STAT 383C: Statistical Modeling I **Fall 2016**

Lecture 18: October 29

Lecturer: Purnamrita Sarkar *Scribes: Zijian Zeng*

Note: *LaTeX template courtesy of UC Berkeley EECS dept.*

Disclaimer: *These scribe notes have been slightly proofread and may have typos etc.*

18.1 Expectation Maximization (EM)

18.1.1 Missing Data

In many practical situations, we do not have all the data originally tested. Imagine that we have iid observations from a Gaussian. n data is available, but m data points go missing. $y_1, y_2..., y_n, y_{n+1}...y_{n+m} \sim N\left(\mu, \sigma^2\right)$
After the data goes missing, we have only m data points left.
Set $y_{n+1}...y_{n+m}$ as missing; $y_1...y_n$ as observed. What is our μ and σ estimate?
We can use just the weighted sample mean to estimate the mean. $\hat{\mu} = \frac{\sum y_i + \bar{y} m}{n+m}$
This could be done using E-M. While this example is trivial, this is applicable for more difficult situations where the MLE is not trivial. Let us model the missing data as latent data, Z. Then we can use log likelihood.

$$l(y, z; \theta) = log \prod_i^n exp(-(y_i - \mu)^2/(2\sigma^2)) * \prod_{n+1}^{n+m} exp(-(z_i - \mu)^2/(2\sigma^2))$$

$$l(y, z; \theta) = -\sum_i^n (y_i - \mu)^2/(2\sigma^2) - \sum_{n+1}^{n+m} -(z_f i - \mu)^2/(2\sigma^2)$$

$$l(y, z; \theta) = \frac{-\sum_i^n y_i^2 - \sum_{i=n+1}^{n+m} z_i^2}{2\sigma^2}$$

$$-\frac{(n+m)\mu^2}{2\sigma^2} + \frac{\mu(\sum_1^n y_i + \sum_{n+1}^{n+m} z_i)}{\sigma^2}$$

If we model this as E-M iterative estimation, our E step will be:

$$n\log(\sigma) - \frac{\sum^n y_i^2}{2\sigma^2} - (\frac{\sigma_t^2 + \mu_t^2}{2\sigma^2})m - (n+m)\frac{\mu^2}{2\sigma^2} + \frac{\mu}{\sigma^2}(\sum_i^n y_i + m\mu_t)\sigma^2$$

M step:

$$argmax_{\mu,\sigma} E[]$$

Putting it together in an iterative manner:

$$\mu_{t+1} = \frac{\sum^n y_i + \mu_t m}{n+m}$$
$$\sigma^2_{t+1} = \frac{\sum^n y_i^2 + (\sigma_t^2 + \mu_t^2)m}{n+m} - \mu^2_{t+1}$$

18.1.2 EM can be used for finding parameters

Multinomial with P: $(\frac{1}{2} + \frac{\theta}{2}, \frac{\theta}{2}, \frac{\theta}{2}, 1/2 - \theta)$ Introduce a latent variable, z

$$\begin{aligned} 1/2- &> z_i \\ \theta/2- &> y_1 - z_1 \\ \theta/2- &> y_2 \\ 1/2 - \theta- &> y_3 \end{aligned}$$

If you write down the E.M., it converges to MLE.

18.2 Light Bulb Example

Suppose the life expectancy of a light bulb is a known distribution. Our goal here is to estimate the parameter θ. So we do the following two experiments to collect data:

Experiment 1: $Y_1, Y_2, \cdots Y_n$ and $Y_i s$ are iid sample time for a light bulb to die.

Experiment 2: $E_1, E_2, \cdots E_n$ and $E_i s$ are iid where $E_i = 1$ if light bulb i is alive at time T. $(T < \theta)$

How do we estimate θ?

18.2.1 Exponential Distribution Case

Suppose the life expectancy of a light bulb has an exponential distribution Exp(θ).

18.2.1.1 Using EM

We introduce a latent variable $z_i : E_i = 1(z_i \geq T)$
We can remove y from the condition as z is independent of y due to iid.

$$E[z_i | y_{obs}, E, \theta^t] = E[z_i | E_i, \theta_t]$$

When $E_i = 1$, due to memoryless property of Exponential Distributions, we have:

$$E[z_i | E_i = 1, \theta_t] = T + \theta_t$$

When $E_i = 0$, by using the law of total expectation:, we have:

$$E[z_i|E_i = 0, \theta_t]p(E_i = 0; \theta) + E[z_i|E_i = 1, \theta_t]p(E_i = 1, \theta_t) = E[z_i; \theta_t] = \theta_t$$
$$E[z_i|E_i = 0, \theta_t](1 - e^{-T/\theta_t}) + (T + \theta_t)e^{-T/\theta_t} = \theta_t$$
$$E[z_i|E_i = 0, \theta_t] = \theta_t - \frac{Te^{-T/\theta_t}}{1 - e^{-T/\theta_t}}$$

Let $F = \sum E_i$, we have

$$\hat{\theta} = \frac{\sum Y_i + \sum Z_i}{n + m}$$
$$\theta_{t+1} = \frac{\sum Y_i + E[z_i|E_i = 1, \theta_t]F + E[z_i|E_i = 0, \theta_t](m - F)}{n + m}$$

18.2.1.2 Using MLE

$$Loglikelihood = -nlog\theta - \frac{\sum y_i}{\theta} - \sum E_i * T/\theta + (m - \sum E_i)log(1 - e^{-T/\theta})$$

Thus,

$$\hat{\theta} = \frac{\sum Y_i + \sum Z_i}{n + m}$$

The EM and MLE techniques converge for this example.

