Supervised Learning

+ Definitions
+ Linear Regression
+ Batch & stochastic gradient descent
+ Normal Equations

Supervised learning

Given: Training Set

$\{(x^{(1)}, y^{(1)}) \cdots (x^{(n)}, y^{(n)})\}$ $x^{(i)} \in X$, $y^{(i)} \in Y$

Prediction

$h: X \longrightarrow Y$

Do: find "good" $h: X \to Y$ hypothesis

Image — Contains cat

TEXT — IS HATE SPEECH?

House Data — PRICE

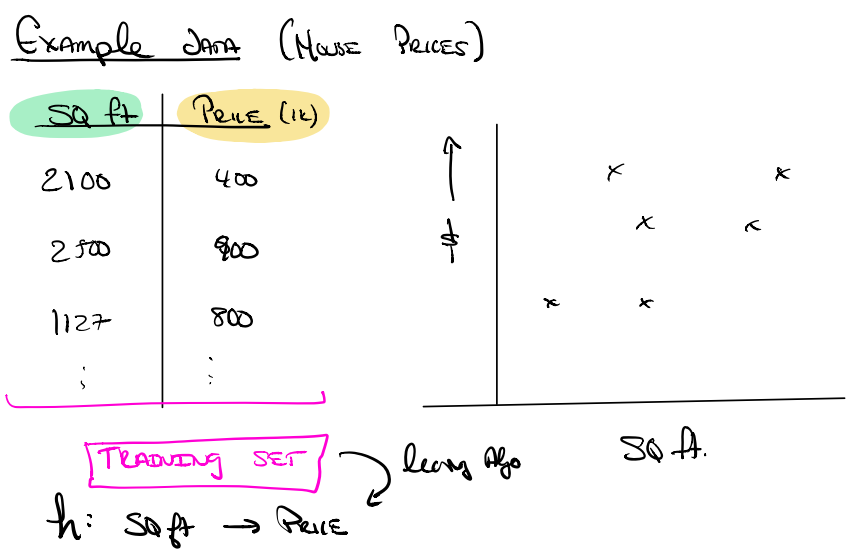
this job of training algorithm

WE USE h ON NEW DATA (x)

CALL this PREDICTION, WE ARE VERY INTERESTED IN x & TRAINING SET

if y IS DISCRETE $\Rightarrow$ classification

y IS continuous $\Rightarrow$ Regression



How do we represent h?

$h(x) = \theta_0 + \theta_1 x_1$ (affine fn.)

	x_1 Size	x_2 Bedroom	x_3 lot size	Price
$x^{(1)}$	2104	4 ($x_2^{(1)}$)	45k	400
$x^{(2)}$	2500	3	30k	900

$$h(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots$$
$$= \sum_{j=0}^{3} \theta_j x_j$$

NB x_0 identically 1

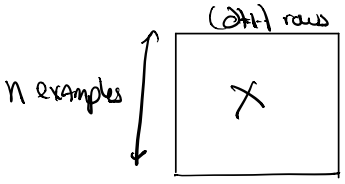
$$\theta = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \\ \theta_3 \end{bmatrix} \quad x^{(i)} = \begin{bmatrix} x_0^{(i)} \\ x_1^{(i)} \\ x_2^{(i)} \\ x_3^{(i)} \end{bmatrix} \begin{matrix} 1 \\ \text{SIZE} \\ \text{BEDROOMS} \\ \text{lot SIZE} \end{matrix} \quad y^{(i)} \text{ is Price}$$

