NOTES ON THE EM ALGORITHM

IAN JERMYN

Imagine you have data y, and you want to explain y using a distribution from some family of probability distributions $p_\psi(y)$, where ψ is a point in some parameter space. (So you are moving around on a submanifold of the space of probability distributions over the space of y's.) You want to do this by maximizing the likelihood of y (questions as to whether this is the best thing can wait until later): in other words you want to find $\psi^* = \arg\max_\psi p_\psi(y) = \arg\max_\psi \log p_\psi(y)$.

Now look at the quantity $\Delta L_y(\psi', \psi) = \log(p_{\psi'}(y)) - \log(p_\psi(y))$, where ψ' and ψ are two different values of the parameter. If we can find a function $H_y(\psi', \psi)$, such that $\Delta L_y(\psi', \psi) \geq H_y(\psi', \psi)$ and $H_y(\psi, \psi) = 0$, then given ψ, finding a ψ' such that $H_y(\psi', \psi) > 0$, means that $\Delta L_y(\psi', \psi) > 0$ and hence that ψ' gives a higher likelihood to y than does ψ. We can therefore form the following algorithm:

(1) Choose initial ψ.
(2) $\psi_0 := \psi$
(3) While not converged loop
(4) $\psi_1 := \arg\max_{\psi'} H_y(\psi', \psi_0)$
(5) $\psi_0 := \psi_1$
(6) end loop
(7) return ψ_1

Because of the properties of H_y, each successive iteration produces a ψ_1 that is better than the previous one, in the sense that $\Delta L_y(\psi_1, \psi_0) > 0$. Under certain conditions (never met in practice), this is the global solution. More often it is a local solution, or worse things can happen like cycles and so on.

So now we have to find H_y. That goes like this. We suppose that there exists a distribution $p_\psi(y, z)$ over two variables, such that $\sum_z p_\psi(y, z) = p_\psi(y)$. The function ΔL_y becomes:

$$
\begin{aligned}
\Delta L_y(\psi', \psi) &= \log \frac{p_{\psi'}(y)}{p_\psi(y)} \\
&= \log \frac{\sum_z p_{\psi'}(y, z)}{p_\psi(y)} \\
&= \log \sum_z \frac{p_{\psi'}(y, z) p_\psi(z|y)}{p_\psi(y, z)} \\
&\geq \sum_z p_\psi(z|y) \log \frac{p_{\psi'}(y, z)}{p_\psi(y, z)} \\
&=: H_y(\psi', \psi)
\end{aligned} \tag{1}
$$

where the first line is the definition, the second follows from the definition of the 2D distribution, the third follows from Bayes' theorem, the fourth follows because log is a concave function, and the fifth is the definition of H_y. Note that $H_y(\psi, \psi) = 0$. In expositions of the EM algorithm, the fourth line is usually transformed in the following way:

$$H_y(\psi',\psi) = \sum_z p_\psi(z|y)\log\frac{p_{\psi'}(y,z)}{p_\psi(y,z)} \tag{2}$$
$$= \sum_z p_\psi(z|y)\log p_{\psi'}(y,z) - \sum_z p_\psi(z|y)\log p_\psi(y,z)$$
$$=: Q_y(\psi',\psi) - R_y(\psi)$$

Note that the second term does not depend on ψ', so that $\arg\max_{\psi'} H_y(\psi',\psi) = \arg\max_{\psi'} Q_y(\psi',\psi)$. Since

$$Q_y(\psi',\psi) = \sum_z p_\psi(z|y)\log p_{\psi'}(y,z) \tag{3}$$

$$Q_y(\psi',\psi) = E_\psi[\log p_{\psi'}(y,z)|y] \tag{4}$$

whre the second version is poor notation, but universally used. The EM algorithm typically uses Q_y rather than H_y.

This then leads to the EM algorithm:

(1) Expectation: calculate or compute

$$Q_y(\psi',\psi) = E_\psi[\log p_{\psi'}(y,z)|y]\sum_z p_\psi(z|y)\log p_{\psi'}(y,z) \tag{5}$$

(2) Maximization: calculate or compute

$$\psi^* = \arg\max_{\psi'} Q_y(\psi',\psi) \tag{6}$$