1) Sigmoid function derivatives $\sigma(\eta)$

The sigmoid function is written as $\sigma(\eta) = \frac{1}{1+e^{-\eta}} = \frac{e^{\eta}}{1+e^{\eta}}$, where $0 < \sigma(\eta) < 1$.
Show that $\frac{d\sigma(\eta)}{d\eta} = \sigma(\eta)\left[1 - \sigma(\eta)\right]$ and $\frac{d\log\sigma(\eta)}{d\eta} = 1 - \sigma(\eta)$.

Solution

$$\sigma(\eta) = \frac{1}{1+e^{-\eta}} = \frac{e^{\eta}}{1+e^{\eta}} \quad , \quad 0 < \sigma(\eta) < 1$$

$$\frac{d\sigma(\eta)}{d\eta} = -\frac{-e^{-n}}{(1+e^{-n})^2} = \frac{e^{-n}}{(1+e^{-n})^2} = \frac{1}{1+e^{-n}}\left(\frac{e^{-n}}{1+e^{-n}}\right) = \frac{1}{1+e^{-n}}\left(1 - \frac{1}{1+e^{-n}}\right) = \sigma(\eta)\left[1-\sigma(\eta)\right]$$

$$\frac{d\log\sigma(\eta)}{d\eta} = \frac{1}{\sigma(\eta)} \cdot \frac{d\sigma(\eta)}{d\eta} = 1 - \sigma(\eta)$$

2) Logistic Regression Likelihood & Cross-Entropy

Let $\mathcal{D} = \{(\mathbf{x_1}, y_1), \ldots, (\mathbf{x_N}, y_N)\}$, where $\mathbf{x_n} \in \mathbb{R}^D$ and $y_n \in \mathbb{R}$, be the training data of a binary logistic regression model with weights $\mathbf{w} \in \mathbb{R}^D$. The probability of sample $(\mathbf{x_n}, y_n)$ belonging to class 1 is $p(y=1|\mathbf{x},\mathbf{w}) = \sigma(\mathbf{w}^T\mathbf{x})$, while the probability of belonging to class 0 is $p(y=0|\mathbf{x},\mathbf{w}) = 1 - \sigma(\mathbf{w}^T\mathbf{x})$. Compute the likelihood $\mathcal{L}(\mathcal{D}|\mathbf{w})$ of data $\mathcal{D}$ given the model parameters $\mathbf{w}$, as well as the cross-entropy error $\mathcal{E}(\mathbf{w}) = -\log\mathcal{L}(\mathcal{D}|\mathbf{w})$.

Solution
Input: $\mathbf{x} \in \mathbb{R}^D$
Output: $y \in \{0, 1\}$
Training data: $\mathcal{D} = \{(\mathbf{x_1}, y_1), \ldots, (\mathbf{x_N}, y_N)\}$
Model:

$$p(y=1|\mathbf{x},\mathbf{w}) = \sigma(\mathbf{w}^T\mathbf{x})$$

$$p(y=0|\mathbf{x},\mathbf{w}) = 1 - \sigma(\mathbf{w}^T\mathbf{x}), \;\; \sigma(\eta) = \frac{1}{1+e^{-\eta}}$$

$$f(\mathbf{x}) : \mathbf{x} \to y, \;\; f(\mathbf{x}) = \begin{cases} 1, & p(y=1|\mathbf{x},\mathbf{w}) > 0.5 \\ 0, & \text{otherwise} \end{cases} = \begin{cases} 1, & \sigma(\mathbf{w}^T\mathbf{x}) > 0.5 \\ 0, & \text{otherwise} \end{cases}$$

Model parameters: Weights $\mathbf{w} \in \mathbb{R}^D$ (to be learned)
Data likelihood for 1 training sample:

$$p(y_n|\mathbf{x_n},\mathbf{w}) = \left\{ \begin{array}{ll} \sigma(\mathbf{w}^T\mathbf{x_n}), & y_n = 1 \\ 1 - \sigma(\mathbf{w}^T\mathbf{x_n}), & y_n = 0 \end{array} \right\} = \left[\sigma(\mathbf{w}^T\mathbf{x_n})\right]^{y_n} \left[1 - \sigma(\mathbf{w}^T\mathbf{x_n})\right]^{1-y_n}$$

Data likelihood for all training data:

$$L(\mathcal{D}|\mathbf{w}) = \prod_{n=1}^{N} p(y_n|\mathbf{x_n},\mathbf{w}) = \prod_{n=1}^{N} \left[\sigma(\mathbf{w}^T\mathbf{x_n})\right]^{y_n} \left[1 - \sigma(\mathbf{w}^T\mathbf{x_n})\right]^{1-y_n}$$

Log-likelihood for all training data:

$$l(\mathcal{D}|\mathbf{w}) = \sum_{n=1}^{N} \left\{ y_n \log\left[\sigma(\mathbf{w}^T\mathbf{x_n})\right] + (1-y_n)\log\left[1-\sigma(\mathbf{w}^T\mathbf{x_n})\right] \right\}$$

