1 Binary Classification with Sum of Squared Error

For given x and W,

$$x = \begin{bmatrix} x_1 & x_2 & \cdots & x_d \end{bmatrix} \tag{1}$$

$$W = \begin{bmatrix} W_1 \\ W_2 \\ \vdots \\ W_d \end{bmatrix} \tag{2}$$

1. Linear Model

$$z = xW \tag{3}$$

2. Derivative of linear model: For $j = 1, 2, \cdots, d$,

$$\frac{\partial}{\partial W_j} z = \frac{\partial}{\partial W_j}(xW) \tag{4}$$

$$= \frac{\partial}{\partial W_j}(x_1 W_1 + \cdots + x_d W_d) \tag{5}$$

$$= x_j \tag{6}$$

3. Sigmoid

$$\hat{y} = \sigma(z) = \frac{1}{1 + e^{-z}} \tag{7}$$

4. Derivative of sigmoid

$$\sigma'(z) = \left[\frac{1}{1 + e^{-z}} \right]' \tag{8}$$

$$= \left(-\frac{1}{(1 + e^{-z})^2} \right) \cdot \left(-e^{-z} \right) \tag{9}$$

$$= \frac{e^{-z}}{(1 + e^{-z})^2} \tag{10}$$

$$= \frac{1}{1 + e^{-z}} \frac{e^{-z}}{1 + e^{-z}} \tag{11}$$

$$= \frac{1}{1 + e^{-z}} \frac{1 + e^{-z} - 1}{1 + e^{-z}} \tag{12}$$

$$= \frac{1}{1 + e^{-z}} \left(1 - \frac{1}{1 + e^{-z}} \right) \tag{13}$$

$$= \sigma(z)(1 - \sigma(z)) \tag{14}$$

5. FeedForward Model

$$\hat{y} = \sigma(xW) \tag{15}$$

6. Derivative of FeedForward Model

$$\frac{\partial}{\partial W_j} \hat{y} = \frac{\partial}{\partial W_j} \sigma(xW) \tag{16}$$

$$= \frac{\partial}{\partial z} \sigma(z) \frac{\partial z}{\partial W_j} \tag{17}$$

$$= \sigma'(z) \frac{\partial z}{\partial W_j} \tag{18}$$

$$= \sigma'(z) \frac{\partial (xW)}{\partial W_j} \tag{19}$$

$$= \sigma(z)(1 - \sigma(z)) \frac{\partial (xW)}{\partial W_j} \tag{20}$$

$$= \sigma(z)(1 - \sigma(z)) x_j \tag{21}$$

$$= \hat{y}(1 - \hat{y}) x_j \tag{22}$$

7. Loss function, $E(y, \hat{y})$ where $y \in \{0, 1\}$.

$$E(y, \hat{y}) = \frac{1}{2}|y - \hat{y}|^2 \tag{23}$$

Backpropagation

$$E(y, \hat{y}) = \frac{1}{2}|y - \hat{y}|^2 \tag{24}$$

For $j = 1, 2, \cdots, d$,

$$\frac{\partial}{\partial W_j} E(y, \hat{y}) = \frac{\partial}{\partial W_j} \left[\frac{1}{2}|y - \hat{y}|^2 \right] \tag{25}$$

$$= \frac{1}{2} 2(y - \hat{y}) \frac{\partial \hat{y}}{\partial W_j} \tag{26}$$

$$= (y - \hat{y}) \frac{\partial}{\partial W_j} \sigma(xW) \tag{27}$$

$$= (y - \hat{y}) \sigma(z)(1 - \sigma(z)) \frac{\partial}{\partial W_j}(xW) \tag{28}$$

$$= (y - \hat{y}) \hat{y}(1 - \hat{y}) \frac{\partial}{\partial W_j}(xW) \tag{29}$$

$$= (y - \hat{y}) \hat{y}(1 - \hat{y}) x_j \tag{30}$$

2 Binary Classification with Cross-Entropy

For given x and W,

$$x = \begin{bmatrix} x_1 & x_2 & \cdots & x_d \end{bmatrix} \tag{31}$$

$$W = \begin{bmatrix} W_1 \\ W_2 \\ \vdots \\ W_d \end{bmatrix} \tag{32}$$

1. Linear Model

$$z = xW \tag{33}$$

2. Derivative of linear model: For $j = 1, 2, \cdots, d$,

$$\frac{\partial}{\partial W_j} z = \frac{\partial}{\partial W_j}(xW) \tag{34}$$

$$= \frac{\partial}{\partial W_j}(x_1 W_1 + \cdots + x_d W_d) \tag{35}$$

$$= x_j \tag{36}$$

3. Sigmoid

$$\hat{y} = \sigma(z) = \frac{1}{1 + e^{-z}} \tag{37}$$

4. Derivative of sigmoid

$$\sigma'(z) = \left[\frac{1}{1 + e^{-z}} \right]' \tag{38}$$

$$= \left(-\frac{1}{(1 + e^{-z})^2} \right) \cdot \left(-e^{-z} \right) \tag{39}$$

