A short notes on EM algorithm and gradient descent

Tong Zhou*

October 6, 2020

This notes is based on the following materials:

– *All of Statistics: A consice course in Statistical Inference*, by *Larry Wasserman*, Page 143 - 146.

– Lecture notes from the course STAT 535: Statistical Machine Learning: lecture 13: EM Algorithm and Gradient Descent, by *Yen-chi Chen*.

Maximum likelihood estimation is one of the mostly used estimation methods. As long as we know the density function $f(x;\theta)$ that data come from, θ can be estimated by maximizing the log-likelihood function. However, $f(x;\theta)$ sometimes is hard to maximize owing to its complexity or impracticality. But suppose adding another random variable Z makes the log-likelihood of $f(x,z;\theta)$ easy to maximize, such that $f(x;\theta) = \int f(x,z;\theta)\,\mathrm{d}z$. The problem is that Z is often unobserved, which is usually called hidden variable, missing data or heterogeneity depending on specific contexts. Therefore, the first step, before implementing the usual maximization, is to "fill in" the missing data (or latent variable, unobserved heterogeneity).

Notations:

– $\boldsymbol{x}$: the complete data (including the latent variables)

– $\boldsymbol{y}$: the observed data.

– $L(\theta|\boldsymbol{x}) = p(\boldsymbol{x};\theta)$: the likelihood function (on the complete data).

EM ALGORITHM

- **(0)** Pick a starting value θ^0. Then for $n = 1, 2, \cdots$ repeat the E-step and M-step below:

- **E-step:** evaluate $Q\left(\theta;\theta^{(n)}|\boldsymbol{y}\right) = \mathbf{E}[\log L(\theta|\boldsymbol{x})|\boldsymbol{y};\theta^{(n)}]$,

- **M-step:** update $\theta^{(n+1)} = \arg\max_\theta Q(\theta;\theta^{(n)}|\boldsymbol{y})$,

*Johns Hopkins University, tzhou11@jhu.edu

until certain criterion is met (e.g., $\|\theta^{(n-1)} - \theta^{(n)}\|_\infty < \varepsilon$).

Note that in the E-step, the expectation means that $\boldsymbol{x}$ is from the distribution $p(\cdot; \theta^{(n)})$ conditional on $\boldsymbol{y}$!

> **Remark**
>
> 1. The EM algorithm does not guarantee the likelihood value is always increasing.
> 2. The EM algorithm does not guarantee to find the global maximizer.
> 3. Only certain MLEs can be found by the EM algorithm; not all MLE can be found using the EM.
> 4. The critical points problem.

Example 1 (Mixture of Normals)**.** *Think of heights of people being a mixture of men and women's heights. Let $\phi(y; \mu, \sigma)$ denote a normal density with mean μ and standard deviation σ. The density of a mixture of two Normals is*

$$f(y; \theta) = (1 - p)\phi\left(y; \mu_0, \sigma_0\right) + p\phi(y : \mu_1, \sigma_1). \tag{1}$$

The idea is that an observation is drawn from the first normal with probability p and the second with probability $1 - p$. However, we don't know which Normal it was drawn from. The parameters are $\theta = (\mu_0, \sigma_0, \mu_1, \sigma_1, p)$. The likelihood function is

$$\mathcal{L}(\theta|y) = \prod_{i=1}^{n} \left[(1 - p)\phi(y_i; \mu_0, \sigma_0) + p\phi(y_i; \mu_1, \sigma_1)\right]. \tag{2}$$

Maximizing this function over the five parameters is hard!. Now imaging we were given extra information telling us which of the two normals every observation came from, i.e. the data $\{(Y_i, Z_i)\}_{i=1}^n$ is "complete" data, where $Z_i = 0$ represents the first normal and $Z_i = 1$ represents the second. Note that $\mathbf{P}(Z_i = 1) = p$. Show that the likelihood for the complete *data $(Y_1, Z_1), \cdots, (Y_n, Z_n)$ is much simpler than that for the observed data $Y_1, \cdots, Y_n$.*

The likelihood function for the complete data is

$$\mathcal{L}(\theta|y, z) = \prod_{i=1}^{n} \left[(1 - p)\phi(y_i; \mu_0, \sigma_0)\right]^{1-Z_i} \bullet \left[p\phi(y_i; \mu_1, \sigma_1)\right]^{Z_i}. \tag{3}$$

In practice, however, $Z_1, \cdots, Z_n$ may not be observed. Folks often call Z is a latent variable in this model.

