Gradient Descend

Friday, 16 March 2018 20:17

m: # training set

θ: parameter. $h_\theta(x)$: hypothesis function.

loss function: $J(\theta) = \frac{1}{2}\sum_{j=1}^{m}(h(x)-y)^2$

min: $J(\theta)$

$\theta_i = \theta_i - \alpha \frac{\partial}{\partial \theta_i} J(\theta)$.

$\frac{\partial}{\partial \theta_i} J(\theta) = \frac{1}{2}\cdot 2\,(h_\theta(x)-y)\cdot \frac{\partial}{\partial \theta_i}(h_\theta(x)-y)$

$= (h_\theta(x)-y)\cdot \frac{\partial}{\partial \theta_i}(\theta_0 x_0 + \cdots + \theta_n x_n - y)$

$= (h_\theta(x)-y)\cdot x_i$

update: $\theta_i := \theta_i - \underset{\text{learning rate}}{\alpha} \sum_{j=1}^{m}(h_\theta(x^{(j)}) - y^{(j)})\cdot x_i^{(j)}$

repeat until convergence. (Batch Gradient Descent)

Stochastic Gradient Descent

Repeat {
  For $j = 1$ to m {
    $\theta_i := \theta_i - \alpha(h_\theta(x^{(j)}) - y^{(j)})\cdot x_i^{(j)}$ (for all i)
  }
}

Linear Algebra:

$\nabla_\theta J = [\frac{\partial J}{\partial \theta_0}, \ldots \frac{\partial J}{\partial \theta_n}]^T \in R^{n+1}$

Thus: Gradient descent:

$\theta := \theta - \alpha \nabla_\theta J$ // $\theta, J \in R^{n+1}$

$f: R^{m\times n} \to R$ $f(A)$

$$\nabla_A f(A) = \begin{bmatrix} \frac{\partial f}{\partial A_{11}} & \cdots & \frac{\partial f}{\partial A_{1n}} \\ \vdots & \ddots & \vdots \\ \frac{\partial f}{\partial A_{m1}} & \cdots & \frac{\partial f}{\partial A_{mn}} \end{bmatrix}$$

Fact:

$tr(AB) = tr(BA)$ $tr(ABC) = tr(CAB) = tr(BCA)$

$\nabla_A tr AB = B^T$

$\nabla_A tr ABA^TC = CAB + C^TAB^T$

Design Matrix:

$$X\theta = \begin{bmatrix} (x^{(1)})^T \\ (x^{(2)})^T \\ \vdots \\ (x^{(m)})^T \end{bmatrix} \theta = \begin{bmatrix} x^{(1)T}\theta \\ x^{(2)T}\theta \\ \vdots \\ x^{(m)T}\theta \end{bmatrix} = \begin{bmatrix} h_\theta(x^{(1)}) \\ \vdots \\ h_\theta(x^{(m)}) \end{bmatrix}$$

$\vec{y} = [y^{(1)} \ldots y^{(m)}]^T$

$X\theta - y = [h(x^{(1)}) - y^{(1)}, \ldots h(x^{(m)}) - y^{(m)}]^T$

Recall: $z^Tz = \sum_i z_i^2$

$\frac{1}{2}(X\theta - y)^T(X\theta - y) = \frac{1}{2}\sum_{j=1}^{m}(h(x^{(j)}) - y)^2 = J(\theta)$

Recall: min $J(\theta)$

$\nabla_\theta \frac{1}{2}(X\theta - y)^T(X\theta - y)$

$= \frac{1}{2}\nabla_\theta(\theta^TX^TX\theta - \theta^TX^Ty - y^TX\theta + y^Ty)$

$= \frac{1}{2}\nabla_\theta\, tr(\theta^TX^TX\theta - \theta^TX^Ty - y^TX\theta + y^Ty)$

$= \frac{1}{2}[\nabla_\theta tr\,\theta\theta^TX^TX - \nabla_\theta tr\,y^TX\theta - \nabla_\theta tr\,y^TX\theta]$

$\nabla_\theta tr\,\underset{A}{\theta}\underset{B}{I}\underset{A^T}{\theta^T}\underset{C}{X^TX} = X^TX\theta + X^TX\theta$

$\nabla_\theta tr\,\underset{B}{y^TX}\underset{A}{\theta} = X^Ty$

$\nabla_\theta J(\theta) = \frac{1}{2}[X^TX\theta + X^TX\theta - X^Ty - X^Ty] = 0$

$X^TX\theta - X^Ty = 0$

$X^TX\theta = X^Ty$ Normal Equation.

$\theta = (X^TX)^{-1}X^Ty$