Statistical Machine Learning EM algorithm

Lecture Notes 13: EM algorithm

Professor: Zhihua Zhang *Scribe:*

13 Expectation-Maximization Algorithm

13.1 Description and Examples

In statistics, an expectationmaximization (EM) algorithm is an iterative method for finding maximum likelihood or maximum a posteriori (MAP) estimates of parameters in statistical models, where the model depends on unobserved latent variables. The EM iteration alternates between performing an expectation (E) step, which creates a function for the expectation of the log-likelihood evaluated using the current estimate for the parameters, and a maximization (M) step, which computes parameters maximizing the expected log-likelihood found on the E step. These parameter-estimates are then used to determine the distribution of the latent variables in the next E step.

Given a statistical model consisting of a set X of observed data, a set of latent data or missing values Z and a vector of unknown parameters θ. Let $Y = (X, Z)$, called complete data. The maximum likelihood estimate (MLE) of the unknown parameters is determined by the marginal likelihood of the observed data

$$L(\theta; X) = p(X|\theta) = \sum_{Z} p(X, Z|\theta)$$

However, this quantity is often intractable (e.g. if Z is a sequence of events, so that the number of values grows exponentially with the sequence length, making the exact calculation of the sum extremely difficult).

The EM algorithm seeks to find the MLE of the marginal likelihood by iteratively applying the following two steps:

1. Expectation step (E step): Calculate the expected value of the log likelihood function, with respect to the conditional distribution of Z given X under the current estimate of the parameters $\theta^{(t)}$:

$$Q(\theta|\theta^{(t)}) = \mathrm{E}_{Z|X,\theta^{(t)}} \left[\log L(\theta; X, Z)\right]$$

2. Maximization step (M step): Find the parameter that maximizes this quantity:

$$\theta^{(t+1)} = \arg\max_{\theta} Q(\theta|\theta^{(t)})$$

until $|\theta^{(t+1)} - \theta^{(t)}| < \varepsilon$ or $|Q(\theta^{(t+1)}|\theta^{(t)}) - Q(\theta^{(t)}|\theta^{(t-1)})| < \varepsilon$.

Here we will show how EM algorithm comes from.

$$\begin{aligned}
& \log p(X|\theta) \\
= & \log \int p(X,Z|\theta)dz \\
= & \log \int \frac{p(X,Z|\theta)}{q(z)} q(z)dz \\
\geq & \int \log \frac{p(X,Z|\theta)}{q(z)} q(z)dz \quad \text{(Jensen's inequality)}
\end{aligned}$$

Let $q(z) = p(Z|X, \theta^{(t)})$,

$$\begin{aligned}
& \log p(X|\theta) \\
\geq & \int \log \frac{p(X,Z|\theta)}{p(Z|X,\theta^{(t)})} p(Z|X,\theta^{(t)})dz \\
= & \int \log p(X,Z|\theta)p(Z|X,\theta^{(t)})dz - \int \log p(Z|X,\theta^{(t)})p(Z|X,\theta^{(t)})dz
\end{aligned}$$

Since $\int \log p(Z|X,\theta^{(t)})p(Z|X,\theta^{(t)})dz$ is unrelated to θ, so maximize $p(X|\theta)$ is equivalent to maximize $\int \log p(X,Z|\theta)p(Z|X,\theta^{(t)})dz$. And

$$\begin{aligned}
& \int \log p(X,Z|\theta)p(Z|X,\theta^{(t)})dz \\
= & E[\log p(X,Z|\theta)|X,\theta^{(t)}]
\end{aligned}$$

The above expectation is called Q function.

Example 13.1 (Toy Example 1)**.** *Assume y_1, y_2 are iid and have p.d.f $\theta \exp(-y\theta)$. Observed $y_1 = 5$, y_2 is missing. Try to estimate θ and y_2.*

Solution.

