Pre-class activities
Artifical Neural Networks

Emmanuel Rachelson

1 Derivation: the chain rule and total derivatives

1. Consider a function $f : \mathbb{R}^p \to \mathbb{R}^q$. Recall the expression of this function's Jacobian matrix.

   The Jacobian matrix of f in $\hat{x}$ is the $q \times p$ matrix of partial derivatives.

   $$\begin{bmatrix} \frac{\partial f_1}{\partial x_1}(\hat{x}) & \cdots & \frac{\partial f_1}{x_p}(\hat{x}) \\ \vdots & \ddots & \vdots \\ \frac{\partial f_q}{\partial x_1}(\hat{x}) & \cdots & \frac{\partial f_q}{x_p}(\hat{x}) \end{bmatrix}.$$

   The kth row of the Jacobian of f evaluated in $\hat{x}$ is the vector of partial derivatives of f_k in $\hat{x}$, that is the transpose of the gradient of f_k in $\hat{x}$.

2. The total derivative $Df_{\hat{x}}(h)$ of f in $\hat{x}$ is a linear operator. Recall its definition (for the same $f : \mathbb{R}^p \to \mathbb{R}^q$ function as above).

   The total derivative of f in $\hat{x}$ is the linear mapping: $Df_{\hat{x}} : \mathbb{R}^p \to \mathbb{R}^q$, where $h \in \mathbb{R}^p$ such that

   $$f(\hat{x} + h) = f(\hat{x}) + Df_{\hat{x}}(h) + o(\|h\|).$$

   The matrix of the total derivative is the Jacobian of f in $\hat{x}$. Thus, for $h = [h_1, \ldots, h_p]^T$:

   $$Df_{\hat{x}}(h) = \begin{bmatrix} \frac{\partial f_1}{\partial x_1}(\hat{x}) & \cdots & \frac{\partial f_1}{x_p}(\hat{x}) \\ \vdots & \ddots & \vdots \\ \frac{\partial f_q}{\partial x_1}(\hat{x}) & \cdots & \frac{\partial f_q}{x_p}(\hat{x}) \end{bmatrix} \begin{bmatrix} h_1 \\ \vdots \\ h_p \end{bmatrix} = Df_{\hat{x}} \cdot h.$$

3. Consider two functions, $g : \mathbb{R} \to \mathbb{R}^p$ and $f : \mathbb{R}^p \to \mathbb{R}$. Let $F = f \circ g$ be the composite function such that $F(x) = f(g(x))$. Write the derivative of F with respect to x as an expression of the partial derivatives of f and g.

   The chain rule tells us that:

   $$DF_{\hat{x}} = D(f \circ g)_{\hat{x}} = Df_{g(\hat{x})} Dg_{\hat{x}}.$$

This can be interpreted directly in terms of matrix multiplication.

But $f : \mathbb{R}^p \to \mathbb{R}$, so:

$$Df_{g(\hat{x})} = \begin{bmatrix} \frac{\partial f}{\partial x_1}(g(\hat{x})) & \cdots & \frac{\partial f}{\partial x_p}(g(\hat{x})) \end{bmatrix}.$$

Similarly, $g : \mathbb{R} \to \mathbb{R}^p$, so:

$$Dg_{\hat{x}} = \begin{bmatrix} \frac{dg_1}{dx}(\hat{x}) \\ \vdots \\ \frac{dg_p}{dx}(\hat{x}) \end{bmatrix}.$$

Consequently, and as expected, $DF_{\hat{x}}$ is a scalar:

$$DF_{\hat{x}} = \frac{dF}{dx}(\hat{x}) = \sum_{k=1}^{p} \frac{\partial f}{\partial x_k}(g(\hat{x}))\frac{dg_k}{dx}(\hat{x}).$$

4. Now suppose that in the example above, all the g_k functions are identity functions, that is $g(x) = [x, \ldots, x]^T$. How does the total derivative of F simplify?

   In this case, $\frac{dg_k}{dx}(\hat{x}) = 1$ and thus:

$$DF_{\hat{x}} = \frac{dF}{dx}(\hat{x}) = \sum_{k=1}^{p} \frac{\partial f}{\partial x_k}(g(\hat{x})).$$

5. Finally, consider two functions $g : \mathbb{R} \to \mathbb{R}^p$ and $f : \mathbb{R}^p \to \mathbb{R}^q$. As previously, lets $F = f \circ g$ be the composite function. Write the total derivative of F as an expression of the partial derivatives of f and g.

   Just as before:

$$DF_{\hat{x}} = D(f \circ g)_{\hat{x}} = Df_{g(\hat{x})}Dg_{\hat{x}}.$$

   Now $Df_{g(\hat{x})}$ is the $q \times p$ matrix:

$$Df_{g(\hat{x})} = \begin{bmatrix} \frac{\partial f_1}{\partial x_1}(g(\hat{x})) & \cdots & \frac{\partial f_1}{x_p}(g(\hat{x})) \\ \vdots & \ddots & \vdots \\ \frac{\partial f_q}{\partial x_1}(g(\hat{x})) & \cdots & \frac{\partial f_q}{x_p}(g(\hat{x})) \end{bmatrix}.$$

   Let's write $\frac{\partial f}{\partial x_k}(g(\hat{x}))$ the kth column in this matrix. Then we have:

$$Df_{g(\hat{x})} = \begin{bmatrix} \frac{\partial f}{\partial x_1}(g(\hat{x})) & \cdots & \frac{\partial f}{\partial x_p}(g(\hat{x})) \end{bmatrix}.$$

And:

$$Dg_{\hat{x}} = \begin{bmatrix} \frac{dg_1}{dx}(\hat{x}) \\ \vdots \\ \frac{dg_p}{dx}(\hat{x}) \end{bmatrix}.$$

So the same writing as above still holds:

$$DF_{\hat{x}} = \sum_{k=1}^{p} \frac{\partial f}{\partial x_k}(g(\hat{x}))\frac{dg_k}{dx}(\hat{x}).$$

Recall that $\frac{\partial f}{\partial x_k}(g(\hat{x})) \in \mathbb{R}^q$ *and so* $DF_{\hat{x}} \in \mathbb{R}^q$ *as expected.*