STA250 Lecture-13 Notes

Shuo Li

November 13, 2013

Recap:

- We saw that EM can be used to maximize certain forms of complicated likelihood.
- EM:

$$\theta^{(t+1)} = \underset{\theta}{argmax}\ Q(\theta|\theta^{(t)})$$

where

$$Q(\theta|\theta^{(t)}) = E[log\ P(Y_{obs}, Y_{mis}|Y_{obs}, \theta^{(t)})]$$
$$= \int log\ P(Y_{obs}, Y_{mis}|\theta) * P(Y_{mis}|Y_{obs}, \theta^{(t)}) dY_{mis}$$

Note: EM maximize $log\ P(Y_{obs}|\theta)$ by expanded log-likelihood $log\ P(Y_{obs}, Y_{mis}|\theta)$ where the observed data likelihood preserved
i.e.

$$\int P(Y_{obs}, Y_{mis}|\theta) dY_{mis} = P(Y_{obs}|\theta)$$

Two key points:

- Y_{mis} dos not have to correspond to "real" missing data
- The choice of Y_{mis} is not unique

Example:

- Model-1:
  $Y_{obs}|\theta \sim N(\theta, v+1)$
  Goal: find MLE for θ (answer is Y_{obs})
  No missing data!
  Consider a "complete" model s.t.

  $$Y_{obs}|Y_{mis} \sim N(Y_{mis}, 1)$$

  $$Y_{mis} \sim N(\theta, v)$$

  Need to check:

  $$\int P(Y_{obs}, Y_{mis}|\theta) dY_{mis} = P(Y_{obs}|\theta)$$

  We can show (standard result) that this is true here.
  Here Y_{mis} is not "real" missing data.

What is we instead used a different "complete" data model?

- Model-2:

  $$Y_{obs}|\widetilde{Y}_{mis}, \theta \sim N(\widetilde{Y}_{mis} + \theta, v)$$
  $$\widetilde{Y}_{mis} \sim N(0, 1)$$

  We can show that again:

  $$\int P(Y_{obs}, \widetilde{Y}_{mis}|\theta) d\widetilde{Y}_{mis} = P(Y_{obs}|\theta)$$

  For this "complete" data model, the EM algorithm is:

  $$\theta^{(t+1)} = (\frac{v}{v+1})\theta^{(t)} + (\frac{1}{v+1})Y_{obs}$$

We have tow EMs corresponding to two "complete" data models. Both give same MLE, which is better?

- M-1 has linear convergence rate $\dfrac{1}{v+1}$
- M-2 has linear convergence rate $\dfrac{v}{v+1}$

Lower is better, depend on v.

- M-1 is know as a sufficient augmentation scheme (Y_{mis} is a sufficient statistic for θ in the "complete" data model)
- M-2 is know as an ancillary augmentation scheme (Since $\widetilde{Y}_{mis}$ does not depend on θ)

It turns out that the EM algorithm has an important property: Monotone convergence.
i.e.

$$l(\theta^{(t+1)}) \geqslant l(\theta^{(t)})$$

where

$$l(\theta) = log\ P(Y_{obs}|\theta)$$

This makes EM very stable (& popular); N-R, Bisection, Scoring etc. do not have this property.

Proof:

Note:

$$P(Y_{obs}, Y_{mis}|\theta) = P(Y_{obs}|\theta)P(Y_{mis}|Y_{obs}, \theta)$$
$$\Rightarrow l_{obs}(\theta) = log\ P(Y_{obs}, Y_{mis}|\theta) - log\ P(Y_{mis}|Y_{obs}, \theta)$$

Integrate both sides w.r.t. $P(Y_{mis}|Y_{obs}, \theta^{(t)})$

$$l_{obs}(\theta) = Q(\theta|\theta^{(t)}) + H(\theta|\theta^{(t)})$$

where

$$H(\theta|\theta^{(t)}) = -\int log\ P(Y_{mis}|Y_{obs}, \theta)P(Y_{mis}|Y_{obs}, \theta^{(t)})dY_{mis}$$

So,

$$l_{obs}(\theta^{(t+1)}) - l_{obs}(\theta^{(t)}) = [Q(\theta^{(t+1)}|\theta^{(t)}) - Q(\theta^{(t)}|\theta^{(t)})] + [H(\theta^{(t+1)}|\theta^{(t)}) - H(\theta^{(t)}|\theta^{(t)})]$$

First term ΔQ is $\geqslant 0$ by definition of Q function. We only need to show $\Delta H = H(\theta^{(t+1)}|\theta^{(t)}) - H(\theta^{(t)}|\theta^{(t)}) \geqslant 0$

$$\Delta H = \int log\ (\frac{P(Y_{mis}|Y_{obs}, \theta^{(t)})}{P(Y_{mis}|Y_{obs}, \theta^{(t+1)})})P(Y_{mis}|Y_{obs}, \theta^{(t)})dY_{mis}$$