1. Using MLE,
$$\log p(y_1|\theta) = \log \theta - y_1\theta$$
By taking derivative and let it equals zero, we can get $\theta = \frac{1}{y_1} = 0.2$. Then the estimation of y_2 is the expectation, $y_2 = \frac{1}{\theta} = 5$.

2. Using EM,
$$\begin{aligned}
& p(y_1, y_2|\theta) = \theta^2 \exp(-y_1\theta)\exp(-y_2\theta) \\
\Rightarrow \quad & \log p(y_1, y_2|\theta) = 2\log\theta - (y_1 + y_2)\theta
\end{aligned}$$

Hence,

$$
\begin{aligned}
Q(\theta|\theta^{(t)}) &= \int \log p(y_1, y_2|\theta) p(y_2|y_1, \theta^{(t)}) dy_2 \\
&= \int (2\log\theta - (y_1 + y_2)\theta)(\theta^{(t)} \exp(-y_2\theta^{(t)})) dy_2 \\
&= 2\log\theta - y_1\theta - \theta \int y_2\theta^{(t)} \exp(-y_2\theta^{(t)}) dy_2 \\
&= 2\log\theta - y_1\theta - \frac{\theta}{\theta^{(t)}}
\end{aligned}
$$

Then let $Q' = 0$,

$$
\begin{aligned}
\frac{2}{\theta^{(t+1)}} - y_1 - \frac{1}{\theta^{(t)}} = 0 \\
\Rightarrow \quad \theta^{(t+1)} = \frac{2\theta^{(t)}}{y_1\theta^{(t)} + 1}
\end{aligned}
$$

So, we if the iteration convergence, it convergence to 0.2 and $E(y_2) = \frac{1}{\theta^{(t)}} = 5$.

Example 13.2 (Toy example 2)**.** *Imagine you ask n kids to choose a toy out of four choices. Let $Y = [Y_1 \cdots Y_4]^T$. T denote the histogram of their n choices, where Y_i is the number of the kids that chose toy i, for $i = 1, \cdots, 4$. We can model this random histogram Y as being distributed according to a multinomial distribution. The multinomial has two parameters: the number of kids asked, denoted by $n \in N$, and the probability that a kid will choose each of the four toys, denoted by $p \in [0,1]^4$, where $p_1 + p_2 + p_3 + p_4 = 1$. Then the probability of seeing some particular histogram y is:*

$$P(y|p) = \frac{n!}{y_1!y_2!y_3!y_4!} p_1^{y_1} p_2^{y_2} p_3^{y_3} p_4^{y_4}.$$

Next, say that we have reason to believe that the unknown probability p of choosing each of the toys is parameterized by some hidden value $\theta \in (0,1)$ such that

$$p_\theta = \left[\frac{1}{2} + \frac{1}{4}\theta \quad \frac{1}{4}(1-\theta) \quad \frac{1}{4}(1-\theta) \quad \frac{1}{4}\theta\right]^T, \theta \in (0,1)$$

The estimation problem is to guess the θ that maximizes the probability of the observed histogram y of toy choices.

Solution. we can write the probability of seeing the histogram y as

$$P(y|\theta) = \frac{n!}{y_1!y_2!y_3!y_4!} \left(\frac{1}{2} + \frac{1}{4}\theta\right)^{y_1} \left(\frac{1}{4}(1-\theta)\right)^{y_2} \left(\frac{1}{4}(1-\theta)\right)^{y_3} \left(\frac{1}{4}\theta\right)^{y_4}.$$

For this simple example, one could directly maximize the log-likelihood $\log P(y|\theta)$, but here we will instead illustrate how to use the EM algorithm to find the maximum likelihood estimate of θ.

