Gradient derivation for FFNN for binary data

Model - No hidden layers & L neurons

$$z_{\ell i} = \phi\left(b_\ell + \sum_{j=1}^{p} w_{\ell j} X_{ij}\right) \quad \text{for } i \in \{1, \dots, n\}, \; \ell \in \{1, \dots, L\}$$

$$n_i = \beta_0 + \sum_{\ell=1}^{L} z_{\ell i} \beta_\ell$$

$$p_i = \text{Prob}(Y_i = 1 \mid X_i) = \frac{e^{n_i}}{1 + e^{n_i}} \qquad \theta = \{b_\ell, w_{\ell j}, \beta_\ell\}$$

Likelihood

$$\log f(Y \mid \theta) = \sum_{i=1}^{n} Y_i \log(p_i) + (1 - Y_i) \log(1 - p_i)$$

$$= \sum_{i=1}^{n} Y_i n_i - Y_i \log(1 + e^{n_i}) - (1 - Y_i) \log(1 + e^{n_i})$$

(1)
$$= \sum_{i=1}^{n} Y_i n_i - \log(1 + e^{n_i})$$

(2)
$$= \sum_{i=1}^{n} Y_i \left(\beta_0 + \sum_\ell z_{\ell i} \beta_\ell\right) + \sum_{i=1}^{n} \log\left(1 + e^{\beta_0 + \sum_\ell z_{\ell i} \beta_\ell}\right)$$

(3)
$$= \sum_{i=1}^{n} Y_i \left(\beta_0 + \sum_\ell \beta_\ell \phi\left(b_\ell + \sum_j w_{\ell j} X_{ij}\right)\right)$$
$$- \sum_{i=1}^{n} \log\left(1 + e^{\beta_0 + \sum_\ell \phi\left(b_\ell + \sum_j w_{\ell j} X_{ij}\right) \beta_\ell}\right)$$

Output layer gradients (using (2))

$$\frac{\partial \log(f(Y|\theta))}{\partial \beta_0} = \sum_{i=1}^{n} Y_i - \sum_{i=1}^{n} \frac{e^{\eta_i}}{1+e^{\eta_i}} = \sum_{i=1}^{n} (Y_i - p_i)$$

$$\frac{\partial \log(f(Y|\theta))}{\partial \beta_\ell} = \sum_{i=1}^{n} Y_i Z_{\ell i} - \sum_{i=1}^{n} \frac{e^{\eta_i}}{1+e^{\eta_i}} Z_{\ell i}$$

$$= \sum_{i=1}^{n} Z_{\ell i} (Y_i - p_i)$$

Input layer gradients

$$\frac{\partial \log(f(Y|\theta))}{\partial b_\ell} = \sum_{i=1}^{n} Y_i \frac{\partial \eta_i}{\partial b_\ell} - \sum_{i=1}^{n} \frac{e^{\eta_i}}{1+e^{\eta_i}} \frac{\partial \eta_i}{\partial b_\ell}$$

$$= \sum_{i=1}^{n} (Y_i - p_i) \frac{\partial \eta_i}{\partial b_\ell}$$

$$\frac{\partial \eta_i}{\partial b_\ell} = \frac{\partial}{\partial b_\ell} \beta_0 + \sum_{\ell=1}^{L} Z_{\ell i} \beta_\ell$$

$$= \frac{\partial}{\partial b_\ell} Z_{\ell i} \beta_\ell$$

$$= \beta_\ell \frac{\partial}{\partial b_\ell} \phi\left(b_\ell + \sum_{j=1}^{p} w_{\ell j} X_{ij}\right)$$

$$= \beta_\ell \, \phi'\left(b_\ell + \sum_{j=1}^{p} w_{\ell j} X_{ij}\right)$$

$$\cancel{= \beta_\ell \, \phi'\left(b_\ell + \sum_{j=1}^{p} w_{\ell j} X_{ij}\right)}$$

$$\frac{\partial \log(f(Y|\theta))}{\partial w_{\ell j}} = \sum_{i=1}^{n} (y_i - p_i) \frac{\partial \eta_i}{\partial w_{\ell j}}$$

as in previous result

$$\frac{\partial \eta_i}{\partial w_{\ell j}} = \frac{\partial}{\partial w_{\ell j}} \beta_0 + \sum_{\ell=1}^{L} z_{\ell i} \beta_\ell$$

$$= \beta_\ell \frac{\partial z_{\ell i}}{\partial w_{\ell j}}$$

$$= \beta_\ell \frac{\partial}{\partial w_{\ell j}} \phi\left(b_\ell + \sum_{j=1}^{p} w_{\ell j} x_{ij}\right)$$

$$= \beta_\ell \phi'\left(b_\ell + \sum_{j=1}^{p} w_{\ell j} x_{ij}\right) x_{ij}$$