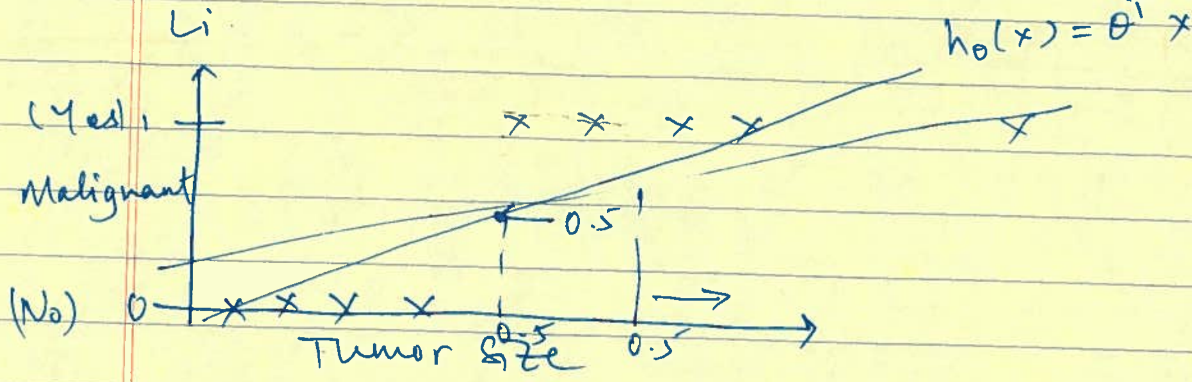


Threshold classifier output $h_\theta(x)$ at 0.5

If $h_\theta(x) \geq 0.5$, predict "$y=1$"

If $h_\theta(x) < 0.5$, predict "$y=0$"

Logistic Regression

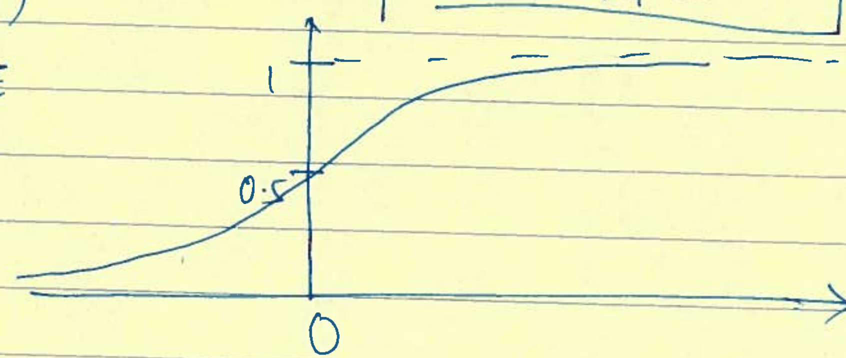
Hypothesis Representation

want $0 \leq h_\theta(x) \leq 1$

$$\boxed{h_\theta(x) = \frac{1}{1+e^{-\theta^T x}}}$$

$h_\theta(x) = g(\theta^T x)$

sigmoid or Logistic function $g(z) = \frac{1}{1+e^{-z}}$



Parameters $\underline{\theta}$

$h_\theta(x)$ = estimated probability that $y=1$ on input x

if $x = \begin{bmatrix} x_0 \\ x_1 \end{bmatrix} = \begin{bmatrix} 1 \\ \text{tumorsize} \end{bmatrix}$

$h_\theta(x) = 0.7$ 70% chance.