To use EM, we need to specify what the complete data X is. We will choose the complete data to enable us to specify the probability mass function (pmf) in terms of only θ and $1-\theta$. To that end, we define the complete data to be $X = [X_1 \cdots X_5]^T$, where X has a multinomial distribution with number of trials n and the probability of each event is:

$$q_\theta = \begin{bmatrix} \frac{1}{2} & \frac{1}{4}\theta & \frac{1}{4}(1-\theta) & \frac{1}{4}(1-\theta) & \frac{1}{4}\theta \end{bmatrix}^T, \ \theta \in (0,1)$$

By defining X this way, we can then write the observed data Y as:

$$Y = T(X) = [X_1 + X_2, X_3, X_4, X_5]^T.$$

The likelihood of a realization x of the complete data is

$$P(x|\theta) = \frac{n!}{x_1!x_2!x_3!x_4!x_5!} \left(\frac{1}{2}\right)^{x_1} \left(\frac{\theta}{4}\right)^{x_2+x_5} \left(\frac{1-\theta}{4}\right)^{x_3+x_4}$$

For EM, we need to maximize the Q-function:

$$\theta^{(t+1)} = \arg\max_\theta Q(\theta|\theta^{(t)}) = \arg\max_{\theta\in(0,1)} \mathrm{E}_{X|y,\theta^{(t)}}\left[\log p(X;\theta)\right]$$

To solve the above equation, we actually only need the terms of $\log p(x|\theta)$that depend on θ, because the other terms are irrelevant as far as maximizing over θ is concerned. Hence,

$$\begin{aligned} \theta^{(t+1)} &= \arg\max_\theta \mathrm{E}_{X|y,\theta^{(t)}}\left[(X_2+X_5)\log\theta + (X_3+X_4)\log(1-\theta)\right] \\ &= \arg\max_\theta \left(\mathrm{E}_{X|y,\theta^{(t)}}[X_2] + \mathrm{E}_{X|y,\theta^{(t)}}[X_5]\right)\log\theta + \left(\mathrm{E}_{X|y,\theta^{(t)}}[X_3] + \mathrm{E}_{X|y,\theta^{(t)}}[X_4]\right)\log(1-\theta) \end{aligned}$$

To solve the above maximization problem, we need the expectation of the complete data X conditioned on the already known incomplete data y, which only leaves the uncertainty about X_1 and X_2. Since we know that $X_1 + X_2 = y_1$, we can use the indicator function $\mathbf{1}_{\{\cdot\}}$ to write that given y_1, the pair $(X_1; X_2)$ is binomially distributed with X_1 ”successes” in y_1 events:

$$\begin{aligned} p(x|y,\theta^{(t)}) &= \frac{y_1}{x_1!x_2!}\left(\frac{\frac{1}{2}}{\frac{1}{2}+\frac{\theta^{(t)}}{4}}\right)^{x_1}\left(\frac{\frac{\theta^{(t)}}{4}}{\frac{1}{2}+\frac{\theta^{(t)}}{4}}\right)^{x_2}\mathbf{1}_{x_1+x_2=y_1}\prod_{i=3}^{5}\mathbf{1}_{x_i=y_i-1} \\ &= \frac{y_1}{x_1!x_2!}\left(\frac{2}{2+\theta^{(t)}}\right)^{x_1}\left(\frac{\theta^{(t)}}{2+\theta^{(t)}}\right)^{x_2}\mathbf{1}_{x_1+x_2=y_1}\prod_{i=3}^{5}\mathbf{1}_{x_i=y_i-1}. \end{aligned}$$

Then the conditional expectation of X given y and $\theta^{(m)}$ is

$$\mathrm{E}_{X|y,\theta^{(t)}}[X] = \begin{bmatrix} \frac{2}{2+\theta^{(t)}}y_1 & \frac{\theta^{(t)}}{2+\theta^{(t)}}y_2 & y_3 & y_4 \end{bmatrix}^T.$$

and the M-step becomes

$$\begin{aligned}
\theta^{(t+1)} &= \arg\max_{\theta}\left(\left(\frac{\theta^{(t)}}{2+\theta^{(t)}}y_1+y_4\right)\log\theta+(y_2+y_3)(1-\log\theta)\right)\\
&= \frac{\frac{\theta^{(t)}}{2+\theta^{(t)}}y_1+y_4}{\frac{\theta^{(t)}}{2+\theta^{(t)}}y_1+y_2+y_3+y_4}
\end{aligned}$$

