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

$$\begin{aligned} \frac{\partial}{\partial W_j} z &= \frac{\partial}{\partial W_j}(xW) &(4)\\ &= \frac{\partial}{\partial W_j}(x_1 W_1 + \cdots + x_d W_d) &(5)\\ &= x_j &(6) \end{aligned}$$

3. Sigmoid

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

4. Derivative of sigmoid

$$\begin{aligned} \sigma'(z) &= \left[\frac{1}{1+e^{-z}}\right]' &(8)\\ &= \left(-\frac{1}{(1+e^{-z})^2}\right) \cdot \left(-e^{-z}\right) &(9)\\ &= \frac{e^{-z}}{(1+e^{-z})^2} &(10)\\ &= \frac{1}{1+e^{-z}} \frac{e^{-z}}{1+e^{-z}} &(11)\\ &= \frac{1}{1+e^{-z}} \frac{1+e^{-z}-1}{1+e^{-z}} &(12)\\ &= \frac{1}{1+e^{-z}} \left(1 - \frac{1}{1+e^{-z}}\right) &(13)\\ &= \sigma(z)(1-\sigma(z)) &(14) \end{aligned}$$

5. FeedForward Model

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

6. Derivative of FeedForward Model

$$\begin{aligned} \frac{\partial}{\partial W_j}\hat{y} &= \frac{\partial}{\partial W_j}\sigma(xW) &(16)\\ &= \frac{\partial}{\partial z}\sigma(z)\frac{\partial z}{\partial W_j} &(17)\\ &= \sigma'(z)\frac{\partial z}{\partial W_j} &(18)\\ &= \sigma'(z)\frac{\partial (xW)}{\partial W_j} &(19)\\ &= \sigma(z)(1-\sigma(z))\frac{\partial (xW)}{\partial W_j} &(20)\\ &= \sigma(z)(1-\sigma(z))x_j &(21)\\ &= \hat{y}(1-\hat{y})x_j &(22) \end{aligned}$$

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

$$\begin{aligned}
\frac{\partial}{\partial W_j} E(y, \hat{y}) &= \frac{\partial}{\partial W_j} \left[\frac{1}{2}|y - \hat{y}|^2 \right] && (25)\\
&= \frac{1}{2} 2(y - \hat{y}) \frac{\partial \hat{y}}{\partial W_j} && (26)\\
&= (y - \hat{y}) \frac{\partial}{\partial W_j} \sigma(xW) && (27)\\
&= (y - \hat{y}) \sigma(z)(1 - \sigma(z)) \frac{\partial}{\partial W_j}(xW) && (28)\\
&= (y - \hat{y}) \hat{y}(1 - \hat{y}) \frac{\partial}{\partial W_j}(xW) && (29)\\
&= (y - \hat{y}) \hat{y}(1 - \hat{y}) x_j && (30)
\end{aligned}$$

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

$$\begin{aligned} \frac{\partial}{\partial W_j} z &= \frac{\partial}{\partial W_j}(xW) && (34) \\ &= \frac{\partial}{\partial W_j}(x_1 W_1 + \cdots + x_d W_d) && (35) \\ &= x_j && (36) \end{aligned}$$

3. Sigmoid

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

4. Derivative of sigmoid

$$\begin{aligned} \sigma'(z) &= \left[\frac{1}{1 + e^{-z}}\right]' && (38) \\ &= \left(-\frac{1}{(1 + e^{-z})^2}\right) \cdot \left(-e^{-z}\right) && (39) \\ &= \frac{e^{-z}}{(1 + e^{-z})^2} && (40) \\ &= \frac{1}{1 + e^{-z}} \frac{e^{-z}}{1 + e^{-z}} && (41) \\ &= \frac{1}{1 + e^{-z}} \frac{1 + e^{-z} - 1}{1 + e^{-z}} && (42) \\ &= \frac{1}{1 + e^{-z}} \left(1 - \frac{1}{1 + e^{-z}}\right) && (43) \\ &= \sigma(z)(1 - \sigma(z)) && (44) \end{aligned}$$

5. FeedForward Model

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

6. Derivative of FeedForward Model

$$\begin{aligned} \frac{\partial}{\partial W_j} \hat{y} &= \frac{\partial}{\partial W_j} \sigma(xW) && (46) \\ &= \frac{\partial}{\partial z} \sigma(z) \frac{\partial z}{\partial W_j} && (47) \\ &= \sigma'(z) \frac{\partial z}{\partial W_j} && (48) \\ &= \sigma'(z) \frac{\partial (xW)}{\partial W_j} && (49) \\ &= \sigma(z)(1 - \sigma(z)) \frac{\partial (xW)}{\partial W_j} && (50) \\ &= \sigma(z)(1 - \sigma(z)) x_j && (51) \\ &= \hat{y}(1 - \hat{y}) x_j && (52) \end{aligned}$$