$$= \frac{e^{-z}}{(1 + e^{-z})^2} \tag{40}$$

$$= \frac{1}{1 + e^{-z}} \frac{e^{-z}}{1 + e^{-z}} \tag{41}$$

$$= \frac{1}{1 + e^{-z}} \frac{1 + e^{-z} - 1}{1 + e^{-z}} \tag{42}$$

$$= \frac{1}{1 + e^{-z}} \left(1 - \frac{1}{1 + e^{-z}} \right) \tag{43}$$

$$= \sigma(z)(1 - \sigma(z)) \tag{44}$$

5. FeedForward Model

$$\hat{y} = \sigma(xW) \tag{45}$$

6. Derivative of FeedForward Model

$$\frac{\partial}{\partial W_j} \hat{y} = \frac{\partial}{\partial W_j} \sigma(xW) \tag{46}$$

$$= \frac{\partial}{\partial z} \sigma(z) \frac{\partial z}{\partial W_j} \tag{47}$$

$$= \sigma'(z) \frac{\partial z}{\partial W_j} \tag{48}$$

$$= \sigma'(z) \frac{\partial (xW)}{\partial W_j} \tag{49}$$

$$= \sigma(z)(1 - \sigma(z)) \frac{\partial (xW)}{\partial W_j} \tag{50}$$

$$= \sigma(z)(1 - \sigma(z)) x_j \tag{51}$$

$$= \hat{y}(1 - \hat{y}) x_j \tag{52}$$

7. Loss function, $E(y, \hat{y})$ where $y \in \{0, 1\}$.

$$E(y, \hat{y}) = -y \log(\hat{y}) - (1 - y) \log((1 - \hat{y})) \tag{53}$$

Backpropagation

$$E(y, \hat{y}) = -y \log(\hat{y}) - (1 - y) \log((1 - \hat{y})) \tag{54}$$

For $j = 1, 2, \cdots, d$,

$$\frac{\partial}{\partial W_j} E(y, \hat{y}) = \frac{\partial}{\partial W_j} \left[-y \log(\hat{y}) - (1 - y) \log((1 - \hat{y})) \right] \tag{55}$$

$$= \frac{\partial}{\partial \hat{y}} \left[-y \log(\hat{y}) - (1 - y) \log((1 - \hat{y})) \right] \frac{\partial \hat{y}}{\partial W_j} \tag{56}$$

$$= \left[-\frac{y}{\hat{y}} + \frac{1 - y}{1 - \hat{y}} \right] \frac{\partial \hat{y}}{\partial W_j} \tag{57}$$

$$= \left[-\frac{y}{\hat{y}} + \frac{1 - y}{1 - \hat{y}} \right] \frac{\partial}{\partial W_j} \sigma(xW) \tag{58}$$

$$= \left[-\frac{y}{\hat{y}} + \frac{1 - y}{1 - \hat{y}} \right] \frac{\partial}{\partial z} \sigma(z) \frac{\partial z}{\partial W_j} \tag{59}$$

$$= \left[-\frac{y}{\hat{y}} + \frac{1 - y}{1 - \hat{y}} \right] \sigma(z)(1 - \sigma(z)) \frac{\partial z}{\partial W_j} \tag{60}$$

$$= \left[-\frac{y}{\hat{y}} + \frac{1 - y}{1 - \hat{y}} \right] \sigma(xW)(1 - \sigma(xW)) \frac{\partial z}{\partial W_j} \tag{61}$$

$$= \left[-\frac{y}{\hat{y}} + \frac{1 - y}{1 - \hat{y}} \right] \hat{y}(1 - \hat{y}) \frac{\partial z}{\partial W_j} \tag{62}$$

$$= \left[-\frac{y}{\hat{y}} + \frac{1 - y}{1 - \hat{y}} \right] \hat{y}(1 - \hat{y}) \frac{\partial(xW)}{\partial W_j} \tag{63}$$

$$= \left[-\frac{y}{\hat{y}} + \frac{1 - y}{1 - \hat{y}} \right] \hat{y}(1 - \hat{y}) \frac{\partial(xW)}{\partial W_j} \tag{64}$$

$$= \left[-\frac{y}{\hat{y}} + \frac{1 - y}{1 - \hat{y}} \right] \hat{y}(1 - \hat{y}) x_j \tag{65}$$

$$= \left[-y(1 - \hat{y}) + (1 - y)\hat{y} \right] x_j \tag{66}$$

$$= \left[-y + y\hat{y} + \hat{y} - y\hat{y} \right] x_j \tag{67}$$

$$= (-y + \hat{y}) x_j \tag{68}$$

$$= -(y - \hat{y}) x_j \tag{69}$$