CO2774
Machine Learning
Aug 27, 2021

Logistics:-
Make-up class-
Aug 28, 2021
8am - 9am

Last class:-

$\theta^{(t)} \leftarrow$ initially ($t=0$)

Stochastic Gradient Descent

SGD:-

do {

$\theta^{(t+1)} \leftarrow \theta^{(t)} - \eta \cdot \nabla_\theta f_{B(t)}(\theta) \big|_{\theta^{(t)}}$

$t \leftarrow t+1$;

} (while ! converged)

↳ Batch of examples randomly sampled) r

$m = 1000$

$r = 10$

1000

$x^{(1)}, y^{(1)}$ - - - - - $x^{(m)}, y^{(m)}$

$\underbrace{\quad}_{r}$ $\underbrace{\quad}_{r}$ $\underbrace{\quad}_{r}$ $\underbrace{\quad}_{r}$

epoch → 1000

(m/r) $r = 10$

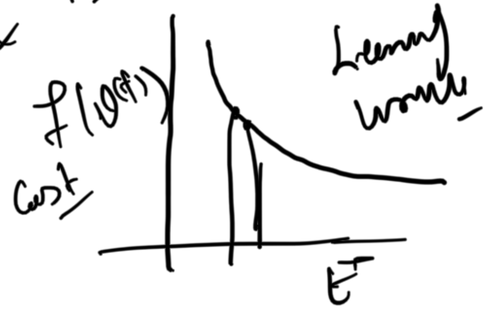
→ 100 updates in epoch

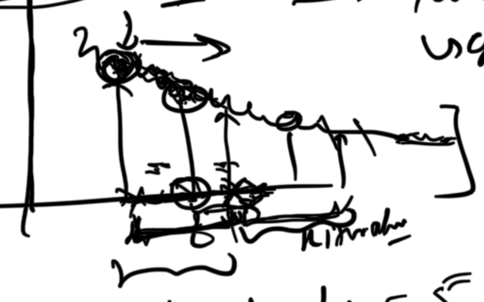
$\left[f(\theta^{(t+1)}) - f(\theta^{(t)}) \right] < \delta$

$| f_{B(t+2)}(\theta) - f_{B(t+1)}(\theta) | \le \delta$

$\Downarrow$

$J_{B(t+2)}(\theta)$ $f_{B(t)}(\theta)$





k iterations $= \delta$

Andrew Ng :-

"Validation"

$$J(\theta) = \frac{1}{2m} \sum_{i=2}^{m} \left[y^{(i)} - h_\theta(x^{(i)}) \right]^2$$

Analytical solution for Least Squares Regression

$$J(\theta) = \frac{1}{2m} \sum_{i=1}^{m} \left[y^{(i)} - h_\theta(x^{(i)}) \right]^2$$

→ $\arg\min_\theta J(\theta)$:- Gradient Descent

Analytical solution :- $h_\theta(x) = \theta^T x$

$$[f(\theta) = (\theta - 3)^2 + 1] \Rightarrow f'(\theta) = 2(\theta - 3)$$

$\theta = 3$

$f''(\theta) > 0$ global minima.

$\theta \in \mathbb{R}^{n+1} \rightarrow$ gradient

$\nabla_\theta J(\theta) = 0$ $\quad [\ \ ]$ $n+1$ dimensional vector. $\mathbb{R}^{(n+1)}$

Design:-

$X \rightarrow$ design matrix $(n+1)$

(Data Matrix) $\in \mathbb{R}^{m \times (n+1)}$

$\forall i\ x_0^{(i)} = 1$

$$X = \begin{bmatrix} — x^{(1)T} — \\ — x^{(i)T} — \\ \vdots \\ — x^{(m)T} — \end{bmatrix} \quad m \times n+1$$

ith row

$$Y = \begin{bmatrix} y^{(1)} \\ \vdots \\ y^{(m)} \end{bmatrix} \in \mathbb{R}^m \qquad \theta = \begin{bmatrix} \theta_0 \\ \vdots \\ \theta_n \end{bmatrix}$$

$y^{(i)}$ ith row

$J(\theta)$ as a function of X, Y, θ

$\mathbb{R}^{m \times (n+1)}$, $\mathbb{R}^{(n+1) \times 1}$

$$\frac{1}{2m}(X\theta - Y)^T (X\theta - Y) = J(\theta)$$

vector of "errors"

Difference between predictions & actual label

vector of model predictions.

matrix

Sum of squared errors

$$= \frac{1}{2m} \sum_{i=1}^{m} [y^{(i)} - \theta^T x^{(i)}]^2$$

$$\nabla_\theta J(\theta) = 0$$

$$\frac{1}{2m} \nabla_\theta \left[(X\theta - y)^T (X\theta - y) \right]$$

number

$$\frac{1}{2m} \nabla_\theta \left[(X\theta)^T (X\theta) - (X\theta)^T y - y^T X\theta + y^T y \right]$$

$$\frac{1}{2m} \nabla_\theta \left[\theta^T (X^T X) \theta - \theta^T X^T y - (y^T X\theta)^T + y^T y \right]$$

$$= \frac{1}{2m} \nabla_\theta \left[\theta^T (X^T X) \theta - \theta^T X^T y - \underbrace{(X\theta)^T (y^T)^T}_{\theta^T X^T y} + y^T y \right]$$

$$= \frac{1}{2m} \nabla_\theta \left[\theta^T X^T X \theta - 2\theta^T X^T y + y^T y \right]$$

$$\nabla_\theta \left[\theta^T X^T X \theta \right] = 2\left[X^T X \theta \right]$$

$$2 \nabla_\theta \theta^T X^T y = 2 X^T y$$

$$\nabla_\theta y^T y = 0$$

$$= \nabla_\theta \left[\boxed{\theta^T A \theta} \right] \rightarrow \text{matrix } (n+1) \times (n+1) \text{ (symmetric)}$$

$$= 2A\theta$$

Homework to prove

$\Rightarrow$ Result:

$\theta \in \mathbb{R}^{n+1}$

$$\nabla_\theta \sum_{j,k} \theta_j A_{jk} \theta_k$$

$$\frac{\partial}{\partial \theta_j} \sum_{j,k} \theta_j A_{jk} \theta_k$$

$$\nabla_\theta J(\theta) = 0$$

$$2X^T X\theta + 2X^T y = 0$$

$$\Rightarrow X^T X \theta = -X^T y$$

$$(X^T X)^{-1} X^T y$$

$\theta =$

Analytical soln for θ.