This is the KL divergence $KL(P(Y_{mis}|Y_{obs},\theta^{(t)}))||P(Y_{mis}|Y_{obs},\theta^{(t+1)})$
$\Rightarrow$ By properties of KL divergence $\Delta H \geqslant 0$ with $\Delta H = 0$
iff.

$$P(Y_{mis}|Y_{obs},\theta^{(t+1)}) = P(Y_{mis}|Y_{obs},\theta^{(t)})$$

Therefore,

$$l_{obs}(\theta^{(t+1)}) - l_{obs}(\theta^{(t)}) \geqslant 0$$

Aside:

We can also use EM to find posterior modes not just MLE's.

- To maximize $log\ P(\theta|Y_{obs})$,
  Let
  $$Q_{MAP}(\theta|\theta^{(t)}) = E[log\ P(\theta, Y_{mis}|Y_{obs})|Y_{obs},\theta^{(t)}]$$
  $$= \int log\ P(\theta, Y_{mis}|Y_{obs})P(Y_{mis}|Y_{obs},\theta^{(t)})dY_{mis}$$
- "MAP estimate" maximize a posterior value (i.e. posterior mode)

Example:

- Probit Regression
  $$Y_i|X_i \sim Bin(1, g(X_i^T\beta))$$

  For logistic regression: $g(u) = \dfrac{e^u}{1+e^u}$
  For probit regression: $g(u) = \Phi(u)$, CDF of $N(0,1)$
  Form a complete data model:

  $$Y_i|Z_i,\beta \sim 1_{\{z_i \geqslant 0\}}$$
  $$Z_i|\beta \sim N(X_i^T\beta, 1)$$

  Parameter: β
  Complete data: $\{(Y_i, Z_i), i = 1,2,...,n\}$
  Observed data: $\{(Y_i), i = 1,2,...,n\}$
  Missing data: $\{(Z_i), i = 1,2,...,n\}$

- Check:

$$\int P(Y_i, Z_i|\beta)dZ_i = P(Y_i|\beta)$$

$$P(Y_i = 1|\beta) = \int_{Z>0} \frac{1}{\sqrt{2\pi}} exp(-\frac{1}{2}(Z - X_i^T\beta)^2)dZ_i = \Phi(X_i^T\beta)$$

$\Rightarrow$ preserves observed data log-likelihood

Let's derive the EM algorithm for this model:

$$Q(\theta|\theta^{(t)}) = E[log\ P(Y_{obs}, Y_{mis}|\theta)|Y_{obs}, \theta^{(t)}]$$
$$Q(\beta|\beta^{(t)}) = E[log\ P(Y, Z|\beta)|Y, \beta^{(t)}]$$

Take the expectations, we need to know $Z_i|Y_i, \beta^{(t)}$

$$Z_i|Y_i = 0, \beta^{(t)} \sim TN(X_i^T\beta^{(t)}, 1, (-\infty, 0])$$
$$Z_i|Y_i = 1, \beta^{(t)} \sim TN(X_i^T\beta^{(t)}, 1, [0, +\infty))$$
$$Q(\beta|\beta^{(t)}) = -E[\frac{1}{2}(Z_i - X_i^T\beta)^2|Y, \beta^{(t)}]$$

$\Rightarrow$Maximizer of $Q(\beta|\beta^{(t)})$
We can show,
If $Y_i = 1$

$$Z_i^{(t+1)} = X_i^T\beta^{(t)} + \frac{\Phi(X_i^T\beta^{(t)})}{1 - \Phi(-X_i^T\beta^{(t)})}$$

If $Y_i = 0$

$$Z_i^{(t+1)} = X_i^T\beta^{(t)} + \frac{\Phi(X_i^T\beta^{(t)})}{\Phi(-X_i^T\beta^{(t)})}$$

The maximizer of $Q(\beta|\beta^{(t)})$ *w.r.t.* β is seen to be the LSE of β when regressing $Z^{(t+1)}$ on X.
i.e.

$$\beta^{(t+1)} = (X^TX)^{-1}X^TZ^{(t+1)}$$

where

$$Z^{(t+1)} = [\begin{matrix} {Z_1}^{(t+1)} \\ \vdots \\ {Z_n}^{(t+1)} \end{matrix}]$$

E-Step: Compute $Z^{(t+1)}$
M-Step: Compute $\beta^{(t+1)} = (X^T X)^{-1} X^T Z^{(t+1)}$