EM Algorithms

Yinan Huang

Oct 25, 2020

Contents

1 Introduction

MLE (Maximum likelihood expectation) is a widely used method for paramter estimation. But for mixture model with latent variables, usually MLE does not have a closed form solution. EM (Expectation-Maximization) algorithms is an iterative method that converges to local maximum of likelihood. Tough local maximum can be found by gradient descent, but for model with latent variables, EM is more efficient.

2 Mixture model

A **Mixture model** is a probabilistic model that random variables can be from certain distribution class with different parameters with different probability. Concretely, let X be sample from mixture model, and $X \sim p(x|\theta_i)$ with probability π_i:

$$p(x|\theta, \pi) = \sum_i p(x|\theta_i)\pi_i. \tag{1}$$

We use $\Theta = (\theta, \pi)$ to represent all the paramters. And we can define **latent variables** Z, for which $Z = i$ represents $X \sim p(x|\theta_i)$. Then the distribution of X can be written in terms of z:

$$p(x|\Theta) = \sum_z p(x|z, \Theta)p(z|\Theta) = \sum_z p(x, z|\Theta), \tag{2}$$

where by definition $p(x|\theta_z) = p(x|z, \Theta)$ and $p(z|\Theta) = \pi_z$. In the problem of paramter estimation, we are asked to estimate Θ (including θ and π) given data $X = (x_1, ..., x_N)$. Note that we are not assuming $Z = (z_1, ..., z_N)$ is known (actually we assume the data of latent variables are missing or unobversable). Data missing of latent variable makes our estimation difficult.

2.1 MLE of complete data

We first assume our data is complete, i.e. (X, Z) is given. Then we will see that MLE of this mixture model is the same as MLE of pure model, which is reasonable since we have already known which distribution each sample comes from.

We begin with the log likelihood of complete data:

$$\ln p(X, Z|\Theta) = \sum_n^N \ln p(x_n, z_n|\Theta) = \sum_n^N \ln \left(p(x_n|z_n, \Theta)p(z_n|\Theta)\right) = \sum_n^N \ln p(x_n|\theta_{z_n}) + \sum_n^N \ln \pi_{z_n}. \tag{3}$$

In the last step, we see that π is seperated from θ, so they can be maximized seperatly (as expected):

$$\begin{aligned} \hat{\theta} &= \arg\max_\theta \sum_n^N \ln p(x_n|\theta_{z_n}), \\ \hat{\pi} &= \arg\max_\pi \sum_n^N \ln \pi_{z_n}, \quad \sum_i \pi_i = 1. \end{aligned} \tag{4}$$

Note that we can classify X into different distribution based on Z:

$$\sum_n^N \ln p(x_n|\theta_{z_n}) = \sum_i \sum_{n|z_n=i}^{N_i} \ln p(x_n|\theta_i) \quad \Rightarrow \quad \hat{\theta}_i = \arg\max_{\theta_i} \sum_{n|z_n=i}^{N_i} \ln p(x_n|\theta_i). \tag{5}$$

As we expected, for estimation of θ, we can estimate θ_i seperately, since we knew Z already. This estimation is only based on single model, so it is much easier compared to the case of missing data we are going to discuss later.

The estimation of π is also simple. We can use Lagrange Multiplier Method to calculate maximum of $\sum_n^N \ln \pi_{z_n}$ under constraint $\sum_i \pi_i = 1$. Then let derivative of π_j equals zero, getting

$$\hat{\pi}_j = \sum_{n|z_n=j}^{N_j} \frac{1}{\pi_j} + \lambda = \frac{N_j}{\pi_j} + \lambda = 0. \quad \Rightarrow \quad \lambda = -N. \tag{6}$$

So the estimation of π_j, as expected, is the fraction of data in j^{th} distribution:

$$\hat{\pi}_j = \frac{N_j}{N}. \tag{7}$$

We saw that if the complete data of mixture model are given, we can treat them just like the single model. For example, considering Gaussian mixture model, as long as we can solve single Gaussian model, then we can solve Gaussian mixture model given the complete data.

