The Expectation-Maximization Algorithm

Kenneth Chu Study Notes January 28, 2013

1 The Expectation-Maximization Algorithm

The Expectation-Maximization (EM) Algorithm is an algorithm that solves the optimization (maximization) problem for a marginal likelihood (or probability):

$$L(\theta; X) \;=\; p(X \,|\, \theta) \;=\; \int p(X, Z \,|\, \theta)\, \mathrm{d}Z$$

More specifically, the EM Algorithm *attempts* to compute:

$$\widehat{\theta} \;:=\; \underset{\theta}{\operatorname{argmax}}\, \{\, L(\theta\,;\, X)\,\} \;=\; \underset{\theta}{\operatorname{argmax}}\, \{\, p(X \,|\, \theta)\,\} \;=\; \underset{\theta}{\operatorname{argmax}} \left\{ \int p(X, Z \,|\, \theta)\, \mathrm{d}Z \right\}$$

Here, $L(\theta; X, Z) = p(X, Z \,|\, \theta)$ is a likelihood, where θ is the random vector of model parameters, X is the (non-random) vector of observed data, and Z is the random vector of unobservable variables. In practice, the EM Algorithm should produce estimates of a local maximum $\widehat{\theta}$ of $L(\theta; X)$.

The Expectation-Maximization (EM) Algorithm

Choose (arbitrarily) an initial value θ_0 for θ. Choose (arbitrarily) a termination threshold $\tau > 0$. Generate the sequence $\{\,\theta_t\,\}$, for $t = 1, 2, 3, \ldots$, by iterating through the following two-step procedure:

1. **Expectation Step:** Compute the following expectation value (as a function of θ):

$$Q(\theta \,|\, \theta_t) \;:=\; E_{Z|X,\theta_t}\{\, \log L(\theta\,;\, X, Z)\,\} \;=\; \int \left[\log L(\theta\,;\, X, Z)\right] p(Z \,|\, X, \theta_t)\mathrm{d}Z \tag{1.1}$$

2. **Maximization Step:** Solve the following optimization (maximization) problem to obtain θ_{t+1}:

$$\theta_{t+1} \;:=\; \underset{\theta}{\operatorname{argmax}}\, \{\, Q(\theta \,|\, \theta_t)\,\} \tag{1.2}$$

Terminate the EM Algorithm when

$$\left| \frac{\log p(X \,|\, \theta_{t+1}) - \log p(X \,|\, \theta_t)}{\log p(X \,|\, \theta_t)} \right| \;\leq\; \tau \tag{1.3}$$

Remark 1.1 *We remark that the "Expectation Step" is really an integration "along the Z-direction," with respect to the measure $p(Z \,|\, X, \theta_t)\mathrm{d}Z$. This yields a function $Q(\theta \,|\, \theta_t)$ of θ. The "Maximization Step" then produces a (local) maximum $\widehat{\theta}$ of the function $Q(\theta \,|\, \theta_t)$.*

Theorem 1.2 *The sequence θ_1, θ_2, θ_3, ... produced by the EM Algorithm satisfies the following:*

$$\log p(X \,|\, \theta_{t+1}) \;\geq\; \log p(X \,|\, \theta_t), \quad \text{for each } \; t = 1, 2, 3, \ldots$$

Proof First, observe that:

$$\begin{aligned}\log p(X\,|\,\theta) &= \log\left(\frac{p(X,\theta)}{p(\theta)}\right) = \log\left(\frac{p(X,Z,\theta)}{p(\theta)}\frac{p(X,\theta)}{p(X,Z,\theta)}\right)\\ &= \log\left(p(X,Z\,|\,\theta)\right) - \log\left(p(Z\,|\,X,\theta)\right)\end{aligned}$$

Taking expectation on both sides with respect to $p(Z\,|\,X,\theta_t)\,\mathrm{d}Z$ yields:

$$\begin{aligned}E_{Z|X,\theta_t}\{\log p(X\,|\,\theta)\} &= E_{Z|X,\theta_t}\{\log\left(p(X,Z\,|\,\theta)\right)\} - E_{Z|X,\theta_t}\{\log\left(p(Z\,|\,X,\theta)\right)\}\\ \int\{\log p(X\,|\,\theta)\}\,p(Z\,|\,X,\theta_t)\,\mathrm{d}Z &= \int\{\log\left(p(X,Z\,|\,\theta)\right)\}\,p(Z\,|\,X,\theta_t)\,\mathrm{d}Z - \int\{\log\left(p(Z\,|\,X,\theta)\right)\}\,p(Z\,|\,X,\theta_t)\,\mathrm{d}Z\\ \log p(X\,|\,\theta) &= Q(\theta\,|\,\theta_t) + H(\theta\,|\,\theta_t)\end{aligned}$$

where $H(\theta\,|\,\theta_t)$ is defined as follows:

$$H(\theta\,|\,\theta_t) := -\int\{\log\left(p(Z\,|\,X,\theta)\right)\}\,p(Z\,|\,X,\theta_t)\,\mathrm{d}Z$$

CLAIM 1: $H(\theta\,|\,\theta_t) \geq H(\theta_t\,|\,\theta_t)$, for any θ.
Note that **CLAIM 1** is an immediate consequence of Gibb's Inequality (see Appendix).