For simplicity, suppose $p = \frac{1}{2}$ and $\sigma_1 = \sigma_2 = 1$, which means the only unknown parameters are (μ_0, μ_1). Then the likelihood function becomes:

$$\mathcal{L}(\mu_0, \mu_1|y, z) = \frac{1}{2} \prod_{i=1}^{n} \phi(y_i; \mu_0, 1)^{1-Z_i} \cdot \phi(y_i; \mu_1, 1)^{Z_i}.$$

The log-likelihood of $\mathscr{L}(\mu_0, \mu_1|y, z)$ is:

$$\ell(\mu_0, \mu_1|y, z) = \log(\mathscr{L}(\mu_0, \mu_1|y, z)) = -\frac{1}{2}\sum_{i=1}^{n}(1 - Z_i)(y_i - \mu_0)^2 - \frac{1}{2}\sum_{i=1}^{n} Z_i(y_i - \mu_1)^2.$$

To implement the EM algorithm, first picking up an initial value $\theta^{(0)} = \left(\mu_0^{(0)}, \mu_1^{(0)}\right)$. Then proceed to the E-step:

$$\begin{aligned} Q(\theta, \theta^{(0)}|y) &= \mathbf{E}[\ell(\mu_0, \mu_1|y, z)|y, \theta^{(0)}] && (4)\\ &= \mathbf{E}_{\theta^{(0)}}[\ell(\mu_0, \mu_1|y, z)|y] && (5)\\ &= -\frac{1}{2}\sum_{i=1}^{n}\left(1 - \mathbf{E}_{\theta^{(0)}}[Z_i|y]\right)(y_i - \mu_0)^2 - \frac{1}{2}\sum_{i=1}^{n}\mathbf{E}_{\theta^{(0)}}[Z_i|y]\left(y_i - \mu_1\right)^2. && (6)\end{aligned}$$

Observe that Z is binary. So

$$\mathbf{E}_{\theta^{(0)}}[Z_i|y] = \mathbf{P}_{\theta^{(0)}}\left(Z_i = 1|y\right)$$

Note that Z_i is a discrete r.v., while Y is continuous. By the law of total probability,

$$f_Y(y) = \sum_{j\in\{0,1\}} f_{Y|Z_i=j}(y)\mathbf{P}(Z_i = j) = f_{Y|Z_i=1}(y)\mathbf{P}\left(Z_i = 1\right) + f_{Y|Z_i=0}(y)\mathbf{P}\left(Z_i = 0\right)$$

Then by Bayes rule:

$$\begin{aligned}\mathbf{P}_{\theta^{(0)}}\left(Z_i = 1|y\right) &= \frac{f_{Y=y|Z_i=1}(y)\mathbf{P}_{\theta^{(0)}}(Z_i = 1)}{f_Y(y)}\\ &= \frac{f_{Y=y|Z_i=1}(y)\mathbf{P}_{\theta^{(0)}}(Z_i = 1)}{f_{Y=y|Z_i=1}(y)\mathbf{P}_{\theta^{(0)}}(Z_i = 1) + f_{Y=y|Z_i=0}(y)\mathbf{P}_{\theta^{(0)}}(Z_i = 0)}\\ &= \frac{f_{Y_i=y_i|Z_i=1}(y_i)}{f_{Y_i=y_i|Z_i=1} + f_{Y_i=y_i|Z_i=0}}\\ &= \frac{\phi\left(y_i; \mu_1^{(0)}, 1\right)}{\phi\left(y_i; \mu_1^{(0)}, 1\right) + \phi\left(y_i; \mu_0^{(0)}, 1\right)}\\ &\equiv \tau(i)\end{aligned}$$

Plugging back to the E-step function $Q(\theta, \theta^{(0)}|y)$:

$$Q\left(\theta, \theta^{(0)}|y\right) = -\frac{1}{2}\sum_{i=1}^{n}(1 - \tau(i))(y_i - \mu_0)^2 - \frac{1}{2}\sum_{i=1}^{n}\tau_i(y_i - \mu_1)^2.$$

Taking first derivative over μ_1 and μ_0:

$$\mu_1^{(1)} = \frac{\sum_{i=1}^{n} \tau_i y_i}{\sum_{i=1}^{n} \tau_i};$$
$$\mu_0^{(1)} = \frac{\sum_{i=1}^{n} (1-\tau_i) y_i}{\sum_{i=1}^{n} (1-\tau_i)}$$

The iteration proceeds for $j = 2, \cdots$ until convergence criterion is hit.

Stochastic EM algorithm (Monte Carlo EM algorithm)

The motivation is that sometimes the E-step is complex and does not admit a closed form solution. That is, the $Q(\theta|\theta^{(i)})$ cannot be computed explicitly. A solution is by first evaluating the Q function by Monte Carlo methods.

Following the above notations, suppose we observe $X = (Y, Z)$, where we observe Y and Z is a latent variable. Then the MCEM algorithm can be implemented as follows:

- **E-Step:** On the $(i+1)$-th iteration, draw $z^1, \cdots, z^M$ from $f_{Z|Y,\Theta}(\cdot|y, \theta^{(i)})$. Approximate the Q-function as

$$\widetilde{Q}(\theta|\theta^{(i)}) = \frac{1}{M}\sum_{m=1}^{M} \log f_{X|\Theta}\left(\left(y, z^m\right)|\theta\right).$$

- **M-Step:** Maximize the approximate $\widetilde{Q}(\theta|\theta^{(i)})$ and put $\theta^{(i+1)}$ as the maximizer.