$h_\theta(x) = P(y=1 \mid x;\theta)$ $y=1$, give x parameterized by θ

$$P(y=0 \mid x;\theta) + P(y=1 \mid x;\theta) = 1$$

$$P(y=0 \mid x;\theta) = 1 - P(y=1 \mid x;\theta)$$

$h_\theta(x) = g(\theta^T x) = P(y=1|x;\theta)$

$g(z) = \dfrac{1}{1+e^{-z}}$

$y=1$ if $h_\theta(x) \geq 0.5 \longrightarrow$ $g(z) \geq 0.5$ when $z \geq 0$

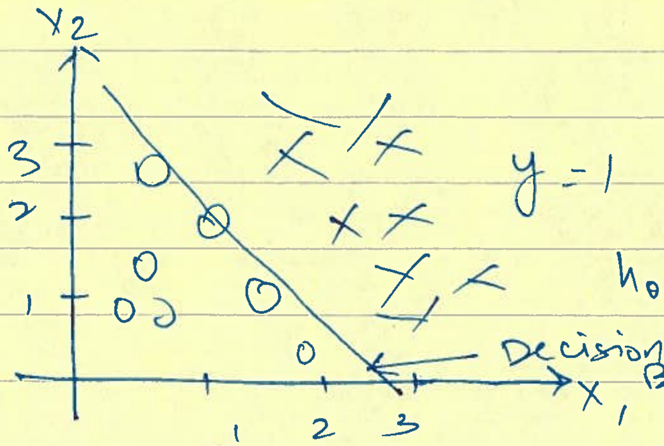
$h_\theta(x) = g(\theta^T x) \geq 0.5$ whereever $\theta^T x \geq 0$

$y=0$ if $h_\theta(x) < 0.5$

$g(z) < 0.5$

$h_\theta(x) = g(\theta^T x)$

$\theta^T x < 0$



$h_\theta(x) = g(\theta_0 + \theta_1 x_1 + \theta_2 x_2)$

$\theta_0 \to 3$, $\theta_1 \to 1$, $\theta_2 \to 1$

"$y=1$" if $\underbrace{-3 + x_1 + x_2 \geq 0}_{\theta^T x}$ $\quad \theta = \begin{bmatrix} -3 \\ 1 \\ 1 \end{bmatrix}$

$\longrightarrow x_1 + x_2 \geq 3 \quad x_1 + x_2 < 3$

$y = 1 \qquad y = 0$

$\boxed{x_1 + x_2 = 0} \longrightarrow h_\theta(x) = 0.5.$

$\theta = \begin{bmatrix} 5 \\ -1 \\ 0 \end{bmatrix}$ $\quad h_\theta(x) = g(5 - x_1)$

$5 - x_1 \geq 0$

$x_1 \geq 5$

$y=1$ if $5 + (-1)x_1 + 0(x_2) \geq 0$

$5 - x_1 \geq 0$

$-x_1 \geq -5 \quad x \leq 5$

$$h_\theta(x) = g(\theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_1^2 + \theta_4 x_3^2)$$

$$\theta = \begin{bmatrix} -1 \\ 0 \\ 0 \\ 1 \\ 1 \end{bmatrix}$$

$y=1$ if $-1 + x_1^2 + x_2^2 \geq 0$

$x_1^2 + x_2^2 \geq 1$ Decision Boundary.

Unit circle.

Inside circle $y=0$

Outside circle $y=1$

Decisio Boundary → is a property of hypothesis and parameters not the training set.

Cost function

Training set $\{(x^{(1)}, y^{(2)}), (x^{(2)}, y^{(2)}), \ldots, (x^{(m)}, y^{(m)})\}$

m examples $x \in \begin{bmatrix} x_0 \\ x_1 \\ \vdots \\ x_n \end{bmatrix} \mathbb{R}^{n+1}$ $\quad x_0 = 1, y \in \{0, 1\}$

$$h_\theta(x) = \frac{1}{1 + e^{-\theta^T x}}$$

~~Linear~~ Regression

↓

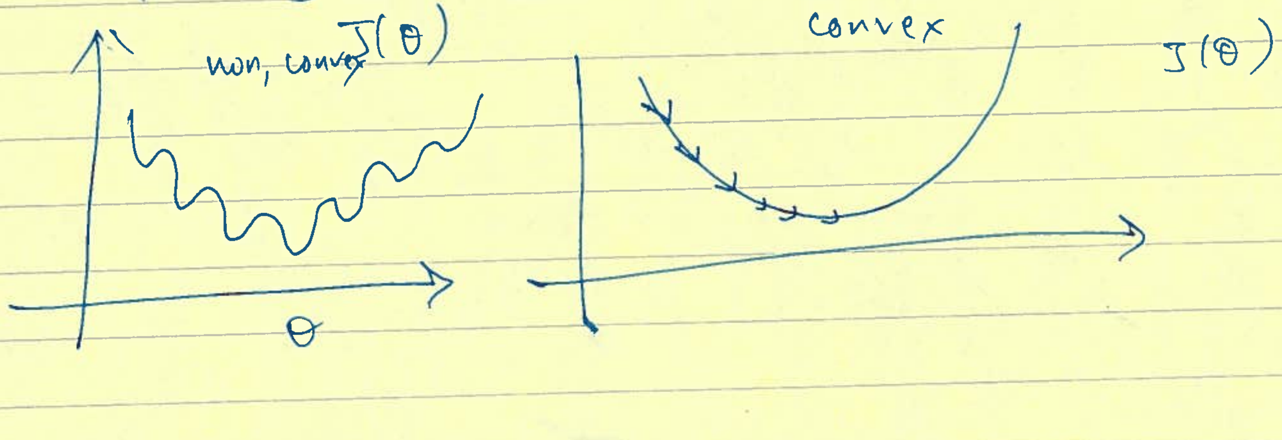
logistic.