18.2.2 Uniform Distribution Case

Suppose the life expectancy of a light bulb has a uniform distribution Unif(0,θ)

18.2.2.1 Using EM

In order to use EM, we need to introduce latent variables $Z_i s$ where $E_i = \mathbb{1}_{Z_i \geq T}$.

E-step: Assume at least one $E_i = 1$, calculate the expectation of Z_i for given θ_t.

$$\mathbb{E}[Z_i|E_i = 1, \theta_t] = \frac{T + \theta_t}{2}$$
$$\mathbb{E}[Z_i|E_i = 0, \theta_t] = \frac{T}{2}$$

M-step: Maximize the conditional expectation of the log likelihood given Y_i and E_i.

$$l(\theta|Y_i, Z_i) = \log\left(\prod_{i=1}^{n} \frac{1}{\theta}\mathbb{1}_{Y_i \in (0,\theta]} \prod_{i=1}^{m} \frac{1}{\theta}\mathbb{1}_{Z_i \in (0,\theta]}\right) = -(n + m)\log\theta \mathbb{1}_{Y_{\max} \in (0,\theta]}\mathbb{1}_{Z_{\max} \in (0,\theta]}$$

$$\mathbb{E}_{Z_i}[\hat{\theta}] = \mathbb{E}_{Z_i}[\max(Y_{\max}, Z_{\max})] = \max(Y_{\max}, \mathbb{E}_{Z_i}[Z_{\max}]) = \max\left(Y_{\max}, \frac{T + \theta_t}{2}\right)$$

Combine E-step and M-step, we have:

$$\theta_{t+1} = \max\left(Y_{\max}, \frac{T+\theta_t}{2}\right)$$

18.2.2.2 Using MLE

$$L(\theta|Y_i, E_i) = \prod_{i=1}^{n} \frac{1}{\theta} \mathbb{1}_{Y_i \in (0,\theta]} \prod_{i=1}^{m} \left(1 - \frac{T}{\theta}\right)^{E_i} \left(\frac{T}{\theta}\right)^{1-E_i}$$

$$l(\theta|Y_i, E_i) = -n \log\theta + \sum_{i=1}^{m} E_i \cdot \log\left(1 - \frac{T}{\theta}\right) + \sum_{i=1}^{m} (1 - E_i) \cdot \log\left(\frac{T}{\theta}\right)$$

$$\frac{dl(\theta|Y_i, E_i)}{d\theta} = -\frac{n}{\theta} + \sum_{i=1}^{m} E_i \cdot \frac{T}{\theta(\theta - T)} + \sum_{i=1}^{m} (1 - E_i)\left(-\frac{1}{\theta}\right)$$

$$\hat{\theta} = \max\left(\frac{n+m}{n+m-\sum_{i=1}^{m} E_i} \cdot T, \quad Y_{\max}\right)$$

This estimator makes sense: it combines the usual MLE for uniform distribution with extra information we get from $E_i s$. The more $1s$ we observe from $E_i s$, the greater $\hat{\theta}$ is, but it cannot be greater than $\max\left(\frac{n+m}{n} \cdot T, \quad Y_{\max}\right)$.

18.2.3 What is wrong with EM here

It is easily seen that if we use the EM algorithm and start with any θ_0, this procedure will converge to $\hat{\theta}_{EM} = \max(Y_{\max}, T)$, which is obviously wrong. So what is the problem here?

It turns out, the reason for the apparent EM algorithm not resulting in the MLE is that the E-step is wrong. In the E-step, we are supposed to find the conditional expectation of likelihood function given $Y_i s$ and $E_i s$ at current parameter values. Now given the data with assumption that at least one $E_i = 1$, we have $\theta \geq T$ and hence the conditional distributions of Z_i are uniform in $[T, \theta_t]$. Thus for $\theta < \theta_t$ the conditional density of Z_i takes value 0 with positive probability and hence the conditional expected value of the likelihood we are seeking does not exist.

18.2.4 Nonapplicability of EM and The Generalized EM

Can we fix the EM by restricting the likelihood function here?

$$\mathbb{E}[l(\theta|Y_i, Z_i)] = \begin{cases} -\infty, & \text{if } \theta < \theta_t \\ -(n+m)\log\theta & \text{if } \theta \geq \theta_t \end{cases}$$

The answer is, sadly, no. From the log likelihood, we can see that when it is not $-\infty$, it is maximized when $\theta = \theta_t$. In other words, θ_t is always the maximum of the lower bound we are trying to maximize, so the EM algorithm will stuck at θ_t and not go anywhere. Therefore, the EM does not apply to this particular example.

It is useful to note the Generalized Expectation Maximization (GEM) algorithm where in M-step, it is not necessary to maximize the likelihood, but just to seek θ_{t+1} such that it leads to an increase in the expectation of conditional likelihood. It is often useful in cases where the maximization is difficult. Here we cannot increase the the lower bound, so GEM/EM does not work.