Supervised Learning

+ Definitions
+ Linear Regression
+ Batch & Stochastic Gradient
+ Normal Equations

Supervised Learning

Prediction

$h : \mathcal{X} \to \mathcal{Y}$

Images   CAT

TEXT   IS HATE SPEECH?

House Data   Price

Given Training Set

$\{(x^{(1)}, y^{(1)}), \ldots (x^{(n)}, y^{(n)})\}$

$x^{(i)} \in \mathcal{X}, \quad y^{(i)} \in \mathcal{Y}$

Do: find good $h : \mathcal{X} \to \mathcal{Y}$ (hypothesis)

WE CARE ABOUT NEW x VALUES NOT IN training SET.

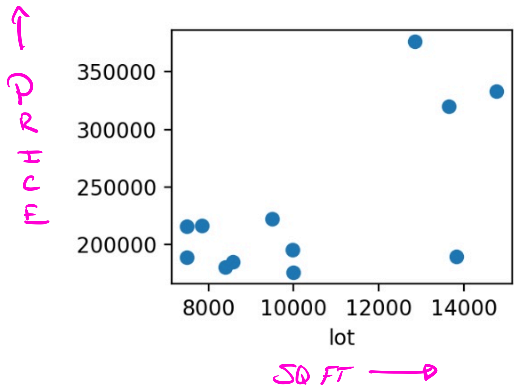
if y IS DISCRETE $\Rightarrow$ Classification

y IS Continuous $\Rightarrow$ Regression

Example Data   House prices   (Ames dataset, Kaggle dataset)

h : Lot. Area ⟶ Price

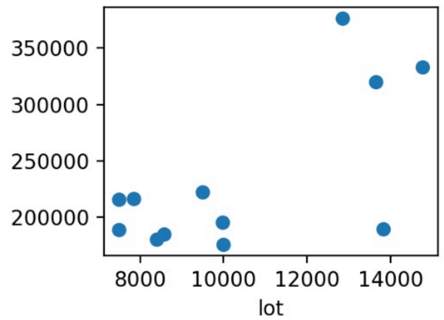
	SalePrice	Lot.Area
4	189900	13830
5	195500	9978
9	189000	7500
10	175900	10000
12	180400	8402
22	216000	7500
36	376162	12858
47	320000	13650
55	216500	7851
56	185088	8577
58	222500	9505
59	333168	14774

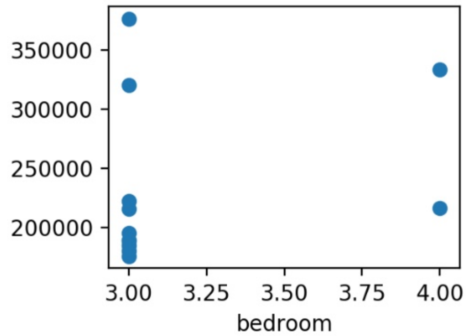


Slightly Richer Example...

	Lot.Area	SalePrice	Bedroom.AbvGr
4	13830	189900	3
5	9978	195500	3
9	7500	189000	3
10	10000	175900	3
12	8402	180400	3
22	7500	216000	3
36	12858	376162	3
47	13650	320000	3
55	7851	216500	4
56	8577	185088	3
58	9505	222500	3
59	14774	333168	4

Price





How do we represent h?

$x_0^{(i)} = 1$

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

	Size	Bedroom	lot size	...	Price (y)
$\rightarrow x^{(1)}$	2104 $x_1^{(1)}$	4 $x_2^{(1)}$	45k		400
$\rightarrow x^{(2)}$	2500	3	36k		900

$$h(x) = \theta_0 x_0 + \theta_1 x_1 + \cdots + \theta_d x_d$$

$$= \sum_{j=0}^{d} \theta_j \cdot x_j \qquad \text{NB } x_0 = 1$$

$$\theta = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \\ \vdots \\ \theta_d \end{bmatrix} \qquad x^{(i)} = \begin{bmatrix} x_0^{(i)} \rightarrow 1 \\ x_1^{(i)} \rightarrow 2104 \\ \vdots \end{bmatrix}$$