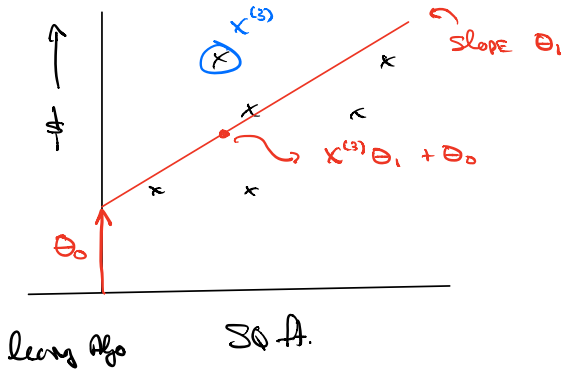
PARAMETERS INPUTS / FEATURES OUTPUT / TARGET

(x, y) is a training example

$(x^{(i)}, y^{(i)})$ is i^{th} example i runs $1 \ldots n$

n examples and d features $\Rightarrow$ $x^{(i)}$'s are $d+1$ dimensional





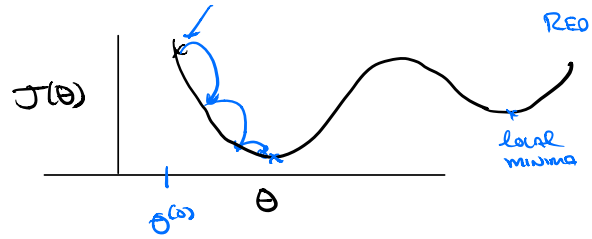
$h_\theta(x) = \sum_{j=0}^{d} \theta_j x_j$ WANT TO CHOOSE θ s.t. $h_\theta(x) \approx y$

IDEA: $J(\theta) = \frac{1}{2} \sum_{i=1}^{n} \left(h_\theta(x^{(i)}) - y^{(i)}\right)^2$ Cost function (least squares)

$\min_\theta J(\theta)$

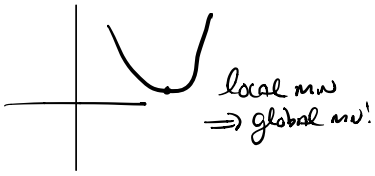
Gradient Descent

START $\theta^{(0)}$ AT RANDOM OR ZERO



REDUCE USING GRADIENT

If J is nice (convex)



$\theta^{(0)} := 0$

$\theta_j^{(t+1)} := \theta_j^{(t)} - \alpha \frac{\partial}{\partial \theta_j} J(\theta^{(t)})$ — α: learning rate

$j = 1 \dots d$

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

$$= \sum_{i=1}^{n} \left(h_\theta(x^{(i)}) - y^{(i)} \right) \frac{\partial}{\partial \theta_j} h_\theta(x^{(i)})$$

$$h_\theta(x^{(i)}) = \theta_0 x_0^{(i)} + \theta_1 x_1^{(i)} + \dots + \theta_d x_d^{(i)}$$

$$\frac{\partial h_\theta(x^{(i)})}{\partial \theta_j} = x_j^{(i)}$$

$$\theta_j^{(t+1)} := \theta_j^{(t)} - \alpha \sum_{i=1}^{n} \left(h_\theta(x^{(i)}) - y^{(i)} \right) x_j^{(i)}$$

SOMETIMES WRITE AS $\theta^{(t+1)} := \theta^{(t)} - \alpha \sum_{i=1}^{n} \left(h_\theta(x^{(i)}) - y^{(i)} \right) x^{(i)}$

vector notation

BATCH VERSUS STOCHASTIC MINIBATCH

$$\theta^{(t+1)} := \theta^{(t)} - \alpha \sum_{i=1}^{n} \left(h_\theta(x^{(i)}) - y^{(i)}\right) x^{(i)}$$

MINIBATCH: Randomly select $b < n$ points AND Estimate gradient

1. Pick b points $\{i_1 \dots i_b\} = B$
2. $$\theta^{(t)} := \theta^{(t)} - \alpha_b \sum_{k \in B} \left(h_\theta(x^{(k)}) - y^{(k)}\right) x^{(k)}$$

ONE DETAIL Scale α AND α_b differently.

TRADEOFF: NOISIER BUT MUCH FASTER

FASTER: IMAGINE IT TRAINING SET CONTAINS 100 COPIES of SAME POINT

$\Rightarrow$ NOT AS REDUCTIONS AS IT SEEMS (LINEAR CODES)

How do you choose B? Sadly, WHATEVER WORKS

Normal Equation

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

$A \in \mathbb{R}^{2\times 2}$

$$A = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \qquad f: A \to \mathbb{R}$$

then

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

Now, we want to find minimum

$$\nabla_\theta J(\theta) = \vec{0} \qquad (\nabla_\theta J(\theta) \in \mathbb{R}^{d+1})$$

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

$$X = \begin{bmatrix} - x^{(1)} - \\ x^{(2)} \\ \vdots \\ x^{(n)} \end{bmatrix} \in \mathbb{R}^{n \times d}$$ Design matrix

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

$$y = \begin{bmatrix} y^{(1)} \\ \vdots \\ y^{(n)} \end{bmatrix} \qquad \text{then } J(\theta) = \frac{1}{2}(X\theta - y)^T (X\theta - y)$$

$$\nabla_\theta J(\theta) = X^T X\theta - X^T y = 0 \Rightarrow \theta = (X^T X)^{-1} X^T y$$

Optimal value.