Outline

- Linear regression
- Batch/ Stochastic gradient descent
- Normal equation

Supervised Learning

X ⟶ y

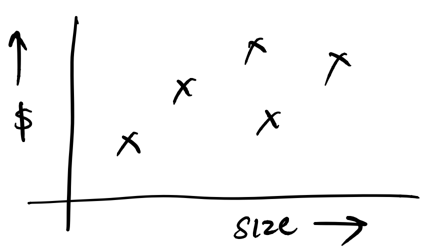
picture — steering direction

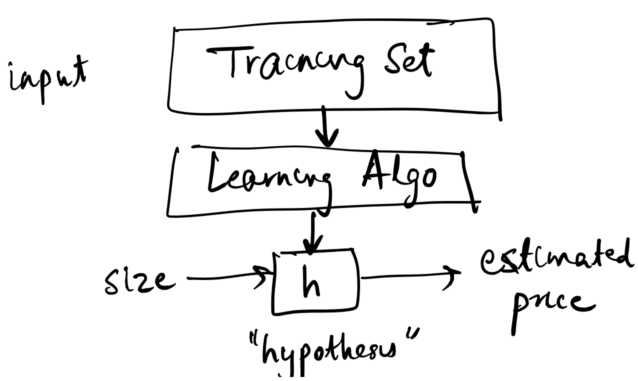
$\boxed{\text{Regression}}$ (o/p continuous)

v/s classification

Housing dataset

Size	Price ($ 1,000s)
2104	400
1416	232
1534	315





How to represent h?

$h(x) = \theta_0 + \theta_1 x$ (technically affine f^n)

More features

		Size	# bedrooms	Price
$x^{(1)}$	1	2104	4	400
$x^{(2)}$	1	1416	3	232

$x_1^{(1)} = 2104$

$x_1^{(2)} = 1416$

$x_1 = \text{size}, \quad x_2 = \#\text{bedrooms}$

$$h(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2$$

$$h(x) = \sum_{j=0}^{2} \theta_j x_j$$

Define $x_0 = 1$

$$\theta = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \end{bmatrix} \qquad x = \begin{bmatrix} x_0 \\ x_1 \\ x_2 \end{bmatrix} \begin{matrix} \text{always } 1 \\ \text{size} \\ \#\text{bedrooms} \end{matrix}$$

parameters.

n = # training examples

x = "inputs" / features.

y = "output" / target variable.

(x, y) = training example

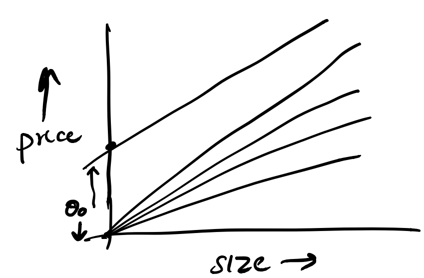
$(x^{(i)}, y^{(i)})$: i^{th} training example

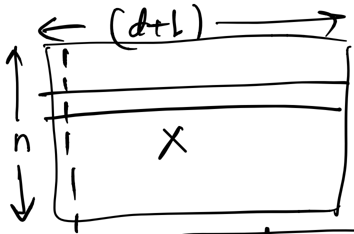
$x_1^{(i)}$: i runs from 1 to n

d = # features

$(d = 2)$

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





$$\boxed{h(x) = \sum_{j=0}^{d} \theta_j x_j}$$

Choose θ s.t. $h(x) \approx y$

$$h_\theta(x) = h(x)$$

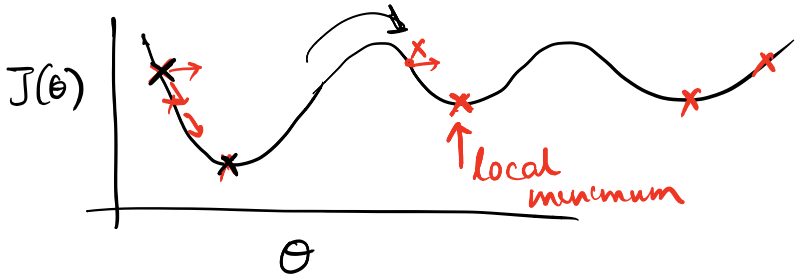
Cost f^n $\quad J(\theta) = \frac{1}{2} \sum_{i=1}^{n} \left(h_\theta(x^{(i)}) - y^{(i)} \right)^2$

