Machine Learning Lecture Notes

Predrag Radivojac

February 19, 2015

1 Newton-Raphson method

A function $f(x)$ in the neighborhood of point x_0, can be approximated using the Taylor series as

$$f(x) = \sum_{n=0}^{\infty} \frac{f^{(n)}(x_0)}{n!}(x - x_0)^n,$$

where $f^{(n)}(x_0)$ is the n-th derivative of function $f(x)$ evaluated at point x_0. Also, $f(x)$ is considered to be infinitely differentiable. For practical reasons, we will approximate this function using the first three terms of the series as

$$f(x) \approx f(x_0) + (x - x_0)f'(x_0) + \frac{1}{2}(x - x_0)^2 f''(x_0).$$

The optimum of this function can be found by finding the first derivative and setting it to zero (technically, one should check the second derivative as well)

$$f'(x) \approx f'(x_0) + (x - x_0)f''(x_0) = 0.$$

Solving this equation for x gives us

$$x = x_0 - \frac{f'(x_0)}{f''(x_0)}.$$

Note that the approach assumes that a good enough solution x_0 already exists. However, this equation, also provides a basis for an iterative process in finding the optimum of function $f(x)$. For example, if $x^{(i)}$ is the value of x in the i-th step, then the value in step $i + 1$ can be obtained as

$$x^{(i+1)} = x^{(i)} - \frac{f'(x^{(i)})}{f''(x^{(i)})}. \tag{1}$$

This method is called the Newton-Raphson method of optimization. We can generalize this approach to functions of vector variables $\mathbf{x} = (x_1, x_2, \ldots, x_k)$. The Taylor approximation for a vector function can be written as

$$f(\mathbf{x}) \approx f(\mathbf{x}_0) + \nabla f(\mathbf{x}_0)^T \cdot (\mathbf{x} - \mathbf{x}_0) + \frac{1}{2} (\mathbf{x} - \mathbf{x}_0)^T \cdot H_{f(\mathbf{x}_0)} \cdot (\mathbf{x} - \mathbf{x}_0),$$

where

$$\nabla f(\mathbf{x}) = \left(\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}, ..., \frac{\partial f}{\partial x_k} \right)$$

is the gradient of function f and

$$H_{f(\mathbf{x})} = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_k} \\ \frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial x_2^2} & & \\ \vdots & & \ddots & \\ \frac{\partial^2 f}{\partial x_k \partial x_1} & & & \frac{\partial^2 f}{\partial x_k^2} \end{bmatrix}$$

is the Hessian matrix of function f. Here, the gradient of f and its Hessian are evaluated at point $\mathbf{x}_0$. Consequently, Eq. 1 is modified into the following form

$$\mathbf{x}^{(i+1)} = \mathbf{x}^{(i)} - \left(H_{f(\mathbf{x}^{(i)})} \right)^{-1} \cdot \nabla f(\mathbf{x}^{(i)}), \tag{2}$$

In Eq. 2, both gradient and Hessian are evaluated at point $\mathbf{x}^{(i)}$.