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

$$\text{Cost}(h_\theta(x^{(i)}, y)$$

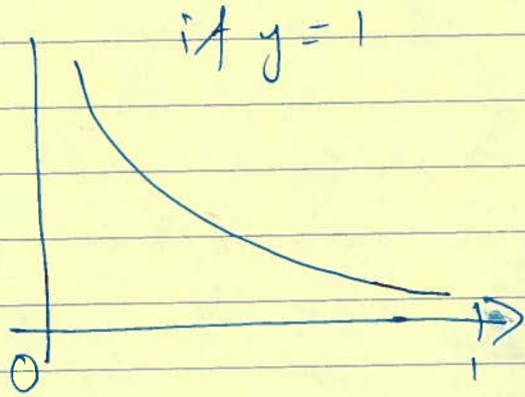
$$\text{Cost}(h_\theta(x^{(i)}), y^{(i)}) = \frac{1}{2} h_\theta(x^{(i)}) - y^{(i)})^2$$

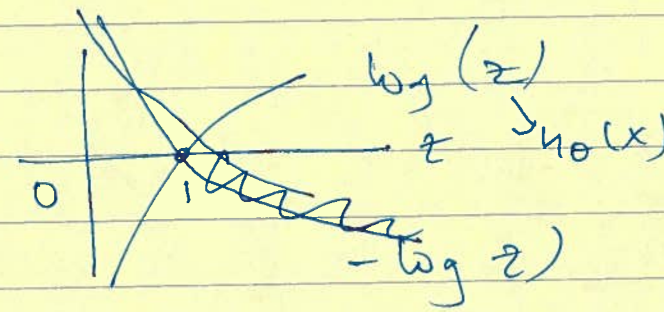
$$\text{Cost}(h_\theta(x, y)) = \frac{1}{2} (h_\theta(x, y))^2$$



Logistic Regression Cost Function

$$\text{Cost}(h_\theta(x), y) = \begin{cases} -\log(h_\theta(x)) & \text{if } y = 1 \\ -\log(1 - h_\theta(x)) & \text{if } y = 0 \end{cases}$$





Cost = 0 if $y = 1$, $h_\theta(x) = 1$

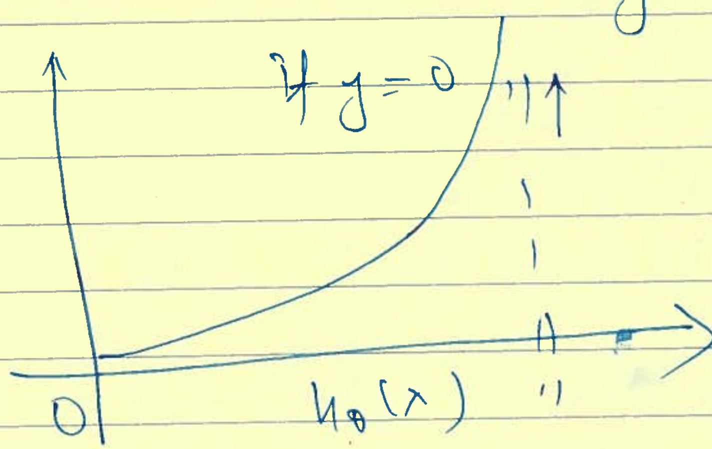
But as $h_\theta(x) \rightarrow 0$

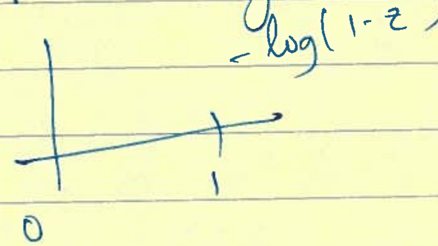
Cost $\rightarrow \infty$

$h_\theta(x) = 0$

Predicts

$P(y = 1 | x; \theta)$ by $= 1$





Simplified cost function and gradient descent

$$J(\theta) = \frac{1}{m} \sum_{i=1}^{m} \text{Cost}\left(h_\theta(x^{(i)}), y^{(i)}\right)$$

$$\text{Cost}(h_\theta(x), y) = \begin{cases} -\log(h_\theta(x)) & \text{if } y = 1 \\ -\log(1 - h_\theta(x)) & \text{if } y = 0 \end{cases}$$

