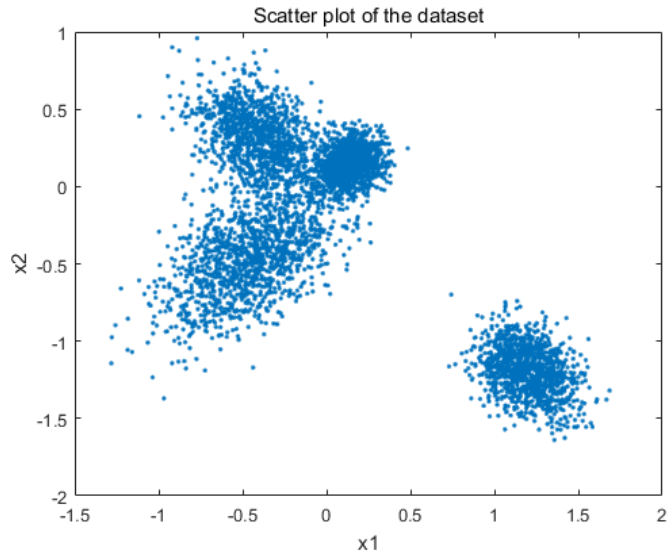

Figure 1: Scatter plot of the dataset

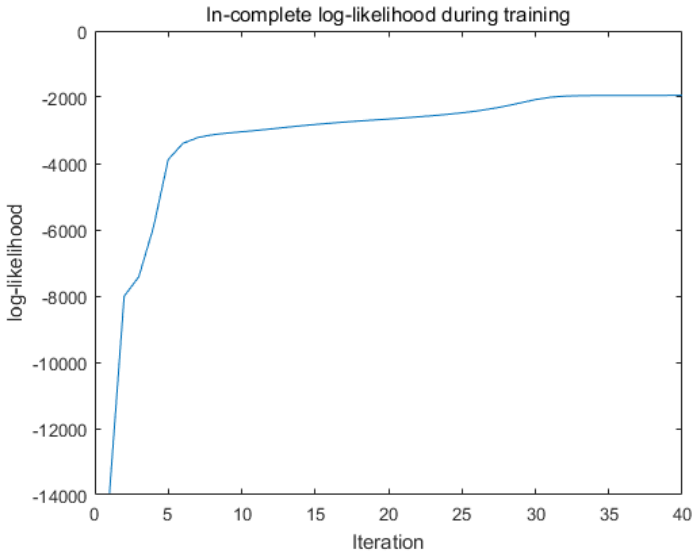

Figure 2: Plot of the in-complete log-likelihood $\ln p(X|\pi, \mu, \Sigma)$

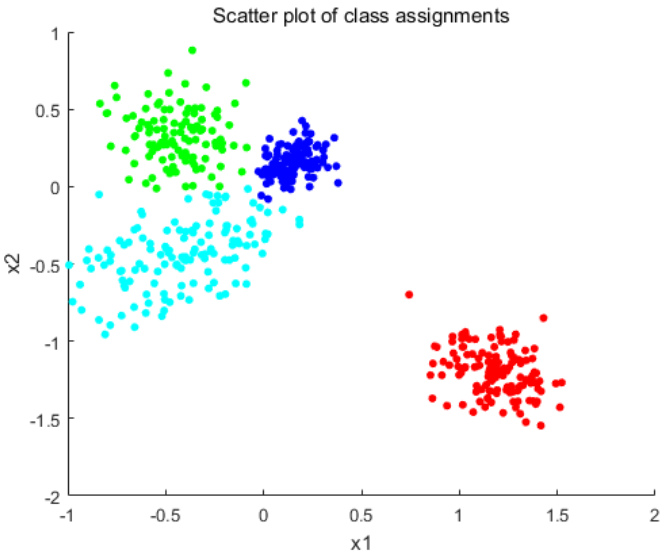

Figure 3: The scatter plot of $\gamma(z^n)$ after training converges

2.2 Solutions to Problem 2)

The learned parameters are:

$$\pi = (0.2300, 0.2287, 0.2731, 0.2682)^T$$

$$\mu_1 = \begin{pmatrix} -0.5231 \\ 0.1750 \end{pmatrix}, \mu_2 = \begin{pmatrix} -0.0511 \\ 0.2182 \end{pmatrix}, \mu_3 = \begin{pmatrix} -0.3396 \\ -0.3706 \end{pmatrix}, \mu_4 = \begin{pmatrix} 1.2035 \\ -1.1969 \end{pmatrix}$$

$$\Sigma_1 = \begin{pmatrix} 0.0203 & 0.0088 \\ 0.0088 & 0.1418 \end{pmatrix}, \Sigma_2 = \begin{pmatrix} 0.0813 & -0.0251 \\ -0.0251 & 0.0192 \end{pmatrix},$$

$$\Sigma_3 = \begin{pmatrix} 0.1011 & 0.0773 \\ 0.0773 & 0.1003 \end{pmatrix}, \Sigma_4 = \begin{pmatrix} 0.0226 & -0.0076 \\ -0.0076 & 0.0236 \end{pmatrix}$$

