Outline

- Linear regression
- Batch / SGD
- Normal equation

Supervised Learning

$X \rightarrow Y$

picture → steering direction

Training Set

↓

Learning Algo

↓

New data

$\underline{x} \rightarrow \boxed{h} \rightarrow \underline{y}$

"hypothesis"

x. input, y, output

(x, y). training example

$(x^{(i)}, y^{(i)})$ i-th training example

$x_1^{(i)}$ 1-st feature from i-th example

d = # of features

$x^{(i)}, \theta$ $(d+1)$ dimensional

Regression (output continuous)

v.s.

Classification output discrete

How to represent h?

$h(x) = \theta_0 + \theta_1 x$ (technically affine func.)

$= \sum_{j=0}^{d} \theta_j x_j$ (where $x_0 = 1$)

Ex:

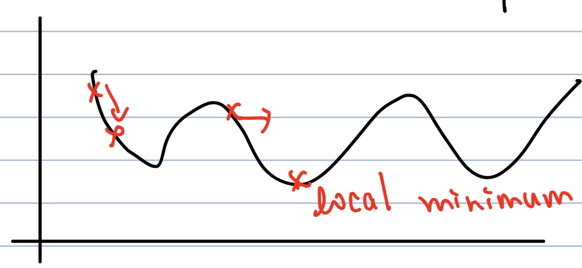
$\theta = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \vdots \\ \theta_d \end{bmatrix}$ $x = \begin{bmatrix} x_0 \\ x_1 \\ \vdots \\ x_d \end{bmatrix}$

Objective Function

cost fn. $J(\theta) = \frac{1}{2} \sum_{i=1}^{n} (h_\theta(x^{(i)}) - y^{(i)})^2$

$\Rightarrow \min_\theta J(\theta)$

Gradient Descent update θ.



(Batch) Gradient Descend

$\theta_j := \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta)$ $(j = 0, 1, \ldots d)$

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

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

$$\theta_j := \theta_j - \underbrace{\sum_{i=1}^{n} \alpha \left(h_\theta(x^{(i)}) - y^{(i)}\right) \cdot x_j^{(i)}}_{\to \frac{\partial}{\partial \theta_j} J(\theta)}$$

Stochastic Gradient Descend

Repeat {
    For $i = 1$ to n {
        For $j = 0$ to d {
            $\theta_j := \theta_j - \alpha (h_\theta(x^i) - y^i)\, x_j^i$
        }
    }
}

Note: SGD could sometimes prevent being trapped in local minimum due to randomness.

Normal Equation

$$\nabla_\theta J(\theta) = \begin{bmatrix} \frac{\partial J}{\partial \theta_0} \\ \frac{\partial J}{\partial \theta_1} \\ \vdots \end{bmatrix}$$

$A \in \mathbb{R}^{2\times 2}$ $\quad A : \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}$ $\quad f(A).\ f: \mathbb{R}^{2\times 2} \to \mathbb{R}$

$$\nabla_A f(A) = \begin{bmatrix} \frac{\partial f}{\partial A_{11}} & \frac{\partial f}{\partial A_{12}} \\ \frac{\partial f}{\partial A_{21}} & \frac{\partial f}{\partial A_{22}} \end{bmatrix}$$

$$\nabla_\theta J(\theta) \overset{set}{=} \vec{0}$$

$$J(\theta) = \frac{1}{2} \sum_{i=1}^{n} \left(h(x^{(i)}) - y^{(i)}\right)^2$$

$$X = \begin{bmatrix} - & (x^{(1)})^T & - \\ & \vdots & \\ - & (x^{(n)})^T & - \end{bmatrix}$$ design matrix

$$\vec{y} = \begin{bmatrix} y^{(1)} \\ \vdots \\ y^{(n)} \end{bmatrix}$$

$$X\theta = \begin{bmatrix} \quad \end{bmatrix} \begin{bmatrix} \theta_0 \\ \theta_1 \\ \vdots \\ \theta_d \end{bmatrix} = \begin{bmatrix} h_\theta(x^{(1)}) \\ \vdots \\ h_\theta(x^{(n)}) \end{bmatrix}$$

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

$$\nabla_\theta J(\theta) = X^T X\theta - X^T y = \vec{0}$$

$$X^T X\theta = X^T y$$ "Normal equation"

Optimal value $\theta = (X^T X)^{-1} X^T y$