Note : $y = 0$ or 1 always.

$$\text{Cost}(h_\theta(x), y) = -y \log(h_\theta(x)) - (1-y) \log(1 - h_\theta(x))$$

if $y = 1$: $\text{Cost}(\;) = -\log h_\theta(x)$

if $y = 0$: $\text{Cost}(\;) = -\log(1 - h_\theta(x))$

$$J(\theta) = \frac{1}{m} \sum_{i=1}^{m} \text{cost}(h_\theta(x^{(i)}, y^{(i)})$$

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

To fit parameters θ:

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

Output $h_\theta(x) = \frac{1}{1+e^{-\theta^T x}}$ $\quad P(y=1|x;\theta)$

Want $\min_\theta J(\theta)$:

Repeat {

$$\theta_j := \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta)$$

for i=0 to n.

$$\theta = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \vdots \\ \theta_n \end{bmatrix}$$

$$\frac{\partial}{\partial \theta_j} J(\theta) = \frac{1}{m} \sum_{i=1}^{m} h_\theta(x^{(i)} - y^{(i)}) x_j^{(i)}$$

$$h_\theta(x) = \theta^T x$$

$$h_\theta(x) = \frac{1}{1+e^{-\theta^T x}}$$

Vectorized implementation

$$\theta : \theta - \alpha \frac{1}{m} \Sigma_{i=1}^{m} [(h_\theta(x^{(i)}) - y^{(i)}) \cdot x^{(i)}]$$

vectorized

$$\theta := \theta - \frac{\alpha}{m} X^T (g(X\theta - \vec{y})$$

Optimization Algorithm.

$J(\theta)$

$$\frac{\partial}{\partial \theta_j} J(\theta) \quad (\text{for } j = 0, 1, \ldots, n)$$

Gradient descent

{ Repeat

$$\theta_j : \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta)$$

}

$$\theta = \begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix}$$

$$J(\theta) = (\theta_1 - 5)^2 + (\theta_2 - 5)^2$$

$\theta_1 = 5, \theta_2 = 5$

$$\frac{\partial J(\theta)}{\partial \theta_1} = 2(\theta_1 - 5)$$

$$\frac{\partial J(\theta)}{\partial \theta_2} = 2(\theta_2 - 5)$$

2

$$\begin{bmatrix} -6 \\ 0 \\ 1 \end{bmatrix}$$

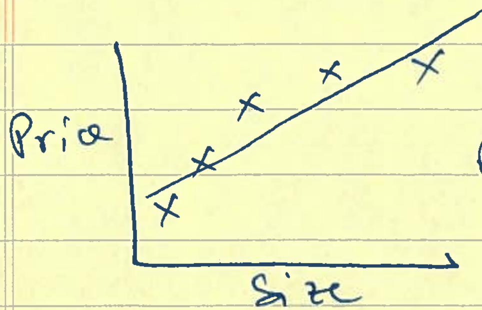
$-6 + x_2 \geq 0$

~~$x_2 \leq 6$~~

$x_2 \geq 6.$

Binary Classification

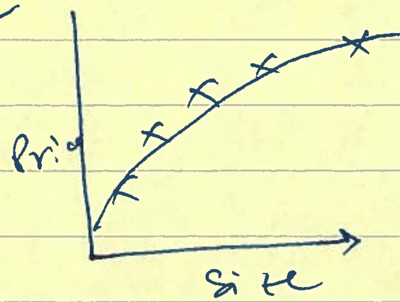
The Problem of Overfitting



$\theta_0 + \theta_1 x$

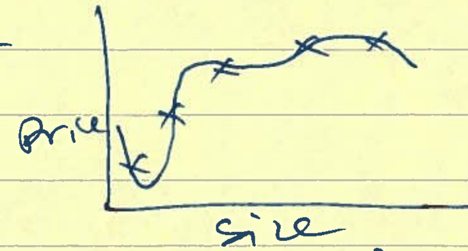
↓ ↓

"Underfit" "High bias"



$\theta_0 + \theta_1 x + \theta_2 x^2$

"just Right"