2.3 Solutions to Problem 3)

The scatter plot of $\gamma(z^n)$ after training converges is depicted in Fig.3.

2.4 Solutions to Problem 4)

If we initialize μ, Σ, π by other values, then the corresponding results are displayed in Fig.4.

Note that under two experiment settings, the training process converges at different speed, but the convergence seems the same.

Although it seems no obvious difference in the figures above, we can observe the variation through the values of μ. Here we show the values of μ from two experiments.

- Experiment 1

$$\mu_1 = \begin{pmatrix} -0.5231 \\ 0.1750 \end{pmatrix}, \mu_2 = \begin{pmatrix} -0.0511 \\ 0.2182 \end{pmatrix}, \mu_3 = \begin{pmatrix} -0.3396 \\ -0.3706 \end{pmatrix}, \mu_4 = \begin{pmatrix} 1.2035 \\ -1.1969 \end{pmatrix}$$

- Experiment 2

$$\mu_1 = \begin{pmatrix} -0.4423 \\ 0.4523 \end{pmatrix}, \mu_2 = \begin{pmatrix} -0.4658 \\ 0.3215 \end{pmatrix}, \mu_3 = \begin{pmatrix} 0.1444 \\ 0.1461 \end{pmatrix}, \mu_4 = \begin{pmatrix} 1.2035 \\ -1.1969 \end{pmatrix}$$

Note that μ_4 is the same in two experiments as group 4 departs clearly from other groups, which is consistent with the figures above. However, μ_1, μ_2, μ_3 are quite different as group $1 \sim 3$ mixes together

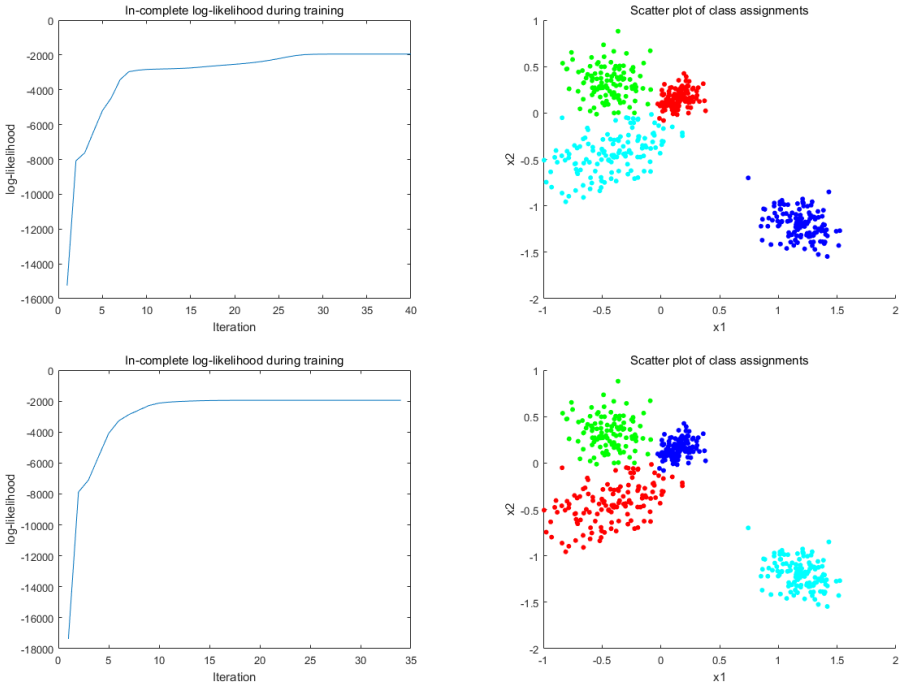

Figure 4: log-likelihood and class assignments for two experiments

and thus hard to classify.

3 Conclusions

EM algorithm can be used to derive the solution for Gaussian Mixture Models. Furthermore, the algorithm is quite sensitive to the initialization of the parameters.