2.2 MLE of Missing data

But in real world latent variables Z might not be given (and this is why they are called "latent"). Only given X, we need to do the MLE of log likelihood of incomplete data:

$$\hat{\Theta} = \arg\max_{\Theta} \ln p(X|\Theta). \tag{8}$$

This time we do not know Z. We can expand $p(X|\Theta)$ in jointly probability $p(X, Z|\Theta)$:

$$\ln p(X|\Theta) = \sum_n^N \ln \sum_z p(x_n, z|\Theta) = \sum_n^N \ln \sum_z p(x_n|\theta_z)\pi_z. \tag{9}$$

Note that we cannot seperate π and θ directly since there is a sum in log. So in this case optimization of log likelihood will be complicated. For example, we can calculate MLE estimation of π:

$$\frac{\partial}{\partial \pi_j}\left[\ln p(X|\Theta) + \lambda\left(\sum_i \pi_i - 1\right)\right] = 0, \quad \Rightarrow \quad \sum_n^N \frac{p(x_n|\theta_j)}{\sum_z \hat{\pi}_z p(x_n|\theta_z)} + \lambda = 0. \tag{10}$$

Using $\sum_i \pi_i = 1$ we can find $\lambda = -N$. So there are several equations that we should solve to obtain $\hat{\pi}$. We can recast $\hat{\pi}_j$ in terms of posterior distribution of Z. Note that

$$\begin{aligned} 0 &= \sum_n^N \frac{p(x_n|\theta_j)}{\sum_z \hat{\pi}_z p(x_n|\theta_z)} + \lambda = \frac{1}{\hat{\pi}_j}\sum_n \frac{\hat{\pi}_j p(x_n|\theta_j)}{\sum_z \hat{\pi}_z p(x_n|\theta_z)} - N \\ &= \frac{1}{\hat{\pi}_j}\sum_n \frac{p(z_n = j|\Theta)p(x_n|z_n = j, \Theta)}{\sum_z p(z|\Theta)p(x_n|z, \Theta)} - N = \frac{1}{\hat{\pi}_j}\sum_n \frac{p(x_n, z_n = j|\Theta)}{p(x_n|\Theta)} - N \\ &= \frac{1}{\hat{\pi}_j}\sum_n p(z_n = j|x_n, \Theta) - N, \end{aligned} \tag{11}$$

so

$$\hat{\pi}_j = \frac{\sum_n^N p(z_n = j|x_n, \Theta)}{N}. \tag{12}$$

Note that we do not obtain a closed-form solution for $\hat{\pi}_j$ here, since the posterior distribution $p(z_n = j|x_n, \Theta)$ depends on prior distribution π (but we do not know yet). These equations basically say the MLE estimation of π are $\hat{\pi}$ such that the posterior and the prior distribution of Z are the same.

3 EM Algorithm for missing data

We saw from the previous section that in the case of missing data, MLE of Θ requires us to solve complex equations. Particularly, MLE of $\hat{\pi}$ should satisfy that the posterior and prior distribution of Z is the same:

$$\hat{\pi}_j = \frac{1}{N}\sum_{n}^{N} p(z_n = j|x_n, \hat{\pi}, \hat{\theta}). \tag{13}$$

Though we can solve it directly, but a naive idea is to use certain iteration algorithm to approch to the MLE. For example, we first make a naive guess of $\Theta = \Theta^{\text{odd}}$, and calculate the posterior distribution of Z, and we use this posterior distribution to improve our guess. The concrete algorithm is as follows:

EM Algorithm: To maximize $\ln p(X|\Theta)$ with latent variables Z whose distribution determined by $p(z = j) = \pi_j$, we initialize $\Theta = \Theta^{(0)}$ and do the following iteration:

