

EM for Linear Interpolation: Notes on Berger

These are notes to accompany Berger's "Complexity, Maximum Likelihood, and All That", based originally on some discussion with Noah Smith. (With thanks to Noah!)

If you go through Berger's derivation of the linear interpolation EM update equations, something is amiss: it looks as if the C_i are defined in terms of the λ' values, *and vice versa*.

Here's a more intuitive way to do what Berger did that equates. We're not going to use C_i until the end.

Take the un-numbered equation before the bit of text before [Berger's] equation 9. ... We have:

$$\sum_y \tilde{p}(y) p_\lambda(s = i|y) * (1/\lambda'_i) = \alpha \quad (1)$$

$$(1/\lambda'_i) * \sum_y \tilde{p}(y) p_\lambda(s = i|y) = \alpha \quad (2)$$

$$\sum_y \tilde{p}(y) p_\lambda(s = i|y) = \alpha * \lambda'_i \quad (3)$$

Now sum both sides over all i .

$$\sum_i \sum_y \tilde{p}(y) p_\lambda(s = i|y) = \sum_i \alpha * \lambda'_i \quad (4)$$

$$\sum_i \sum_y \tilde{p}(y) p_\lambda(s = i|y) = \alpha * \sum_i \lambda'_i \quad (5)$$

$$\sum_i \sum_y \tilde{p}(y) p_\lambda(s = i|y) = \alpha \quad (6)$$

(We can do the last step because we have the constraint that that $\sum_i \lambda'_i = 1$.) We now have an expression for α . We can plug this back into (3) and rearrange to get an expression for λ'_i :

$$\sum_y \tilde{p}(y) p_\lambda(s = i|y) = \left[\sum_i \sum_y \tilde{p}(y) p_\lambda(s = i|y) \right] * \lambda'_i \quad (7)$$

$$\left[\sum_y \tilde{p}(y) p_\lambda(s = i|y) \right] / \left[\sum_i \sum_y \tilde{p}(y) p_\lambda(s = i|y) \right] = \lambda'_i \quad (8)$$

If you define

$$C_i = \sum_y \tilde{p}(y) p_\lambda(s = i|y) \quad (9)$$

then this gives you the desired update equation, i.e. how to calculate the new values λ'_i , in the style Berger intended:

$$C_i / \sum_j C_j = \lambda'_i \quad (10)$$

Noah comments: *Berger's way of doing it is a little convoluted; he's trying to do the math and show you the insight that in our (8), the numerator is "like" an expected count of the number of times you picked model i, and the denominator is "like" a sum over all those expected counts. In fact, he's wrong; his C_i is only proportional to an expected count. To get a true expected count you'd have to use $c(y)$ instead of $\tilde{p}(y)$, but that doesn't quite jive with the usual EM formulation, which is in terms of an expected log probability or cross-entropy, instead of a total log probability. The difference amounts to whether you divide by T or not; I think EM is more clear when you don't. I have a suspicion that Berger was trying to be tricky and then forgot about it. The way he does it, I think $\sum_j C_j$ will always be equal to 1. Notice that the denominator in (8) is*

$$\sum_i \sum_y \tilde{p}(y) p_\lambda(s = i|y) = \sum_i \sum_y q(s = i, y) \quad (11)$$

for a joint distribution q over (i, y) which just happens to be defined by $q(i, y) = \tilde{p}(y) * p_\lambda(s = i|y)$. Summing over both y and i should give you 1. I suspect Berger was trying to make things work out nicely so he wouldn't have to renormalize, then forgot about the trick and went back to doing it the more intuitive way.