Logistic Regression & Multi-class classification

Binary classification: $z = w^T x + b$. $y = \begin{cases} 1 & z \geq r \text{ (threshold)} \\ 0 & z < r \end{cases}$

$w^T x + b \geq r$ = decision boundary. $\Rightarrow w^T x + \underbrace{b - r}_{\text{new intercept} / w_0} \geq 0$

e.g.

$x \leftarrow [1, x]$

$x : \in \mathbb{R}^D$; after assignment $x \in \mathbb{R}^{D+1}$; $z = w^T x$.

x_0	x_1	t
1	0	1
1	1	0

when $x_1 = 0$: $z = w_0 x_0 + w_1 x_1 \geq 0 \Rightarrow w_0 \geq 0$

$x_1 = 1$: $z = w_0 x_0 + w_1 x_1 < 0 \Rightarrow w_0 + w_1 < 0$

x_0	x_1	x_2	t
1	0	0	0
1	0	1	0
1	1	0	0
1	1	1	1

$z = w_0 x_0 + w_1 x_1 + w_2 x_2$

$w_0 < 0$

$w_0 + w_2 < 0$

$w_0 + w_1 < 0$

$w_0 + w_1 + w_2 \geq 0$

$\mathcal{L}_{0,1}(y, t) = \begin{cases} 0 & y = t \\ 1 & y \neq t \end{cases}$

$J = \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}_{0,1}(y, t)$

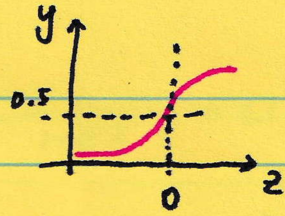
$\frac{\partial \mathcal{L}_{0,1}}{\partial w_j} = \frac{\partial \mathcal{L}_{0,1}}{\partial z} \cdot \frac{\partial z}{\partial w_j}$. observe that $\mathcal{L}_{0,1}$ is "not nice" to the differentiation process given z. $\Rightarrow$ gradient $= 0$

$\mathcal{L}_{SE} = \frac{1}{2}(z - t)^2 \Rightarrow$ residual is large when making a prediction with high confidence (i.e. $z = 10^{100000}$) ; $(z-t)^2$ will be huge.

Logistic function: $\sigma(z) = \frac{1}{1 + e^{-z}}$, σ = activation function.

$y = \sigma(z)$

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



$$\frac{\partial \mathcal{L}}{\partial w_j} = \frac{\partial \mathcal{L}}{\partial z} \cdot \frac{\partial z}{\partial w_j}$$

$\mathcal{L}_{CE}(y, t) = -t\log y - (1-t)\log(1-y)$

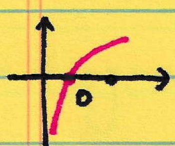
$z = w^T x, \quad y = \sigma(z) = \frac{1}{1+e^{-z}}$

$t=1; \ \mathcal{L}_{CE} = -\log y$

$t=0; \ \mathcal{L}_{CE} = -\log(1-y)$

$\mathcal{L}_{CE} = -t\log y - (1-t)\log(1-y)$

Log graph:



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

$$\mathcal{L}_{LCE} = \mathcal{L}_{CE}(\sigma(z), t) = -t \cdot \log\left(\frac{1}{1+e^{-z}}\right) - (1-t)\log\left(1 - \frac{1}{1+e^{-z}}\right)$$

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

$$= -t(0 - \log(1+e^{-z})) - (1-t)\log\left(\frac{e^{-z}}{1+e^{-z}}\right)$$

$$= -t(-\log(1+e^{-z}) - (1-t)\log\left(\frac{\cancel{e^{-z}}}{\cancel{1}+e^{-z}}\right) \quad \frac{1}{e^{+z}+1}$$

$$= t\log(1+e^{-z}) - (1-t)(\log(1) - \log(e^{+z}+1))$$

$$= t\log(1+e^{-z}) + (1-t)\log(e^{z}+1)$$

$$\mathcal{L}_{CE} = -t\log\left(\frac{1}{1+e^{-z}}\right) - (1-t)\log\left(1-\frac{1}{1+e^{-z}}\right)$$