Parameters   Features

Size

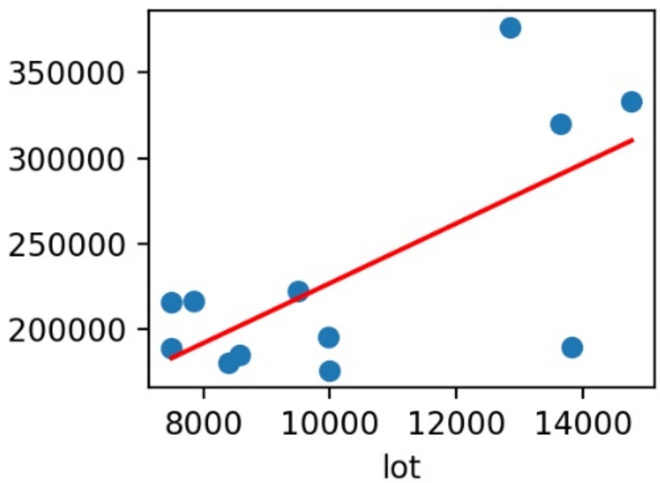
Bedroom

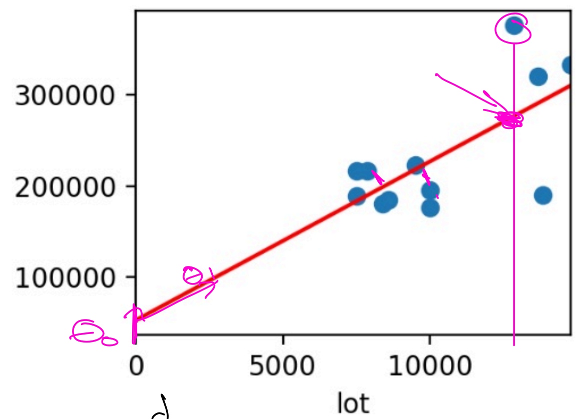
$y^{(i)}$ is Price

$(x^{(i)}, y^{(i)}) \leftarrow$ training example

$$X = \begin{bmatrix} — & x^{(1)} & — \\ — & x^{(2)} & — \\ & \vdots & \\ — & x^{(n)} & — \end{bmatrix} \in \mathbb{R}^{n \times (d+1)}$$

Simplest fit





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

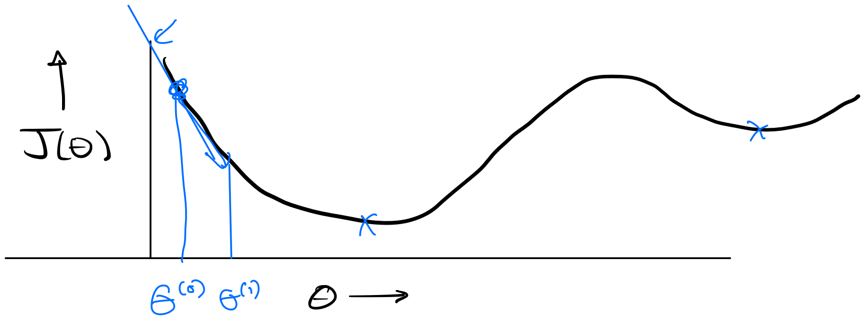
$h_\theta(x) = \sum_{j=0}^{d} \theta_j x_j$ $\qquad h_\theta(x) \approx y$

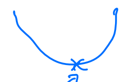
IDEA $J(\theta) = \frac{1}{2} \sum_{i=1}^{n} \left(h_\theta(x^{(i)}) - y^{(i)}\right)^2$ Cost function (Least Squares)

$\longrightarrow \min_{\theta} J(\theta)$

"Optimization"

Gradient Descent





local
$\Rightarrow$ global
for convex

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

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

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

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

Misprediction Error

$$h_\theta(x) = \theta_0 x_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_d x_d$$

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

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

$\Rightarrow$ Vector Equation

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

BATCH VS. STOCHASTIC MINIBATCH

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

MINIBATCH $\quad B$ (AT RANDOM) $\quad B \subseteq [N] \quad B << N$

$$\Theta^{(t+1)} := \Theta^{(t)} - \alpha_B \sum_{i \in B} \left(h_\theta (x^{(i)}) - y^{(i)} \right) x^{(i)}$$

Normal Equation