(3.) Given (4.5): $\ell(\theta) = \prod_{i: y_i = 1} p(x_i) \prod_{i': y_{i'} = 0} (1 - p(x_{i'}))$

where:

(i) θ = vector of parameters on which p depends, which is (β_0, β_1) in above equation in the text

(ii) $p(x) = (e^{\beta_0 + \beta_1 x}) / (1 + e^{\beta_0 + \beta_1 x})$ for the binomial likelihood in the text, although we could generalize it to have $x_1, \dots x_K$ instead of just x easily

(iii) Per above setup, we have $M = 2$ classes, thus y_i takes value in $\{0, 1\}$ in this setup (i.e. this is dummy encoding). Let m denote class and thus $m \in \{0, 1\}$

(iv) We have n samples

Taking the negative of the log of (4.5) yields:

$$-\log(\ell(\theta)) = -\log\left(\prod_{i: y_i = 1} p(x_i) \cdot \prod_{i': y_{i'} = 0} (1 - p(x_{i'}))\right)$$

$$= -\left[\log\left(\prod_{i: y_i = 1} p(x_i)\right) + \log\left(\prod_{i': y_{i'} = 0} (1 - p(x_{i'}))\right)\right]$$ ← property of logs

$$= -\left[\sum_{i: y_i = 1} \log(p(x_i)) + \sum_{i': y_{i'} = 0} \log(1 - p(x_{i'}))\right]$$ ← property of logs

of the target

We note the above encoding is "dummy" (binary) encoding, and (10.14) and thus the below will assume "one-hot" encoding. Thus we now index each actual target as y_{im} = value of target for i^{th} sample in the column for class m. Thus, given the way we've split the sums, each $y_{im} = 1$, and thus we can write:

↓ BACK ↓

$$-\log\big(\ell(\theta)\big) = -\Big[\sum_{i:\, y_i=1} \log\big(p(x_i)\big)^{y_{i1}} + \sum_{i':\, y_{i'}=0} \log\big(1-p(x_{i'})\big)^{y_{i0}}\Big]$$

Further, to align our notation with our new one-hot encoding, we write:

$$f_1(x_i) = p(x_i)$$
$$f_0(x_i) = 1 - p(x_i)$$

Thus,

$$-\log\big(\ell(\theta)\big) = -\Big[\sum_{i:\, y_i=1} \log\big(f_1(x_i)\big)^{y_{i1}} + \sum_{i':\, y_{i'}=0} \log\big(f_0(x_{i'})\big)^{y_{i0}}\Big]$$

↓ property of logs

$$= -\Big[\sum_{i:\, y_i=1} y_{i1} \cdot \log\big(f_1(x_i)\big) + \sum_{i':\, y_{i'}=0} y_{i0} \cdot \log\big(f_0(x_{i'})\big)\Big]$$

Lastly, we can now group these two sums into one sum by realizing that:

(i) There are n total samples

(ii) Given our one-hot encoding for the target, y_{i1} will equal 1 when $y_i = 1$ and 0 when $y_i = 0$, and y_{i0} will equal 1 when $y_{i'} = 0$ and 0 otherwise, (for the i^{th} sample)

Thus, (10.14) when $M=2$

$$-\log\big(\ell(\theta)\big) = -\sum_{i=1}^{n} \sum_{m=0}^{1} y_{im} \cdot \log\big(f_m(x_i)\big)$$

To clarify, since exactly half of the y_{im}'s are 0 (since each sample will have one 1 and one 0, half of these terms will be zero.

which demonstrates that the negative log of the likelihood expression (4.5) is equivalent to the negative multinomial log-likelihood (10.14) when there are $M=2$ classes.