STA 410/2102, Fall 2015 Assignment #2 — Derivations and Discussion

Problem 1: In the E step, we must find the distribution of the unobserved gender of respondents to the first survey, of $n = 130$ people, given their answers to the Minecraft question. The genders of respondents will be independent in this conditional distribution, because they are sampled independently from the population. The probability that a respondent who answered "yes" is a man will be the same for all $x = 75$ such respondents — call this probability q_1. Similarly, let q_2 be the probability that one of the $n - x = 55$ respondents who answered "no" is a man.

Using Bayes' Rule, we find that

$$\begin{aligned} q_1 &= P(\text{man}|\text{yes}) = \frac{P(\text{man})P(\text{yes}|\text{man})}{P(\text{man})P(\text{yes}|\text{man}) + P(\text{woman})P(\text{yes}|\text{woman})} \\ &= (1/2)p_1 \,/\, ((1/2)p_1 + (1/2)p_2) = p_1 \,/\, (p_1 + p_2) \end{aligned}$$

and similarly,

$$\begin{aligned} q_2 &= P(\text{man}|\text{no}) = \frac{P(\text{man})P(\text{no}|\text{man})}{P(\text{man})P(\text{no}|\text{man}) + P(\text{woman})P(\text{no}|\text{woman})} \\ &= (1/2)(1-p_1) \,/\, ((1/2)(1-p_1) + (1/2)(1-p_2)) = (1-p_1) \,/\, (2 - p_1 - p_2) \end{aligned}$$

In the M Step, using q_1 and q_2, we maximize the expected value of the log of what the likelihood would be if we had observed the genders of all respondents to all surveys. Let z_1 be the (unobserved) number of "yes" respondents in the first survey who are men, and z_2 be the (unobserved) number of "no" respondents in this survey who are men. Then, if we had observed z_1 and z_2, the log likelihood function would be

$$\begin{aligned} &\log\left(p_1^{z_1+x_1}(1-p_1)^{z_2+m_1-x_1}p_2^{x-z_1+x_2}(1-p_2)^{n-x-z_2+m_2-x_2}\right) \\ &= (z_1 + x_1)\log(p_1) + (z_2 + m_1 - x_1)\log(1-p_1) \\ &\quad + (x - z_1 + x_2)\log(p_2) + (n - x - z_2 + m_2 - x_2)\log(1-p_2) \end{aligned}$$

Based on the distribution found in the E Step, the expected value of z_1 is xq_1 and the expected value of z_2 is $(n-x)q_2$. So the expected value of the log likelihood above is

$$\begin{aligned} &(xq_1 + x_1)\log(p_1) + ((n-x)q_2 + m_1 - x_1)\log(1-p_1) \\ &\quad + (x - xq_1 + x_2)\log(p_2) + (n - x - (n-x)q_2 + m_2 - x_2)\log(1-p_2) \end{aligned}$$

To maximize this with respect to p_1 and p_2, we set the partial derivatives with respect to p_1 and p_2 to zero and solve the resulting system of equations:

$$\begin{aligned} 0 &= (xq_1 + x_1) \,/\, p_1 - ((n-x)q_2 + m_1 - x_1) \,/\, (1-p_1) \\ 0 &= (x(1-q_1) + x_2) \,/\, p_2 - ((n-x)(1-q_2) + m_2 - x_2) \,/\, (1-p_2) \end{aligned}$$

The solution is

$$\begin{aligned} p_1 &= (xq_1 + x_1) \,/\, (xq_1 + x_1 + (n-x)q_2 + m_1 - x_1) \\ p_2 &= (x(1-q_1) + x_2) \,/\, (x(1-q_1) + x_2 + (n-x)(1-q_2) + m_2 - x_2) \end{aligned}$$

The resulting EM program is simple to write, once the above derivations have been done.

