ML Fundamentals: Normal Equation and Newton's Method

February 2, 2021

1 Normal Equation

Note: This can be proved by setting derivative of cost function, $\nabla_\theta J(\theta)$ as 0, and solving for θ.
But we can illustrate it with simple linear algebra equation with the assumption that global minima would have zero error.

$$\begin{aligned} X\theta &= \vec{y} \\ X^T X\theta &= X^T\vec{y} \\ \theta &= \left(X^T X\right)^{-1} X^T\vec{y} \end{aligned}$$

2 Newton's method

Newton's method is another method to maximize any function, i.e. log likelihood of parameters, $\ell(\theta)$

$$\theta := \theta - \Delta; f'(\theta) = \frac{f(\theta)}{\Delta}$$

$$\Delta := \frac{f(\theta)}{f'(\theta)}$$

$$\theta := \theta - \frac{f(\theta)}{f'(\theta)}$$

We want to maximise log likelihood function, $\ell(\theta)$, so lets take derivative of it, and find parameter where derivative is 0. $f(\theta) = \ell'(\theta)$

$$f(\theta) = \ell'(\theta)$$

Since, θ is vector valued, second derivative would be **Hessian matrix**.

$$\theta := \theta - H^{-1}\nabla_\theta \ell(\theta)$$

$$H_{ij} = \frac{\partial^2 \ell(\theta)}{\partial\theta_i \partial\theta_j}$$

Note, Newton's method is much faster when number of parameters are less than 100 or so. But when they are more than 100, it becomes computationally expensive to compute hessian matrix.