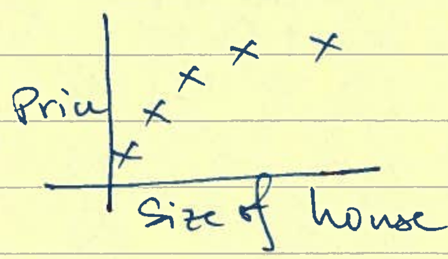


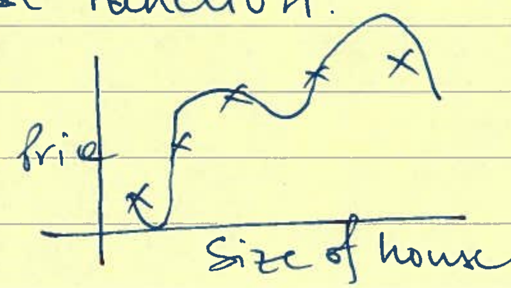
$\theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \theta_4 x^4$

"Overfit"

"High" Variance"

Regularization Cost Function:





θ_3, θ_4 really small

$$\min_\theta \frac{1}{2m} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)} \right)^2 + 1000\,\theta_3^2 + 1000\,\theta_4^2$$

$\theta_3 \approx 0 \quad \theta_4 \approx 0$

$\dfrac{\theta_3, \theta_4}{\approx 0}$

Housing

→ Features $x_1, x_2, \ldots, x_{100}$

Parameter $\theta_0, \theta_1, \theta_2, \ldots, \theta_{100}$

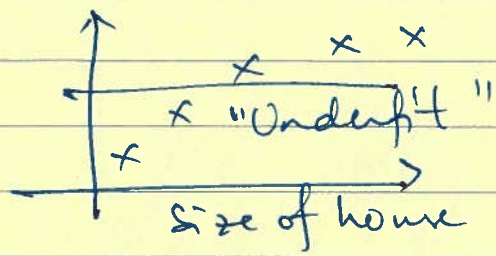
$$J(\theta) = \frac{1}{2m} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)} \right)^2 + \lambda \sum_{i=1}^{1} \theta_j^2$$

$\theta_1, \theta_2, \theta_3, \ldots \theta_{100}$.

Regularization Parameter

fit Training set well)

$\min_\theta J(\theta)$



$\lambda = 10^{10}$ $\theta_1 \approx 0, \theta_2 \approx 0, \theta_3 \approx 0$

$h_\theta(x) = \theta_0$

$$h_\theta(x) = \theta_0 + \cancel{\theta_1 x} + \cancel{\theta_2 x^2} + \cancel{\theta_3 x^3} + \cancel{\theta_4 x^4}$$

Regularized Linear Regression.

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

Repeat {

$$\theta_0 := \theta_0 - \alpha \overbrace{\frac{1}{m} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)} \right) x_0^{(i)}}^{\partial/\partial\theta_0 J(\theta)}$$

$$\theta_j := \theta_j - \alpha \left[\frac{1}{m} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)} \right) x_j^{(i)} + \frac{\lambda}{m} \theta_j \right]$$

$(j = 0, 1, 2, 3, \ldots, n)$

$\frac{\partial}{\partial \theta_j} J(\theta)$ ↳ regularization

$$\theta_j = \theta_j \left(1 - \alpha \frac{\lambda}{m}\right) - \alpha \frac{1}{m} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)} \right) x_j^{(i)}$$

$$\frac{\partial}{\partial \theta_j} J(\theta) \overset{\text{set}}{=} 0$$

$$X = \begin{bmatrix} x^{(1)T} \\ \vdots \\ (x^{(m)})^T \end{bmatrix}_{m \times (n+1)} \qquad y = \begin{bmatrix} y^{(1)} \\ \vdots \\ y^{(m)} \end{bmatrix} \mathbb{R}^m$$

$$\min_\theta J(\theta) = \Big(X^T X + \lambda \underbrace{\begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 1 & & 0 \\ 0 & & \ddots & \\ & & & 1 \end{bmatrix}}_{(n+1) \times (n+1)} \Big)^{-1} X^T y$$

$$\text{Eg } n = 2 \qquad \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

Suppose $\underset{\#\text{examples}}{m} \le n \rightarrow \#$ features

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

↳ non invertible / singular

if $\lambda > 0$

$$\theta = \left(X^T X + \lambda \begin{bmatrix} 0 & & & \\ & 1 & & \\ & & 1 & \\ & & & \ddots & 1 \end{bmatrix} \right)^{-1} X^T y.$$