Generally, one can impose any prior $p()$, and then modify EM to maximize the posterior rather than the likelihood:

$$\hat{\theta}_{MAP}=\arg\max_{\theta}\log p(\theta|y)=\arg\max_{\theta}(\log p(y|\theta)+p(\theta))$$

The EM algorithm is easily extended to maximum a posteriori (MAP) estimation by modifying the M-step:

$$\theta^{(t+1)}=\arg\max_{\theta}\ \left(Q(\theta|\theta^{(t)})+\log p(\theta)\right)$$

Example 13.3 (Exponential Family)**.** *Assume $p(y|\theta)=c_1(y)c_2(\theta)\exp(\theta^Ts(y))$. $s(y)$ is sufficient statistic. Using EM to estimate θ.*

Solution. Ignoring the items not involving θ, we have

$$Q(\theta|\theta^{(t)})=\log c_2(\theta)+\int\theta^Ts(y)p(z|x,\theta^{(t)})dz$$

Then,

$$\begin{aligned}
&Q'=0\\
\Rightarrow\quad&\frac{c_2(\theta)'}{c_2(\theta)}+\int s(y)p(z|x,\theta^{(t)})dz=0
\end{aligned}$$

Also, we have

$$\begin{aligned}
&\int p(y|\theta)=1\\
\Rightarrow\quad&\frac{c_2(\theta)'}{c_2(\theta)}+\int s(y)p(y|\theta)dy=0
\end{aligned}$$

Using above two equations, we can solve the problem.

Example 13.4. *Let $W=(W_1,W_2)^T$ be a bivariate random vector having a normal distribution $W\sim N(\mu,\Sigma)$ with mean $\mu=(\mu_1,\mu_2)^T$ and covariance matrix*

$$\Sigma=\begin{pmatrix}\sigma_{11}&\sigma_{12}\\\sigma_{12}&\sigma_{22}\end{pmatrix}.$$

The bivariate normal density is given by

$$\phi(w;\theta)=(2\pi)^{-1}|\Sigma|^{-\frac{1}{2}}\exp(-\frac{1}{2}(w-\mu)^T\Sigma^{-1}(w-\mu)),$$

where the vector of parameters θ is given by

$$\theta = (\mu_1, \mu_2, \sigma_{11}, \sigma_{12}, \sigma_{22})^T.$$

Suppose we wish to find the MLE of θ on the basis of a random sample of size n taken on W, where the data on the i-th variate W_i are missing in m_i of the units ($i = 1, 2$). We label the data so that $w_j = (w_{1j}, w_{2j})^T (j = 1, ..., m)$denote the fully observed data points, where $m = n - m_1 - m_2$, $w_{2j}(j = m+1, ..., m+m_1)$ denote the m_1 observations with the values of the first variate w_{1j} missing, and $w_{1j}(j = m+m_1+1, ..., n)$ denote the m_2 observations with the values of the second variate w_{2j} missing.

It is supposed that the "missingness" can be considered to be completely random, so that the observed data can be regarded as a random sample of size m from the bivariate normal distribution and an independent pair of independent random samples of size m_i from the univariate normal distributions

$$W_i \sim N(\mu_i, \sigma_{ii})$$

for i = 1,2.