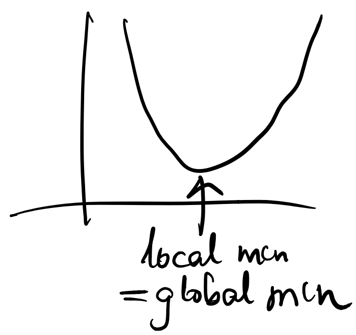
$$\min_\theta J(\theta)$$

Gradient Descent

Start with θ (say $\theta = \vec{0}$)

Keep changing θ to reduce $J(\theta)$





Gradient Descent

Start with θ

Repeat until convergence

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

learning rate

$a := a + 1$ ✓

$a = a + 1$ ✗

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

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

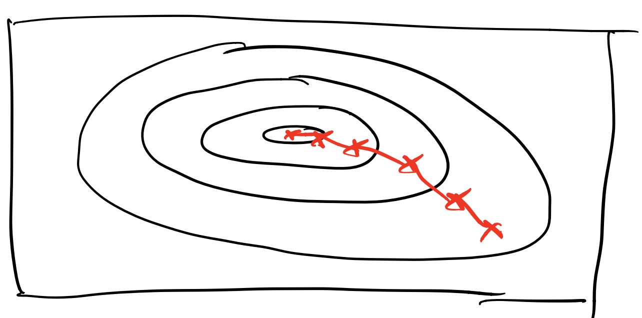
(chain rule)

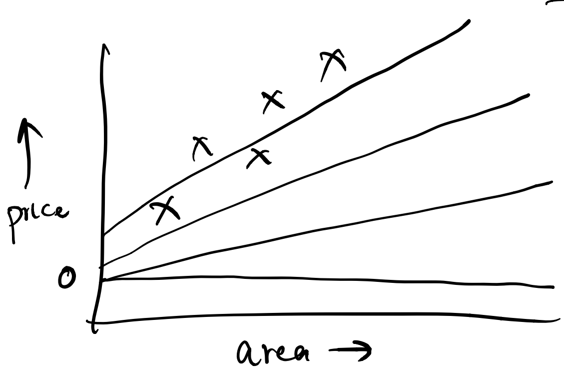
$$= \left(h_\theta(x) - y\right) \frac{\partial}{\partial\theta_j}\left(\theta_0 x_0 + \theta_1 x_1 \cdots + \theta_d x_d - y\right)$$

$$= \left(h_\theta(x) - y\right) x_j$$

$$\theta_j := \theta_j - \alpha\left(h_\theta(x) - y\right) x_j$$

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

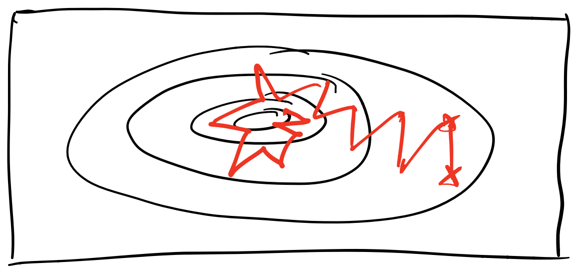




"Batch" gradient Descent

Stochastic gradient Descent

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



mini-batch

Normal Equation

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

$A \in \mathbb{R}^{2 \times 2}$ $\quad A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}$

$f(A)$

$f : \mathbb{R}^{2 \times 2} \longrightarrow \mathbb{R}$

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

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

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

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

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

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

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

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

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

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