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

$$\begin{aligned}
\frac{\partial}{\partial W_j} E(y, \hat{y}) &= \frac{\partial}{\partial W_j} \left[-y \log(\hat{y}) - (1 - y) \log((1 - \hat{y})) \right] && (55) \\
&= \frac{\partial}{\partial \hat{y}} \left[-y \log(\hat{y}) - (1 - y) \log((1 - \hat{y})) \right] \frac{\partial \hat{y}}{\partial W_j} && (56) \\
&= \left[-\frac{y}{\hat{y}} + \frac{1 - y}{1 - \hat{y}} \right] \frac{\partial \hat{y}}{\partial W_j} && (57) \\
&= \left[-\frac{y}{\hat{y}} + \frac{1 - y}{1 - \hat{y}} \right] \frac{\partial}{\partial W_j} \sigma(xW) && (58) \\
&= \left[-\frac{y}{\hat{y}} + \frac{1 - y}{1 - \hat{y}} \right] \frac{\partial}{\partial z} \sigma(z) \frac{\partial z}{\partial W_j} && (59) \\
&= \left[-\frac{y}{\hat{y}} + \frac{1 - y}{1 - \hat{y}} \right] \sigma(z)(1 - \sigma(z)) \frac{\partial z}{\partial W_j} && (60) \\
&= \left[-\frac{y}{\hat{y}} + \frac{1 - y}{1 - \hat{y}} \right] \sigma(xW)(1 - \sigma(xW)) \frac{\partial z}{\partial W_j} && (61) \\
&= \left[-\frac{y}{\hat{y}} + \frac{1 - y}{1 - \hat{y}} \right] \hat{y}(1 - \hat{y}) \frac{\partial z}{\partial W_j} && (62) \\
&= \left[-\frac{y}{\hat{y}} + \frac{1 - y}{1 - \hat{y}} \right] \hat{y}(1 - \hat{y}) \frac{\partial (xW)}{\partial W_j} && (63) \\
&= \left[-\frac{y}{\hat{y}} + \frac{1 - y}{1 - \hat{y}} \right] \hat{y}(1 - \hat{y}) \frac{\partial (xW)}{\partial W_j} && (64) \\
&= \left[-\frac{y}{\hat{y}} + \frac{1 - y}{1 - \hat{y}} \right] \hat{y}(1 - \hat{y}) x_j && (65) \\
&= \left[-y(1 - \hat{y}) + (1 - y)\hat{y} \right] x_j && (66) \\
&= \left[-y + y\hat{y} + \hat{y} - y\hat{y} \right] x_j && (67) \\
&= \left(-y + \hat{y} \right) x_j && (68) \\
&= -\left(y - \hat{y} \right) x_j && (69)
\end{aligned}$$

3 Neural Network Regression

For the simplicity, we assume that we have no bias term. Let $x \in \mathbf{R}^n$, $h \in \mathbf{R}^m$, $W^1 \in \mathbf{R}^{n \times m}$, and $W^2 \in \mathbf{R}^{m \times 1}$.

1. Input layer $x \in \mathbf{R}^n$.
2. Hidden layer $h \in \mathbf{R}^m$ with sigmoid function $\sigma(z) = \frac{1}{1+e^{-z}}$.

$$h = \sigma(xW^1) \tag{70}$$

$$h_k = \sigma(x_1 W^1_{1k} + x_2 W^1_{2k} + \cdots + x_n W^1_{nk}) \qquad \forall k = 1, 2, \cdots, m \tag{71}$$

3. Output layer : $\hat{y} \in \mathbf{R}^1$

$$\hat{y} = hW^2 \tag{72}$$

$$= h_1 W^2_1 + h_2 W^2_2 + \cdots + h_m W^2_m \tag{73}$$

4. Loss function E

$$E = \frac{1}{2}(y - \hat{y})^2 \tag{74}$$

Backpropagation

1. Gradient with respect to $W^2 \in \mathbf{R}^{m \times 1}$. For $i = 1, 2, \cdots, m$,

$$\frac{\partial E}{\partial W^2_i} = \frac{\partial}{\partial W^2_i}\left[\frac{1}{2}(y - \hat{y})^2\right] \tag{75}$$

$$= -(y - \hat{y})\frac{\partial \hat{y}}{\partial W^2_i} \tag{76}$$

$$= -(y - \hat{y})h_i \tag{77}$$

2. Gradient with respect to $W^1 \in \mathbf{R}^{n \times m}$. For $i = 1, 2, \cdots, n$, and $j = 1, 2, \cdots, m$,

$$\frac{\partial E}{\partial W^1_{ij}} = \frac{\partial}{\partial W^1_{ij}}\left[\frac{1}{2}(y - \hat{y})^2\right] \tag{78}$$

$$= -(y - \hat{y})\frac{\partial \hat{y}}{\partial W^1_{ij}} \tag{79}$$

$$= -(y - \hat{y})\sum_{k=1}^{m}\frac{\partial \hat{y}}{\partial h_k}\frac{\partial h_k}{\partial W^1_{ij}} \tag{80}$$

$$= -(y - \hat{y})\sum_{k=1}^{m} W^2_k \frac{\partial h_k}{\partial W^1_{ij}} \tag{81}$$

$$= -(y - \hat{y})\sum_{k=1}^{m} W^2_k h_k(1 - h_k)\frac{\partial (xW^1)_k}{\partial W^1_{ij}} \tag{82}$$

$$= -(y - \hat{y})W^2_j h_j(1 - h_j)x_i \tag{83}$$