The final maximum likelihood estimates found with EM are essentially the same as were found in Assignment #1, differing by only 11×10^{-17} for p_1 and 5×10^{-17} for p_2, which is explainable by small amounts of round-off error. (Note that the exact round-off errors may differ depending on exactly how the program is written.)

EM converges quite slowly however, taking 92 iterations to reach a stable value starting from initial values of $p_1 = x_1/m_1$ and $p_2 = x_2/m_2$. This is slow compared to all the methods tried in Assignment #1, the slowest of which was alternating maximization, which took 24 iterations from the same initial values (although the computation time for each iteration would be significantly higher for alternating maximization than for EM).

As expected, convergence of EM appears to be linear, as shown from the following parameter estimates after every ten iterations:

```
p[1]: 0.83018132635594355  p[2]: 0.28317439082021822   iteration 10
     ^                             ^

p[1]: 0.8300790120498146   p[2]: 0.28328886914259965   iteration 20
               ^                           ^

p[1]: 0.83007558462502806  p[2]: 0.28329270297628018   iteration 30
                ^                           ^

p[1]: 0.8300754698229077   p[2]: 0.28329283139073169   iteration 40
                  ^                           ^

p[1]: 0.83007546597759685  p[2]: 0.28329283569198876   iteration 50
                   ^                           ^

p[1]: 0.83007546584879754  p[2]: 0.28329283583605996   iteration 60
                    ^                           ^

p[1]: 0.83007546584448344  p[2]: 0.28329283584088566   iteration 70
                     ^                           ^

p[1]: 0.830075465844339    p[2]: 0.28329283584104725   iteration 80
                      ^                           ^

p[1]: 0.83007546584433411  p[2]: 0.28329283584105275   iteration 90
                       ^                           ^


p[1]: 0.830075465844334    p[2]: 0.28329283584105286   final answer
```

Above, I have marked with ^ the position of the rightmost digit at which the error (comparing to the final answer) is less than half the digit value. The number of accurate digits grows linearly at the rate of about 1.5 digits for every 10 iterations.

Problem 2: In the E step, we must find the conditional distribution for the species of each beetle, given the observations of mass, length/width ratio, and swamp indicator for that beetle, and whether for that beetle the species is known, only the genus is known, or neither species or genus are known.

When neither species or genus is known, the conditional probabilities for beetle i to be of each species s can be obtained from Bayes' Rule as follows (noting that conditioning on the value of m_i is the same as conditioning on the value $\log(m_i)$, since log is an invertible function):

$$q_{is} = P(z_i = s | \log(m_i), \log(r_i), s_i) = \frac{P(z_i = s)P(\log(m_i)|z_i = s)P(\log(r_i|z_i = s)P(s_i|z_i = s)}{\sum_{j=1}^{10} P(z_i = j)P(m_i|z_i = j)P(r_i|z_i = j)P(s_i|z_i = j)}$$

$$\propto \alpha_s N(\log(m_i); \mu_s, 0.08^2) N(\log(r_i); \nu_s, 0.10^2) \rho_s^{s_i} (1 - \rho_s)^{1 - s_i}$$

When the genus is known, the same formula is used, except that the probabilities are zero for species not in the known genus, and the sum in the denominator above is over only species in that genus. When the species is known, the probability of that species is of course one, and the probabilities of other species are of course zero.

In the M step, we must maximize the expected value, with respect to the distribution over species of beetles found in the E step, of what the log likelihood would be if we had observed the species of all beetles.