(1) **E-step:** Compute posterior distribution of Z,

$$P(z_n = j|x_n, \Theta^{(i)}) = \frac{\pi_j^{(i)} p(x_n|\theta_j^{(i)})}{\sum_k \pi_k^{(i)} p(x_n|\theta_k^{(i)})}.$$

(2) **M-step:** Calculate $\Theta^{(i+1)}$ so that it maximizes the expectation of log likelihood of complete data based on posterior distribution of Z:

$$\Theta^{(i+1)} = \arg\max_{\Theta} \mathbb{E}_Z \ln p(X, Z|\Theta) = \arg\max_{\Theta} \sum_{Z} \ln p(X, Z|\Theta) p(Z|X, \Theta^{(i)}).$$

(3) Repeat step $(1), (2)$ until $\Theta^{(i+1)} \approx \Theta^{(i)}$.

Proof of convergence: To prove EM algorithm converges to MLE, we need to show that $\ln p(X|\Theta^{\text{new}}) \geq \ln p(X|\Theta^{\text{odd}})$. Note that

$$\arg\max_{\Theta} \sum_{Z} \ln p(X, Z|\Theta) p(Z|X, \Theta^{\text{odd}}) = \arg\max_{\Theta} \sum_{Z} \ln \frac{p(X, Z|\Theta)}{p(Z|X, \Theta^{(\text{odd})})} p(Z|X, \Theta^{\text{odd}}),$$

because we simply add a constant unrelated with Θ. We define this as

$$Q(\Theta, \Theta^{\text{odd}}) \equiv \sum_{Z} \ln \frac{p(X, Z|\Theta)}{p(Z|X, \Theta^{(\text{odd})})} p(Z|X, \Theta^{\text{odd}}), \quad \Theta^{\text{new}} = \arg\max_{\Theta} Q(\Theta, \Theta^{\text{odd}}).$$

We are going to show that indeed $\ln p(X|\Theta) \geq Q(\Theta, \Theta^{\text{odd}})$. Consider

$$\begin{aligned} \ln p(X|\Theta) &= \ln \sum_{Z} p(X, Z|\Theta) = \ln \sum_{Z} p(Z|X, \Theta^{\text{odd}}) \frac{p(X, Z|\Theta)}{p(Z|X, \Theta^{\text{odd}})} \\ &\geq \sum_{Z} p(Z|X, \Theta^{\text{odd}}) \ln \frac{p(X, Z|\Theta)}{p(Z|X, \Theta^{\text{odd}})} = Q(\Theta, \Theta^{\text{odd}}), \end{aligned}$$

where we use Jensen inequality in the last second step. Equality holds if $\ln p(X|\Theta^{\text{odd}}) = Q(\Theta^{\text{odd}}, \Theta^{\text{odd}})$. So we can see that when we are currently at Θ^{odd}, $Q(\Theta, \Theta^{\text{odd}})$ is a function of Θ that always lies under $\ln p(X|\Theta)$, and equals to $\ln p(X|\Theta)$ if $\Theta = \Theta^{\text{odd}}$. So we can imagine that if we find a larger $Q(\Theta, \Theta^{\text{odd}})$, then we find a larger $\ln p(X|\Theta)$. Concretely, let $\Theta^{\text{new}} = \arg\max_{\Theta} Q(\Theta, \Theta^{\text{odd}})$, then

$$\ln p(X|\Theta^{\text{new}}) \geq Q(\Theta^{\text{new}}, \Theta^{\text{odd}}) = \max_{\Theta} Q(\Theta, \Theta^{\text{odd}}) \geq Q(\Theta^{\text{odd}}, \Theta^{\text{odd}}) = \ln p(X|\Theta^{\text{odd}}). \quad (14)$$

Therefore $\ln p(X|\Theta)$ will always get larger and reach a local maximum. EM algorithm makes estimation simpler, because in some cases maximizing log likelihood of complete data has closed form solution, which allows us to converge fastly without too many iterations.