⊛ Exercise 1:

a)

The formula of the sigmoid function is:

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

The derivative of the sigmoid function:

$$\sigma'(x) = \left(\frac{1}{1+e^{-x}}\right)'$$

$$= \frac{-1}{(1+e^{-x})^2} \cdot (e^{-x})'$$

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

$$= \frac{1}{1+e^{-x}} \cdot \frac{e^{-x}}{1+e^{-x}}$$

$$= \frac{1}{1+e^{-x}} \cdot \left[1 - \frac{1}{1+e^{-x}}\right]$$

$$= \sigma(x)\left[1-\sigma(x)\right]$$

b) The formula for loss function in logistic regression is:

$$L(w) = \frac{-1}{N}\sum_{i=1}^{N} t^{(i)}\log y^{(i)} + (1-t^{(i)})\log(1-y^{(i)}) \quad (y=\sigma(x))$$

This is the cross-entropy loss function.

c)
$$\frac{\partial L(w)}{\partial w_j} = \frac{\partial}{\partial w_j}\left[\frac{-1}{N}\sum_{i=1}^{N} t^{(i)}\log(\sigma(x_w^{(i)})) + (1-t^{(i)})\log(1-\sigma(x_w^{(i)}))\right]$$

$$= \frac{-1}{N}\sum_{i=1}^{N} t^{(i)}\frac{\partial}{\partial w_j}\left[\log(\sigma(x_w^{(i)}))\right] + (1-t^{(i)})\frac{\partial}{\partial w_j}\left[\log(1-\sigma(x_w^{(i)}))\right] \quad (1)$$

Have:

$$\frac{\partial}{\partial w_j}\log(\sigma(x_w^{(i)})) = \frac{1}{\sigma(x^{(i)})}\frac{\partial}{\partial w_j}\sigma(x_w^{(i)}) \quad (2)$$

$$\frac{\partial}{\partial w_j}\log(1-\sigma(x_w^{(i)})) = \frac{1}{1-\sigma(x^{(i)})}\frac{\partial}{\partial w_j}(1-\sigma(x_w^{(i)}))$$

$$= \frac{-1}{1-\sigma(x^{(i)})}\frac{\partial}{w_j}\sigma(x_w^{(i)}) \quad (3)$$

Plug (2) and (3) to (1):

$$\frac{\partial L(w)}{\partial w_j} = \frac{-1}{N}\sum_{i=1}^{N} t^{(i)}\frac{1}{\sigma(x^{(i)})}\frac{\partial}{\partial w_j}\sigma(x_w^{(i)}) + (1-t^{(i)})\frac{-1}{1-\sigma(x^{(i)})}\frac{\partial}{\partial w_j}\sigma(x_w^{(i)})$$

$$= \frac{-1}{N}\sum_{i=1}^{N}\left[\frac{t^{(i)}}{\sigma(x_w^{(i)})} - \frac{(1-t^{(i)})}{1-\sigma(x_w^{(i)})}\right]\frac{\partial}{\partial w_j}\sigma(x_w^{(i)}) \quad (4)$$

Apply chain rule:

$$\frac{\partial}{\partial w_j}(\sigma(x^{(i)}w)) = \frac{\partial}{\partial z}(\sigma(z))\frac{\partial}{\partial w_j}(z(w))$$

where $z = x^{(i)}w \Rightarrow \frac{\partial}{\partial(z)}\sigma(z) = \sigma(z)(1-\sigma(z))$
$$= \sigma(x^{(i)}w)[1-\sigma(x^{(i)}w)].$$

and $\frac{\partial}{\partial w_j} z(w) = \frac{\partial}{\partial w_j}(x^{(i)}w) = x_j^{(i)}$

$$\Rightarrow \frac{\partial}{\partial w_j}(\sigma(x^{(i)}w)) = \sigma(x^{(i)}w)[1-\sigma(x^{(i)}w)]\, x_j^{(i)} \quad (5)$$

