Logistic Regression & Multi-class classification

binary linear classification $z = w^T x + b$. $y = \begin{cases} 1, & z \geq r \\ 0, & z < r \end{cases}$

$w^T x + b \geq r \Rightarrow w^T x + b - r \geq 0$.

Recall that we can incorporate b into the weight matrix w
$\Rightarrow z = w^T x$; $x_0 = 1$ & $x \in \mathbb{R}^{D+1}$

NOT

x_0	x_1	t
1	0	1
1	1	0

x_0 is always 1 because it's the dummy feature we use to incorporate b into w.

when $x_1 = 0$: $w_0 x_0 + w_1 x_1 \geq 0 \Rightarrow x_0 \geq 0$
$x_1 = 1$: $w_0 x_0 + w_1 x_1 < 0 \Rightarrow w_0 + w_1 < 0$.

there are many possible solutions

AND

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

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

$w_0 < 0$

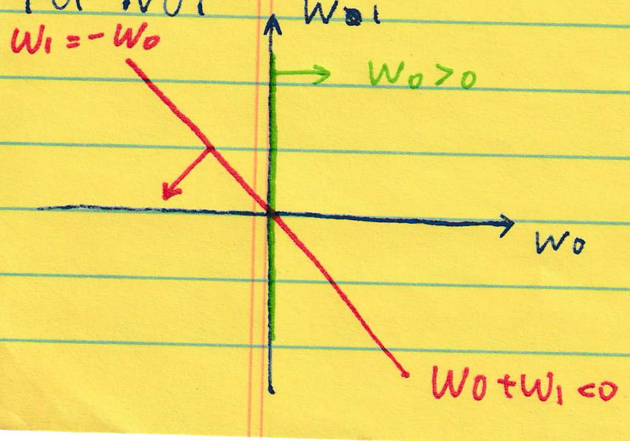
$w_0 + w_2 < 0$

$w_0 + w_1 < 0$

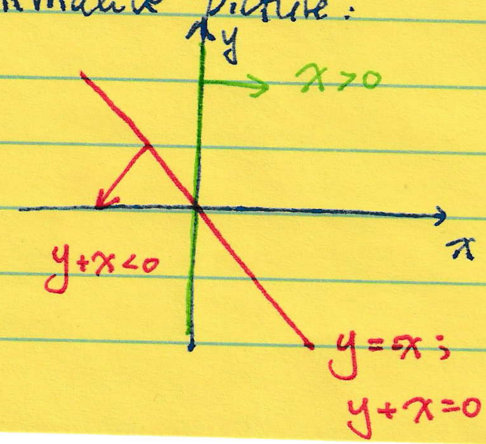
$w_0 + w_1 + w_2 > 0$

again there are many possible solutions

For NOT



an alternative picture:



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

$$\mathcal{J} = \frac{1}{N}\sum_i^N \mathcal{L}^i_{0,1} \quad \boxed{\mathbb{I}(y_i \neq t_i)} \qquad \frac{\partial \mathcal{L}_{0,1}}{\partial w_j} = \frac{\partial \mathcal{L}_{0,1}}{\partial z} \cdot \frac{\partial z}{\partial w_j}$$

$\frac{\partial \mathcal{L}_{0,1}}{\partial z}$ = (almost) zero anywhere it's defined. (refer to the figure on slide 14) $\Rightarrow$ unable to do gradient descent

$\mathcal{L}_{SE} = \frac{1}{2}(z-t)^2$; let's set the final prediction threshold to be 0.5.

$\Rightarrow$ if $z \geq 0.5$, predict positive

if $z < 0.5$, predict negative

Example: ① a sample is "very positive" (the model predicts positive with high confidence. $\Rightarrow$ z is large ($z = 1000$, for example)

② a sample is "not very positive" (the model predicts positive with low confidence. $\Rightarrow$ z is small but still above the threshold. ($z = 0.6$, for example).

in ①: $\mathcal{L}_{SE} = \frac{1}{2}(\cancel{y}\,z - t)^2 = \frac{1}{2}999.5^2$ $\Rightarrow$ the loss function hates when you make (often time) correct preditions with high confidence.

in ②: $\mathcal{L}_{SE} = \frac{1}{2}(z-t)^2 = \frac{1}{2}0.1^2$.

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

$\Rightarrow z = w^T x$, $y = \sigma(z)$, $\mathcal{L}_{SE} = \frac{1}{2}(y-t)^2$

the logistic activation function converts an arbitrary big/small z into the range $[0, 1]$. the bigger the z is, $\sigma(z) = y$ approaches 1, and the smaller the z is, vice versa.

$\frac{\partial \mathcal{L}}{\partial w_j} = \frac{\partial \mathcal{L}}{\partial z} \cdot \frac{\partial z}{\partial w_j}$ $\Rightarrow$ differentiable now!

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Gradient Descent of Logistic Regression.

$$\mathcal{L}(CE) = -t\log y - (1-t)\log(1-y)$$
$$y = \frac{1}{1+e^{-z}}, \quad z = w^{T}x$$

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

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

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

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

② $\frac{\partial y}{\partial z} = \frac{\partial}{\partial z}\left(\frac{1}{1+e^{-z}}\right)$ $\left(\frac{u}{v}\right)' = \frac{u'v - v'u}{v^2}$, whereas $u = 1$, $v = 1+e^{-z}$

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

Same

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

As a result, $\frac{\partial y}{\partial z} = y - y^2 = y(1-y)$

③ $\frac{\partial z}{\partial w_j} = \frac{\partial}{\partial w_j} \vec{w}^T x = \frac{\partial}{\partial w_j}(w_0 x_0 + w_1 x_1 + \cdots + w_j x_j + \cdots + w_{D+1} x_{D+1})$

$$= x_j.$$

$$\Rightarrow \frac{\partial \mathcal{L}_{CE}}{\partial w_j} = \frac{\partial \mathcal{L}_{CE}}{\partial y} \cdot \frac{\partial y}{\partial z} \cdot \frac{\partial z}{\partial w_j}$$

$$= \left(-\frac{t}{y} + \frac{(1-t)}{(1-y)}\right) \cdot y(1-y) \cdot x_j$$

$$w_j \leftarrow w_j - \alpha \cdot \frac{\partial \mathcal{J}}{\partial w_j} = w_j - \alpha \cdot \frac{\partial}{\partial w_j} \frac{1}{N} \sum_i^N \mathcal{L}_{CE}^i$$

$$= w_j - \alpha \cdot \frac{1}{N} \sum_i^N \left(-\frac{t_i}{y_i} + \frac{1-t_i}{1-y_i}\right) \cdot y_i(1-y_i)\, x_j^i$$

$$= w_j \frac{\alpha}{N} \sum_i^N \left(-\frac{t_i}{y_i} + \frac{1-t_i}{1-y_i}\right) \cdot y_i(1-y_i) \cdot x_j^i$$

Multi-class linear classification

for the k^{th} class. do a linear classifier $z_k = \sum_j^D w_{kj} \cdot x_j + b_k$

whereas your $k \in [1, \dots, K]$ whereas there are K classes.

$$y_i = \begin{cases} 1, & \text{if } i = \arg\max_k z_k \\ 0, & \text{if otherwise} \end{cases}$$

We want $\sum_k^K y_k = 1 \Rightarrow y_k = \text{softmax}(z_1 \cdots z_K)_k = \dfrac{e^{z_k}}{\sum_{k'} e^{z_{k'}}}$

$\mathcal{L}_{CE} = - \sum^K t_k \log y_k = - t^T \log y$