Cross-entropy error (negative log-likelihood):

$$\mathcal{E}(\mathbf{w}) = -\sum_{n=1}^{N} \left\{ y_n \log\left[\sigma(\mathbf{w}^T\mathbf{x_n})\right] + (1-y_n)\log\left[1-\sigma(\mathbf{w}^T\mathbf{x_n})\right] \right\}$$

Logistic Regression - Optimization

3a) Show that the first order derivative (i.e., gradient vector) of the cross-entropy function is $\nabla\mathcal{E}(\mathbf{w}) = \frac{\vartheta\mathcal{E}(\mathbf{w})}{\vartheta\mathbf{w}} = \sum_{n=1}^{N} \underbrace{\left(\sigma(\mathbf{w}^T\mathbf{x_n}) - y_n\right)}_{\text{error}} \mathbf{x_n}$

<u>Solution</u>

$$\begin{aligned}
\nabla\mathcal{E}(\mathbf{w}) &= -\sum_{n=1}^{N} \left\{ y_n \left[1-\sigma(\mathbf{w}^T\mathbf{x_n})\right]\mathbf{x_n} - (1-y_n)\left[1-\left(1-\sigma(\mathbf{w}^T\mathbf{x_n})\right)\right]\mathbf{x_n} \right\} \\
&= -\sum_{n=1}^{N} \left\{ y_n \left[1-\sigma(\mathbf{w}^T\mathbf{x_n})\right]\mathbf{x_n} - (1-y_n)\sigma(\mathbf{w}^T\mathbf{x_n})\mathbf{x_n} \right\} \\
&= -\sum_{n=1}^{N} \left[y_n - y_n\sigma(\mathbf{w}^T\mathbf{x_n}) - \sigma(\mathbf{w}^T\mathbf{x_n}) + y_n\sigma(\mathbf{w}^T\mathbf{x_n}) \right]\mathbf{x_n} \\
&= \sum_{n=1}^{N} \underbrace{\left(\sigma(\mathbf{w}^T\mathbf{x_n}) - y_n\right)}_{\text{error}} \mathbf{x_n}
\end{aligned}$$

No closed-form solution that minimizes the cross-entropy function.
We use an approximate method, e.g. gradient descent, so we need to compute $\nabla\mathcal{E}(\mathbf{w})$. Gradient descent update: $\mathbf{w_{k+1}} := \mathbf{w_k} - \alpha(k)\nabla\mathcal{E}(\mathbf{w})$

3b) Show that the Hessian of the cross-entropy function is $\mathbf{H} = \frac{\vartheta^2\mathcal{E}(\mathbf{w})}{\vartheta^2\mathbf{w}} = \nabla\left((\nabla\mathcal{E}(\mathbf{w}))^T\right) = \sum_{n=1}^{N} \sigma(\mathbf{w}^T\mathbf{x_n}) \cdot \left(1-\sigma(\mathbf{w}^T\mathbf{x_n})\right) \cdot \left(\mathbf{x_n} \cdot \mathbf{x_n}^T\right)$ and show that it is positive semi-definite.

<u>Solution</u>
$\mathbf{H} = \frac{\vartheta^2\mathcal{E}(\mathbf{w})}{\vartheta^2\mathbf{w}} = \nabla\left((\nabla\mathcal{E}(\mathbf{w}))^T\right) = \nabla\left(\sum_{n=1}^{N}\left(\sigma(\mathbf{w}^T\mathbf{x_n}) - y_n\right)\mathbf{x_n}^T\right)$

$$\begin{aligned}
\mathbf{H} &= \frac{\vartheta}{\vartheta\mathbf{w}}\left[\sum_{n=1}^{N}\left(\sigma(\mathbf{w}^T\mathbf{x_n}) \cdot \mathbf{x_n}^T - y_n\mathbf{x_n}^T\right)\right] \\
&= \sum_{n=1}^{N} \frac{\vartheta}{\vartheta\mathbf{w}}\left[\sigma(\mathbf{w}^T\mathbf{x_n})\right] \cdot \mathbf{x_n}^T \quad \text{(chain rule)} \\
&= \sum_{n=1}^{N} \underbrace{\sigma(\mathbf{w}^T\mathbf{x_n})}_{\in[0,1]} \cdot \underbrace{\left(1-\sigma(\mathbf{w}^T\mathbf{x_n})\right)}_{\in[0,1]} \cdot \underbrace{\left(\mathbf{x_n} \cdot \mathbf{x_n}^T\right)}_{\in\mathcal{R}^{D\times D}}
\end{aligned}$$

For all $\mathbf{v} \in \mathbb{R}^D$, substituting $\mu_n = \sigma(\mathbf{w}^T \mathbf{x_n}) \left(1 - \sigma(\mathbf{w}^T \mathbf{x_n})\right) \geq 0$, we have:

$$\mathbf{v}^T \mathbf{H} \mathbf{v} = \mathbf{v}^T \left(\sum_{n=1}^{N} \mu_n \mathbf{x_n} \mathbf{x_n}^T \right) \mathbf{v} = \sum_{n=1}^{N} (\mu_n \mathbf{x_n}^T \mathbf{v})^T (\mathbf{x_n}^T \mathbf{v}) = \sum_{n=1}^{N} \mu_n \|\mathbf{x_n}^T \mathbf{v}\|_2^2 \geq 0$$