Unravelling the Noise

Based on lectures by Dr. Steve Presse

Notes taken by Daniel Moore

Fall 2019

Contents

1 Feb 18, 2020

1.1 Expectation Maximization

In order to estimate the parameters, θ, from our observations, y, we must use expectation maximization (EM). The function we're trying to maximize is

$$p(y_{1:N}|\theta) = \sum_{x_{1:N}} p(y_{1:N}, x_{1:N}|\theta)$$

Remember that we want to do this by maximizing the log of this, because we want to avoid numerical underflow. But maximizing the logarithm of a sum is very difficult. Much more difficult, in fact, than maximizing the sum of logs. Expectation Maximization is based around the idea that we don't have to actually maximize $\log(p(y_{1:N}|\theta))$, we can instead maximize the expectation value of $p(y_{1:N}, x_{1:N}|\theta)$, which we also call $Q(\theta, \theta^{old})$,

$$Q(\theta, \theta^{old}) = \sum_{x_{1:N}} \ln(p(y_{1:N}, x_{1:N}|\theta))\, p(y_{1:N}, x_{1:N}|\theta)$$

EM has two (iterative) steps:

(i) First, compute the expectation value,

$$\sum_{x_{1:N}} f_\theta(x_{1:N}) \ln(p(y_{1:N}, x_{1:N}|\theta))$$

The result will usually be a function of θ.

(ii) Next, we maximize the expectation value with respect to the unknown parameters θ.

If we're doing this iteratively, we can write the expectation instead as

$$Q_{\theta^{old}}(\theta) = \sum_{x_{1:N}} \mathrm{d}x_{1:N} f_{\theta^{old}} \ln(p(x_{1:N}, y_{1:N}|\theta))$$

As an example to illuminate how EM works, suppose we have a two-state system, where $x_n \in \{1, 2\}$, so

$$x_n \sim \text{Categorical}[\pi_1, \pi_2]$$

We can expect that

$$y_n \sim \text{Normal}[\mu_{x_n}, \nu_{x_n}]$$

Thus,

$$\theta = \{\mu_1, \mu_2, \nu_1, \nu_2, \pi_1, \pi_2\}$$

So we can write

$$\begin{aligned} p(x_n|\theta) &= \pi_1^{\delta_{1,x_n}} \pi_2^{\delta_{2,x_n}} \\ p(y_n|\theta) &= p(x_n = 1)p(y_n|x_n, \theta) + p(x_n = 2)p(y_n|x_n, \theta)p(x_n = 2)p(y_n|x_n, \theta) \\ &= \pi_1 \text{Normal}(\mu_1, \nu_1) + \pi_2 \text{Normal}(\mu_2, \nu_2) \end{aligned}$$

I got lost in the lecture, to see the rest of this derivation, look in the textbook, pg 70 (chapter 3).

1.2 Bayesian Logic: An Introduction