Now, the following equation

$$\log p(X\,|\,\theta) = Q(\theta\,|\,\theta_t) + H(\theta\,|\,\theta_t) \tag{1.4}$$

holds for any value of θ; in particular, it holds for θ_t:

$$\log p(X\,|\,\theta_t) = Q(\theta_t\,|\,\theta_t) + H(\theta_t\,|\,\theta_t) \tag{1.5}$$

Subtracting Equation (1.5) from Equation (1.4) yields:

$$\log p(X\,|\,\theta) - \log p(X\,|\,\theta_t) = \left(\,Q(\theta\,|\,\theta_t) - Q(\theta_t\,|\,\theta_t)\,\right) + \left(\,H(\theta\,|\,\theta_t) - H(\theta_t\,|\,\theta_t)\,\right) \tag{1.6}$$

Thus, **CLAIM 1** implies:

$$\log p(X\,|\,\theta) - \log p(X\,|\,\theta_t) \geq Q(\theta\,|\,\theta_t) - Q(\theta_t\,|\,\theta_t) \tag{1.7}$$

Since, by definition, $\theta_{t+1} := \underset{\theta}{\operatorname{argmax}}\{\,Q(\theta\,|\,\theta_t)\,\}$, we therefore have:

$$\log p(X\,|\,\theta_{t+1}) - \log p(X\,|\,\theta_t) \geq Q(\theta_{t+1}\,|\,\theta_t) - Q(\theta_t\,|\,\theta_t) \geq 0 \tag{1.8}$$

This proves the Theorem. $\square$

A Gibbs' Inequality & Jensen's Inequality

Theorem A.1 (Jensen's Inequality)
Suppose

- $(\Omega, \mathcal{A}, \mu)$ *is a probability space (i.e. measure space with* $\mu(\Omega) = 1$*).*
- $\varphi : (a,b) \longrightarrow \mathbb{R}$ *is a convex function, i.e.*

  $$\varphi(t\,x_1 + (1-t)x_2) \leq t\,\varphi(x_1) + (1-t)\,\varphi(x_2), \quad \text{for any } t \in [0,1],\ x_1, x_2 \in (a,b),$$

  where $-\infty \leq a < b \leq \infty$*.*
- $g : \Omega \longrightarrow (a,b)$ *is a* μ*-integrable function.*

Then, the following inequality holds:

$$\varphi\left(\int_\Omega g\,\mathrm{d}\mu\right) \;\leq\; \int_\Omega \varphi \circ g\,\mathrm{d}\mu$$

Corollary A.2 (Jensen's Inequality (Expectation Form))
Suppose

- $X : (\Omega, \mathcal{A}, \mu) \longrightarrow (a,b)$ *is a* $\mathbb{R}$*-valued random variable defined on the probability space* $(\Omega, \mathcal{A}, \mu)$ *with range contained in the open interval* (a,b)*, where* $-\infty \leq a < b \leq \infty$*.*
- $\varphi : (a,b) \longrightarrow \mathbb{R}$ *is a convex function.*

Then, the following inequality holds:

$$\varphi\,(\,E[\,X\,]\,) \;\leq\; E[\,\varphi(X)\,]$$

Theorem A.3 (Gibbs' Inequality)
Suppose

- $(\Omega, \mathcal{A})$ *is a measurable space.*
- $f, g : \Omega \longrightarrow [0,\infty)$ *are two nowhere-vanishing probability density functions defined on* $(\Omega, \mathcal{A})$*.*

Then, the following inequality holds:

$$-\int_\Omega (\log f)\,f\,\mathrm{d}x \;\leq\; -\int_\Omega (\log g)\,f\,\mathrm{d}x$$

Proof First, note that $\varphi := -\log : (0,\infty) \longrightarrow \mathbb{R}$ is a convex function defined on the open unit interval $(0,1)$, and that the domain of φ contains the range of f and g. Hence, by Jensen's Inequality, we have:

$$\int_\Omega \left[-\log\left(\frac{g(x)}{f(x)}\right)\right] \cdot f(x)\,\mathrm{d}x \;\geq\; -\log\left(\int_\Omega \frac{g(x)}{f(x)} \cdot f(x)\,\mathrm{d}x\right) = -\log\left(\int_\Omega g(x)\,\mathrm{d}x\right) = -\log\,(\,1\,) = 0$$

The above inequality immediately implies:

$$-\int_\Omega (\log g(x)) \cdot f(x)\,\mathrm{d}x \;\geq\; -\int_\Omega (\log f(x)) \cdot f(x)\,\mathrm{d}x\,,$$

which completes the proof of Gibbs' Inequality. $\square$