This log likelihood would be

$$\sum_{i=1}^{500} \log\Big(\alpha_{z_i} N(\log(m_i);\mu_{z_i},0.08^2) N(\log(r_i);\nu_{z_i},0.10^2) \rho_{z_i}^{s_i}(1-\rho_{z_i})^{1-s_i}\Big)$$

$$= \sum_{i=1}^{500}\sum_{s=1}^{10} I(z_i=s)\Big(\log(\alpha_s) + \log N(\log(m_i);\mu_s,0.08^2) + \log N(\log(r_i);\nu_s,0.10^2) + s_i\log\rho_s + (1-s_i)\log(1-\rho_s)\Big)$$

where $I(z_i = s)$ is one if $z_i = s$ and zero otherwise. The expected value of this is

$$\sum_{i=1}^{500}\sum_{s=1}^{10} q_{is}\Big(\log(\alpha_s) + \log N(\log(m_i);\mu_s,0.08^2) + \log N(\log(r_i);\nu_s,0.10^2) + s_i\log\rho_s + (1-s_i)\log(1-\rho_s)\Big)$$

$$= \sum_{i=1}^{500}\sum_{s=1}^{10} q_{is}\Big(\log(\alpha_s) - 0.5(\log(m_i)-\mu_s)^2/0.08^2 - 0.5(\log(r_i)-\nu_s)^2/0.10^2 + s_i\log\rho_s + (1-s_i)\log(1-\rho_s)\Big)$$

$$+ \text{ constant}$$

Equating the derivative of this with respect to μ_s to zero gives

$$0 = \sum_{i=1}^{500} q_{is}(\log(m_i)-\mu_s)/0.08^2$$

from which one gets that the new value for μ_s should be

$$\mu_s = \sum_{i=1}^{500} q_{is}\log(m_i) \Big/ \sum_{i=1}^{500} q_{is}$$

Similarly, the the new value for ν_s should be

$$\nu_s = \sum_{i=1}^{500} q_{is}\log(r_i) \Big/ \sum_{i=1}^{500} q_{is}$$

Equating the derivative with respect to ρ_s to zero gives

$$0 = \sum_{i=1}^{500} q_{is}\Big(s_i/\rho_s - (1-s_i)/(1-\rho_s)\Big)$$

which gives the new value for ρ_s as

$$\rho_s = \sum_{i=1}^{500} q_{is}s_i \Big/ \sum_{i=1}^{500} q_{is}$$

At the maximum, the partial derivatives of all the α_s must have the same value, say λ, since this implies that no change in the α_s that keeps their sum one will increase the expected log likelihood. So we must have, for all s,

$$\sum_{i=1}^{500} q_{is}/\alpha_s = \lambda$$

and hence the new values for the α_s must have the form $\alpha_s = \sum_{i=1}^{500} q_{is}/\lambda$. From the requirement that $\sum_{s=1}^{10}\alpha_s = 1$, we find that $\lambda = \sum_{s=1}^{10}\sum_{i=1}^{500} q_{is} = 500$.

On the data provided, EM takes 195 iterations to reach a completely stable state, starting with estimates in which all species have the means for log mass and log ratio, and the swamp probability, that

are found from sample means over all the data, and in which the species are equally abundant. In the last few iterations, the log likelihood sometimes decreases, but only very slightly, no more than is explainable by slight round-off errors.

This convergence time is not very fast, but other methods for a problem like this with 39 parameters might have worse problems, such as lack of stability, or difficulty in computing the Hessian matrix. The derivation and implementation of EM were typical of similar problems that are often solved by EM, and hence not too hard to produce (once one is familiar with such applications of EM).

To visualize the results, the data was plotted in one scatterplot of all beetles with mass and length/width ratio as the x and y axes, and with the swamp indicator given by colour. The plot symbol for a beetle was a digit (1234567890) if the species is known, a letter (abcd) if the genus but not species is known, and an asterisk if neither is known. The parameter estimates from EM were added to this plot as large digits, positioned according to the μ and ν parameters, with colour used to represent the swamp probability (as less than 1/3, greater than 2/3, or inbetween). The plot does not show the estimated species abundances, but these are easily understood from the tabular output of the α parameters.

For comparison, the estimates found from the obvious sample means using only data in which the species was known were also computed. They are not drastically different from the EM estimates (which is reassuring regarding the correctness of the EM program), but there are some differences large enough to be of practical importance.