w02-functions-2

Magnus Chr. Hvidtfeldt

Technical University of Denmark, Lyngby, DK,

s255792@dtu.dk

1 Hessian matrix

The double partial derivate of scalar functions $f : \mathbb{R}^n \to \mathbb{R}$ (only one variable functions can we find the hesse matrix). We defined hesse matrix as a $n \times n$ matrix as such

$$\mathbf{H}_f(x) = \begin{bmatrix} \frac{\partial^2 f}{\partial^2 x} & \frac{\partial^2 f}{\partial x \partial y} \\ \frac{\partial^2 f}{\partial y \partial x} & \frac{\partial^2 f}{\partial^2 y} \end{bmatrix} \tag{1.1}$$

Where $\frac{\partial^2 f}{\partial x \partial y} = \frac{\partial^2 f}{\partial y \partial x}$.

2 Jacobian matrix

Given $\mathbf{f} = (f_1, f_2, \ldots, f_k)$ for $\mathbf{x} = (x_1, \ldots, x_n)$, we have gradient vectors for f_i as row vectors for a $k \times n$ matrix as follows

$$\mathbf{J}_f(x) = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \frac{\partial f_2}{\partial x_1} & \frac{\partial f_2}{\partial x_2} & \cdots & \frac{\partial f_2}{\partial x_n} \\ \cdots & & & \\ \frac{\partial f_k}{\partial x_1} & \frac{\partial f_k}{\partial x_2} & \cdots & \frac{\partial f_k}{\partial x_n} \end{bmatrix} \tag{2.1}$$

3 Chain rule

Given $h(x) = g(f(x))$ we have the simple chain rule

$$h'(x) = \nabla g(f(x))^T * f'(x) \tag{3.1}$$

Or the generalized chain rule:

$$\mathbf{J}_h(x_0) = \mathbf{J}_g(y_0) \cdot \mathbf{J}_f(x_0) \tag{3.2}$$

Where $\mathbf{J}$ is the jacobian.

4 Directional derivative

Given $u = v/\|v\|$, we have

$$\nabla_u f(x) = \langle u, \nabla f(x) \rangle \tag{4.1}$$

If u is a unit vector. Remember that you can scale a vector down to be a unit vector, see notes.

5 Differentiability of vector functions

Let $f : U \to \mathbb{R}$, we say f is differentiable at x_0 if $\varepsilon : \mathbb{R}^n \to \mathbb{R}$ s.t. $\varepsilon(h) = 0$ as $h \to 0$. Thus we have

$$f(x_0 + h) - f(x_0) - ch - \varepsilon(h)\|h\| = 0 \tag{5.1}$$

Where $ch = L(h) = \mathbf{J}_f(x_0)h$.

5.1 Continuity

Let $U \subset \mathbb{R}^n$ be an open set and let $f : U \to \mathbb{R}$ be a scalar function. If the partial derivates exist at all points and are continuous, then f is differentiable for all $x \in U$.

Note: Linear and affine maps are differentiable.

5.2 C^1 vector function

Let U be an open set in $\mathbb{R}^n$. A vector function $f : U \to \mathbb{R}^n$ is called differentiable if it is differentiable at all points. If all partial derivatives appearing in the Jacobian matrix are continuous functions, then the vector function f is said to be continuously differentiable, or, f is said to be a C^1 vector function.

6 Training a neural network

Let $\Phi_\theta : \mathbb{R}^{n_0} \to \mathbb{R}^{n_L}$ denote a neural network with L layers, where the input layer has n_0 neurons and the output has n_L neurons, and $\theta = \{A_l, b_l\}_{l=1}^L$ denotes weights and biases. Suppose we want to approximate and unknown target map $y : \mathbb{R}^{n_0} \to \mathbb{R}^{n_L}$ from a finite training set $\{x^{(i)}, y(x^{(i)})\}_{i=1}^m$ by choosing θ so that $\Phi_\theta(x^{(i)}) \approx y(x^{(i)})$ for all i.

To measure approximation quality, introduce a loss function $L : \mathbb{R}^{n_L} \times \mathbb{R}^{n_L} \to \mathbb{R}$, the mean squared error

$$L(\Phi_\theta(x), y(x)) = \frac{1}{2}\|\Phi_\theta(x) - y(x)\|^2 \tag{6.1}$$

The overall loss function $R : \mathbb{R}^N \to \mathbb{R}$ is

$$R(\theta) = \frac{1}{m}\sum_{i=1}^{m} L(\Phi_\theta(x^{(i)}, y(x^{(i)}))) \tag{6.2}$$

Where N is the total number of parameters in θ.

Training the neural network reduces the overall loss $R(\theta)$. This is usually done by iterative gradient-based method: compute the gradient $\nabla_\theta R(\theta)$ and update parameters by gradient descent

$$\theta_{k+1} = \theta_k - \eta_k \nabla_\theta R(\theta_k) \tag{6.3}$$

With learning rate $\eta_k > 0$.