Plug (5) to (4):

$$\frac{\partial L(w)}{\partial w_j} = -\frac{1}{N}\sum_{i=1}^{N}\left[\frac{t^{(i)}}{\sigma(x^{(i)}w)} - \frac{(1-t^{(i)})}{1-\sigma(x^{(i)}w)}\right]\cdot \sigma(x^{(i)}w)[1-\sigma(x^{(i)}w)]x_j^{(i)}$$

$$= \frac{-1}{N}\sum_{i=1}^{N}\left[t^{(i)}(1-\sigma(x^{(i)}w)) - (1-t^{(i)})\sigma(x^{(i)}w)\right]x_j^{(i)}$$

$$= \frac{-1}{N}\sum_{i=1}^{N}\left[t^{(i)} - t^{(i)}\sigma(x^{(i)}w) - \sigma(x^{(i)}w) + t^{(i)}\sigma(x^{(i)}w)\right]x_j^{(i)}$$

$$= \frac{-1}{N}\sum_{i=1}^{N}\left[t^{(i)} - \sigma(x^{(i)}w)\right]x_j^{(i)}$$

$$= \frac{1}{N}\sum_{i=1}^{N}\left[\sigma(x^{(i)}w) - t^{(i)}\right]x_j^{(i)}$$

The vector calculus form:

$$\frac{\partial L}{\partial w} = \frac{1}{N}X^T(\sigma(Xw) - t).$$

(*) Exercise 3:

There are some reasons MSE isn't used as the loss function of logistic regression.

1) While maximizing the probability, if we assume output outcomes from Gaussian distribution, then it can be proven that it equivalent to minimizing MSE loss. But we take output dist as Bernuolli so Binary Cross Entropy loss kicks in. If we use MSE, it would be mismatch in distribution of output.

2) MSE loss in logistic regression is non-convex function while BCE is convex.

To prove these, we need to prove the Hessian matrix of the loss function (2^{nd} order derivative) is positive semidefinite.

+) The 1^{st} order derivative of BCE is:

$$\frac{\partial L}{\partial w_j} = \frac{1}{N}\sum_{i=1}^{N} \left(\sigma(x^{(i)}w) - t^{(i)}\right) x_j^{(i)}.$$

Then:

$$H_{jk} = \frac{\partial^2 L}{\partial w_j \partial w_k} = \frac{1}{N}\sum_{i=1}^{N} \frac{\partial [\sigma(x^{(i)}w) - t^{(i)}] x_j^{(i)}}{\partial w_k}$$

$$= \frac{1}{N}\sum_{i=1}^{N} \frac{\partial \sigma(x^{(i)}w) x_j^{(i)}}{\partial w_k} - \frac{\partial t^{(i)} x_j^{(i)}}{\partial w_k}$$

$$= \frac{1}{N}\sum_{i=1}^{N} x_j^{(i)} \sigma(x^{(i)}w)[1-\sigma(x^{(i)}w)] x_k^{(i)}.$$

In the vector calculus form: $H = X^T \sigma(Xw)[1-\sigma(Xw)]^T X$

$$= \sigma(Xw)[1-\sigma(Xw)]^T XX^T$$

$$= \sigma(Xw)[1-\sigma(Xw)]^T \|X\|_2^2$$

Since $0 \leq \sigma(z) \leq 1 \Rightarrow \sigma(Xw)[1-\sigma(Xw)]^T \geq \vec{0}$

$\|X\|_2^2 \geq 0 \quad \forall x_{i=\overline{1,N}}$

$\Rightarrow \sigma(Xw)[1-\sigma(Xw)]^T \|X\|_2^2 \geq 0.$

$\Leftrightarrow \quad H \geq 0$

Thus the Hessian matrix is semidefinite positive and BCE is convex.

+) The 1st order of MSE loss function is:

$$\frac{\partial L}{\partial w_j} = \frac{1}{N}\sum_{i=1}^{N} \frac{\partial (t^{(i)} - \sigma^{(i)})^2}{\partial w_j}$$

$$= \frac{-2}{N}\sum_{i=1}^{N} (t^{(i)} - \sigma^{(i)}) \frac{\partial \sigma^{(i)}}{\partial x^{(i)}w} \frac{\partial x^{(i)}w}{\partial w_j}$$

$$= \frac{-2}{N}\sum_{i=1}^{N} (t^{(i)} - \sigma^{(i)})\sigma^{(i)}(1-\sigma^{(i)}) \frac{\partial x^{(i)}w}{\partial w_j}$$