Solution. The observed data therefore given by $y = (w_1^T, ..., w_m^T, v^T)^T$, where the vector v is given by

$$v = (w_{2,m+1}, \cdots, w_{2,m+m_1}, w_{2,m+m_1+1}, \cdots, w_{1,n})^T.$$

The log likelihood function for θ based on the observed data y is

$$\begin{aligned} \log L(\theta) &= -n\log(2\pi) - \frac{1}{2}m\log|\Sigma| - \frac{1}{2}\sum_{j=1}^{m}(w_j-\mu)^T\Sigma^{-1}(w_j-\mu) \\ &\quad -\frac{1}{2}\sum_{i=1}^{2} m_i \log\sigma_{ii} - \frac{1}{2}\left(\sigma_{11}^{-1}\sum_{j=m+m_1+1}^{m}(w_{1j}-\mu_1)^2 + \sigma_{22}^{-1}\sum_{j=m+1}^{m+m_1}(w_{2j}-\mu)^2\right). \end{aligned}$$

An obvious choice for the complete data here are the n bivariate observations. The complete-data vector x is then given by

$$x = (w_1^T, \cdots, w_n^T)^T,$$

for which the missing-data vector z is

$$z = (w_{1,m+1}, \cdots, w_{1,m+m_1}, w_{2,m+m_1+1}, \cdots, w_{2,n})^T.$$

The complete-data log likelihood function for θ is

$$\begin{aligned} \log L_c(\theta) &= -n\log(2\pi) - \frac{1}{2}n\log|\Sigma| - frac12\sum_{j=1}^{m}(w_j-\mu)^T\Sigma^{-1}(w_j-\mu) \\ &= -n\log(2\pi) - \frac{1}{2}n\log\xi - \frac{1}{2}\xi^{-1}[\sigma_{22}T_{11} + \sigma_{11}T_{22} - 2\sigma_{12}T_{12} \\ &\quad -2(T_1(\mu_1\sigma_{22} - \mu_2\sigma_{12}) + T_2(\mu_2\sigma_{11} - \mu_1\sigma_{12}) \\ &\quad +n(\mu_1^2\sigma_{22} + \mu_2^2\sigma_{11} - 2\mu_1\mu_2\sigma_{12})], \end{aligned}$$

where

$$T_i = \sum_{j=1}^{n} w_{ij}, \quad (i = 1, 2)$$

$$T_{hi} = \sum_{j=1}^{n} w_{hj} w_{ij}, \quad (h, i = 1, 2)$$

and where

$$\xi = \sigma_{11}\sigma_{22}(1 - \rho^2)$$

and

$$\rho = \sigma_{12}/(\sigma_{11}\sigma_{22})^{\frac{1}{2}}$$

is the correlation between W_1 and W_2.

It can be seen that $L_c(\theta)$ belongs to the regular exponential family with sufficient statistic

$$T = (T_1, T_2, T_{11}, T_{12}, T_{22})^T.$$

We now consider the E-step on the $(t + 1)$th iteration of the EM algorithm, where $\theta^{(t)}$ denotes the value of θ after the tth EM iteration. It can be seen that in order to compute the current conditional expectation of the complete-data log likelihood,

$$Q(\theta|\theta^{(t)}) = E[\log L_c(\theta)|y, \theta^{(t)}],$$

we require the current current conditional expectations of the sufficient statistics T_i and T_{hi}. Thus in effect we require

$$E[W_{1j}|w_{2j}, \theta^{(t)}]$$

and

$$E[W_{1j}^2|w_{2j}, \theta^{(t)}]$$

for $j = m + 1, \cdots, m + m_1$, and

$$E[W_{2j}|w_{1j}, \theta^{(t)}]$$

and

$$E[W_{2j}^2|w_{1j}, \theta^{(t)}]$$

for $j = m + m_1 + 1, \cdots, n$.