$$= -t(\log 1 - \log(1+e^{-z})) - (1-t)\log\frac{e^{-z}}{1+e^{-z}}$$

$$= -t(0-\log(1+e^{-z})) - (1-t)(\log e^{-z} - \log(1+e^{-z}))$$

$$= t\log(1+e^{-z}) - (1-t)(-z-\log(1+e^{-z}))$$

$$= t\log(1+e^{-z}) - [-z-\log(1+e^{-z}) + tz + t\log(1+e^{-z})]$$

$$= t\log(1+e^{-z}) + z + \log(1+e^{-z}) - tz - t\log(1+e^{-z})$$

$$= z - tz + \log(1+e^{-z}) = z(1-t) + \log(1+e^{-z})$$

$$\frac{\partial \mathcal{L}_{CE}}{\partial w_j} = \frac{\partial \mathcal{L}_{CE}}{\partial y}^{①} \cdot \frac{\partial y}{\partial z}^{②} \cdot \frac{\partial z}{\partial w_j}^{③}.$$

$(\log x)' = \frac{1}{x}$

① $\frac{\partial \mathcal{L}_{CE}}{\partial y} = \frac{\partial}{\partial y}(-t\log y - (1-t)\log(1-y))$

$g(a) = \frac{u(a)}{v(a)}$

$= \frac{\partial}{\partial y}(-t\log y) - \frac{\partial}{\partial y}(1-t)\log(1-y)$

$\Rightarrow g' = \frac{u'v \cdot v'u}{v^2}$

$= -t\cdot\frac{1}{y} + (1-t)\cdot\frac{1}{1-y} = \boxed{\frac{-t}{y} + \frac{(1-t)}{(1-y)}}$

② $\frac{\partial y}{\partial z} = \frac{\partial}{\partial z}\frac{1}{1+e^{-z}}$ in which $u=1,\ v=1+e^{-z}$

$$= \frac{0+(1+e^{-z})'}{(1+e^{-z})^2} = \frac{e^{-z}}{(1+e^{-z})^2} = \boxed{y-y^2}$$

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

③ $\frac{\partial z}{\partial w_j} = \frac{\partial}{\partial w_j}(w x)$

$= \frac{\partial}{\partial w_j}[w_1x_1 + w_2x_2 + \cdots + w_jx_j \cdots + w_Dx_D]$

$= \boxed{x_j}$

$$\frac{\partial \mathcal{L}_{CE}}{\partial w_j} = \left(-\frac{t}{y} + \frac{1-t}{1-y}\right)\cdot y(1-y)\cdot x_j$$

Multi-class linear classification

for each output class $k \in K$; i.e. $k = cat$; $K = \{dog, cat, badger\}$

z_k = the z (raw val before activation) of the k = dog case

$$= \sum_{j=1}^{D} w_{k,j} x_j + b_k$$

Remark: Still the linear function but in the case of k. As a result we add the k subscript.

Activation function Softmax (compared to sigmoid)

$$y_k = \text{Softmax}(z_1, \cdots, z_k \cdots z_K)_k = \frac{e^{z_k}}{\sum_{k'} e^{z_{k'}}}$$

$$= \frac{e^{z_{k=cat}}}{e^{z_{k=cat}} + e^{z_{k=dog}} + e^{z_{k=badger}}}$$

Remark: z_k (raw val before the activation) = logits.

In multiclass scenario ($K > 2$); $\mathcal{L}_{CE}(y, t) = -\sum^{K} t_k \log y_k = -t^T \log y$

XOR $\psi(x) = \begin{pmatrix} x_1 \\ x_2 \\ x_1 x_2 \end{pmatrix}$

x_1	x_2	$\psi_1(x)$	$\psi_2(x)$	$\psi_3(x)$	t
0	0	0	0	0	0
0	1	0	1	0	1
1	0	1	0	0	1
1	1	1	1	1	0

$w_2 \psi_2(x) \geq 0$

$w_1 \psi_1(x) \geq 0$

$w_1 \psi_1(x) + w_2 \psi_2(x) + w_3 \psi_3(x) < 0$

$\Rightarrow$ linearly solvable.