Batch gradient descent

$\theta_i = \theta \ominus \alpha \nabla_\theta \ell(\theta)$

Yaxin Li

Stochastic gradient descent (sgd)

$\theta_j := \theta_j \ \cancel{\oplus}\oplus \ \alpha \frac{\partial}{\partial \theta_j} \ell(\theta) \Rightarrow \theta_j = \theta_j + \alpha (y^{(i)} - h_\theta(x^{(i)})) x_j^{(i)}$

~~$\ell(\theta) \nabla \log f(y_n, x_n, \theta$~~

$\theta_n^{sgd} = \theta_{n-1}^{sgd} + a_n \nabla \ell(\theta_{n-1}^{sgd}; x_n, y_n)$ set $\theta_{n-1}^{sgd} = \theta_{n-1}^{im} = \theta$

$\rightarrow \theta_n^{im} = \theta_{n-1}^{im} + a_n \nabla \ell(\theta_n^{im}; x_n, y_n)$

$\ell(\theta) = \nabla \log f(y_n, x_n, \theta_n^{im}) = s_n \nabla \log f(y_n, x_n, \theta_{n-1}^{im})$

$s_n k_{n-1} = \ell'(x_n^T \theta_{n-1}^{im} + \gamma_n s_n k_{n-1} x_n^T C_n x_n; x_n, y_n)$

where $k_{n-1} = \ell'(x_n^T \theta_{n-1}^{im}; x_n, y_n)$

$\rightarrow \theta_n^{im} = \theta_{n-1}^{im} + \gamma_n C_n \nabla \log f(y_n; x_n, \theta_n^{im})$

$P(y | x; \theta) = h_\theta(x)^y + (1 - h_\theta(x))^{1-y}$

$y=1 \Rightarrow P(y=1 | x, \theta) = h_\theta(x)^1 (1 - h_\theta(x))$

$y=0 \Rightarrow P(y=0 | x; \theta) = 1 - h_\theta(x)$

$L(\theta) = P(y | x; \theta) = \prod_i P(y^{(i)} | x^{(i)}; \theta) = \prod_i h_\theta(x^{(i)})^{y^{(i)}} (1 - h_\theta(x^{(i)}))^{1-y^{(i)}}$

Note: $h_\theta(x) = g(\theta^T x) = g(\underbrace{x_1\theta_1 + \dots + x_j\theta_j + \dots + x_n\theta_n}_{z})$

$\frac{\partial}{\partial \theta_j} h_\theta(x) = g'(z) \frac{\partial z}{\partial \theta_j} \Rightarrow x_j$

$g(z) = \frac{1}{1+e^{-z}}$ $\quad g'(z) = \frac{e^{-z}}{(1+e^{-z})^2} = \frac{1}{(1+e^{-z})} \cdot \frac{e^{-z}}{(1-e^{-z})} = g(z)(1-g(z))$

since $1 - g(z) = 1 - \frac{1}{1+e^{z}} = \frac{e^{-z}}{1-e^{-z}}$

$$l(\theta) = \log L(\theta) = \sum_i [y^{(i)} \log h_\theta(x^{(i)}) + (1-y^{(i)}) \log(1-h_\theta(x^{(i)}))]$$

$$\cancel{\frac{\partial}{\partial \theta_j} l(\theta) = \sum_{i=1}^{n} [y^{(i)} - h_\theta(}$$

Maximum log- likelihood.

$$\frac{\partial}{\partial \theta_j} l(\theta) = \sum_i y^{(i)} \frac{1}{h_\theta(x^{(i)})} \frac{\partial}{\partial \theta_j} h_\theta(x^{(i)}) + (1-y^{(i)}) \frac{(-1)}{1-h_\theta(x^{(i)})} \frac{\partial}{\partial \theta_j} h_\theta(x^{(i)})$$