Form the well-known properties of the bivariate normal distribution, the conditional distribution of W_2 given $W_1 = w_1$ is normal with mean

$$\mu_2 + \sigma_{12}\sigma_{11}^{-1}(w_1 - \mu_1)$$

and variance

$$\sigma_{22.1} = \sigma_{22}(1 - \rho^2).$$

Thus,

$$E[W_{2j}|w_{1j}, \theta^{(t)}] = w_{2j}^{(t)}$$

where

$$w_{2j}^{(t)} = \mu_2^{(t)} + (\sigma_{12}^{(t)}/\sigma_{11}^{(t)})(w_{1j} - \mu_1^{(t)}),$$

and

$$E[W_{2j}^2|w_{1j},\theta^{(t)}] = w_{2j}^{(t)^2} + \sigma_{22.1}^{(t)}$$

for $j = m + m_1 + 1, \cdots, n$. Similarly, $E[W_{1j}|w_{2j},\theta^{(t)}]$ and $E[W_{1j}^2|w_{2j},\theta^{(t)}]$ are obtained by interchanging the subscripts 1 and 2.

Note that if we were to simply impute the $w_{2j}^{(t)^2}$ for the missing w_{2j} in the complete-data log likelihood (and likewise or the missing w_{1j}), we would not get the same expression as the Q-function yielded by the E-step, because of the omission of the term $\sigma_{22.1}^{(t)}$.

The M-step on the $(t + 1$th iteration is implemented simply by replacing T_i and T_{hi} by $T_i^{(t)}$ and $T_{hi}^{(t)}$, respectively, where the latter are defined by replacing the missing w_{ij} and w_{ij}^2 with their current conditional expectations. Accordingly, $\theta^{(t+1)}$ is given by

$$\mu_i^{(t+1)} = T_i^{(t)}/n$$

$$\sigma_{hi}^{(t+1)} = (T_{hi}^{(t)} - n^{-1}T_h^{(t)}T_i^{(t)})/n$$

13.2 Convergence

Expectation-maximization works to improve $Q(\theta|\theta^{(t)})$ rather than directly improving $\log p(X|\theta)$. Here we show that improvements to the former imply improvements to the latter.

For any Z with non-zero probability $p(Z|X,\theta)$, we can write

$$\log p(X|\theta) = \log p(X,Z|\theta) - \log p(Z|X,\theta)\,.$$

We take the expectation over values of Z by multiplying both sides by $p(Z|X,\theta^{(t)})$ and summing (or integrating) over Z. The left-hand side is the expectation of a constant, so we get:

$$\begin{aligned}\log p(X|\theta) &= \sum_Z p(Z|X,\theta^{(t)}) \log p(X,Z|\theta) - \sum_Z p(Z|X,\theta^{(t)}) \log p(Z|X,\theta) \\ &= Q(\theta|\theta^{(t)}) + H(\theta|\theta^{(t)})\,,\end{aligned}$$

where $H(\theta|\theta^{(t)})$ is defined by the negated sum it is replacing. This last equation holds for any value of θ including $\theta = \theta^{(t)}$,

$$\log p(X|\theta^{(t)}) = Q(\theta^{(t)}|\theta^{(t)}) + H(\theta^{(t)}|\theta^{(t)})\,,$$

and subtracting this last equation from the previous equation gives

$$\log p(X|\theta) - \log p(X|\theta^{(t)}) = Q(\theta|\theta^{(t)}) - Q(\theta^{(t)}|\theta^{(t)}) + H(\theta|\theta^{(t)}) - H(\theta^{(t)}|\theta^{(t)})\,,$$

However, Gibbs' inequality tells us that $H(\theta|\theta^{(t)}) \geq H(\theta^{(t)}|\theta^{(t)})$, so we can conclude that

$$\log p(X|\theta) - \log p(X|\theta^{(t)}) \geq Q(\theta|\theta^{(t)}) - Q(\theta^{(t)}|\theta^{(t)})\,.$$

In words, choosing θ to improve$Q(\theta|\theta^{(t)})$ beyond $Q(\theta^{(t)}|\theta^{(t)})$ will improve $\log p(X|\theta)$ beyond $\log p(X|\theta^{(t)})$ at least as much.

The origin EM algorithm is linear convergence. A number of methods have been proposed to accelerate the sometimes slow convergence of the EM algorithm, such as those using conjugate gradient and modified NewtonRaphson techniques. Additionally EM can be used with constrained estimation techniques.