Exercise 1

Verify the relation

$$\frac{d\sigma}{da} = \sigma(1-\sigma)$$

For the derivative of the logistic sigmoid function defined by

$$\frac{d\sigma}{da} = \sigma(1-\sigma)\sigma(a) = \frac{1}{1+e^{-a}}$$

For this, the chain rule can be applied:

$$(f(g(a)))' = f'(g(a)) \cdot g'(a)$$

Whereas:

$$\begin{aligned} f(a) = a^{-1} \quad &\Longrightarrow \quad f'(a) = -a^{-2} \\ g(a) = 1 + e^{-a} \quad &\Longrightarrow \quad g'(a) = -e^{-a} \\ \sigma(a) = (1+e^{-a})^{-1} &= f(g(a)) \end{aligned}$$

Therefore, it holds:

$$\begin{aligned} \sigma'(a) &= -(1+e^{-a})^{-2} \cdot -(e^{-a}) \\ &= \frac{e^{-a}}{(1+e^{-a})^{-2}} \\ &= \frac{1}{1+e^{-a}} \cdot \frac{e^{-a}}{1+e^{-a}} \\ &= \sigma(a) \cdot \frac{1+e^{-a}-1}{1+e^{-a}} \\ &= \sigma(a) \cdot (1-\sigma(a)) \end{aligned}$$

□

Exercise 2

By making use of the expression for the derivative of the logistic sigmoid from exercise 1, show that the derivative of the error function for the logistic regression model is given by

$$\nabla E(w) = \sum_{n=1}^{N} (y_n - t_n)\phi(n)$$

Whereas:

$$E(w) = -\sum_{n=1}^{N}\{t_n ln(y_n) + (1-t_n)ln(1-y_n)\}$$
$$\nabla E(w) = \frac{\partial E}{\partial w}$$
$$y_n = \sigma(s_n)$$
$$s_n = w^T \phi_n$$

The simplicity of this derivative is the main reason why the sigmoid (or its multi-class version, the softmax) activation function is used in combination with the cross-entropy (the negative-log likelihood) to train logistic classificators. Especially, it is useful when training neural networks, since the calculation of the backpropagation is kept as well very simple. Again, this property becomes evident when applying the chain rule:

$$\frac{\partial E}{\partial w} = \frac{\partial E}{\partial y}\frac{\partial y}{\partial s}\frac{\partial s}{\partial w}$$

Remembering that $f(x) = ln(x) \Rightarrow f'(x) = x^{-1}$, and what we already proved in Exercise 1, it holds:

$$\frac{\partial E}{\partial y} = \frac{-t}{y} + \frac{1-t}{1-y} = \frac{y-t}{y(1-y)}$$
$$\frac{\partial y}{\partial s} = y(1-y)$$
$$\frac{\partial s}{\partial w} = \phi$$

Finally, multiplying it all, returns the derivative:

$$\frac{\partial E}{\partial w} = \frac{y-t}{y(1-y)} \cdot y(1-y) \cdot \phi = (y-t)\phi$$

This holds for every sample. And since the derivative of a sum is the sum of derivatives, the derivative of $E(w)$ for the whole dataset with N samples, is the sum of this formula over the whole dataset:

$$\frac{\partial}{\partial w}\Big(-\sum_{n=1}^{N}\{t_n ln(y_n) + (1-t_n)ln(1-y_n)\}\Big) = -\sum_{n=1}^{N}\{(y_n - t_n)\phi_n\}$$

□