$$= \frac{-2}{N}\sum_{i=1}^{N} (t^{(i)}\sigma^{(i)} - \sigma^{2(i)})(1-\sigma^{(i)}) \frac{\partial x^{(i)}w}{\partial w_j}$$

$$= \frac{-2}{N}\sum_{i=1}^{N} (t^{(i)}\sigma^{(i)} - t^{(i)}\sigma^{2(i)} - \sigma^{2(i)} + \sigma^{3(i)}) \frac{\partial x^{(i)}w}{\partial w_j} \quad (1)$$

$$= \frac{-2}{N}\sum_{i=1}^{N} (t^{(i)}\sigma^{(i)} - t^{(i)}\sigma^{2(i)} - \sigma^{2(i)} + \sigma^{3(i)}) x_j^{(i)}$$

Then,

$$H_{jk} = \frac{\partial^2 L}{\partial w_j \partial w_k} = \frac{\partial}{\partial w_k} \frac{-2}{N}\sum_{i=1}^{N}\left(t^{(i)}\sigma^{(i)} - t^{(i)}\sigma^{2(i)} - \sigma^{2(i)} + \sigma^{3(i)}\right)x_j^{(i)}$$

$$= \frac{-2}{N}\sum_{i=1}^{N}\left(t^{(i)}\frac{\partial \sigma^{(i)}}{\partial w_k} - t^{(i)}\frac{\partial \sigma^{2(i)}}{\partial w_k} - \frac{\partial \sigma^{2(i)}}{\partial w_k} + \frac{\partial \sigma^{3(i)}}{\partial w_k}\right)x_j^{(i)}$$

$$= \frac{-2}{N}\sum_{i=1}^{N}\left(t^{(i)} - 2t^{(i)}\sigma^{(i)} - 2\sigma^{(i)} + 3\sigma^{2(i)}\right)\frac{\partial \sigma^{(i)}}{\partial w_k}x_j^{(i)}$$

$$= \frac{-2}{N}\sum_{i=1}^{N}\left(t^{(i)} - 2t^{(i)}\sigma^{(i)} - 2\sigma^{(i)} + 3\sigma^{2(i)}\right)\underbrace{\sigma^{(i)}(1-\sigma^{(i)})x_k^{(i)}x_j^{(i)}}_{g}$$

Since $0 \le \sigma^{(i)} \le 1$; $\sigma^{(i)}(1-\sigma^{(i)})x_k^{(i)}x_j^{(i)} \geq 0$.

Since $t^{(i)} \in \{0, 1\}$:

+) If $t^{(i)} = 0$:

$$H_{jk} = \frac{-2}{N}\sum_{i=1}^{N}\left(-2\sigma^{(i)} + 3\sigma^{2(i)}\right)g$$

$$= -\frac{2}{N}\sum_{i=1}^{N}\left(-2 + 3\sigma^{(i)}\right)\sigma^{(i)}g$$

since $(-2 + 3\sigma^{(i)})$ changes sign when $\sigma^{(i)} < \frac{2}{3} \Rightarrow H_{jk} < 0$

$\Rightarrow$ H is not pos semi definite.

+) If $t^{(i)} = 1 \Rightarrow H_{jk} = \frac{-2}{N}\sum_{i=1}^{N}\left(1 - 2\sigma^{(i)} - 2\sigma^{(i)} + 3\sigma^{2(i)}\right)g$

$$= \frac{-2}{N}\sum_{i=1}^{N}\left(1 - 4\sigma^{(i)} + 3\sigma^{2(i)}\right)g$$

$\sigma^{(i)} \in [0;1] \,/\, (1-4\sigma^{(i)}+3\sigma^{2(i)}) \in [-\frac{1}{3};1] \Rightarrow H$ is not positive semidefinite.

Thus, the MSE loss function isn't convex.

3) MSE doesn't penalize misclassification enough.

For example, if we have perfect mismatch which ~~y~~ $t=1$ and $y=0$, then:

$MSE = (1-0)^2 = 1$

$BCE = -1\log(0) - 0\log(1) = -\infty$

$\Rightarrow$ When using gradient descent, models would translate it to steeper gradient and a faster correction of weights

$\Rightarrow$ Faster convergent.