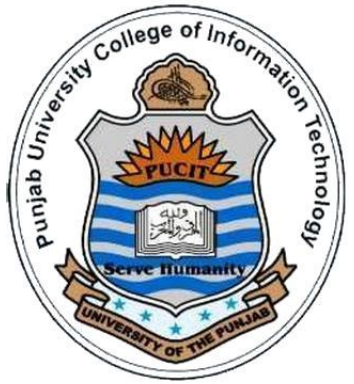


Machine Learning

Muhammad Adeel Nisar

Assistant Professor – Department of IT,
Faculty of Computing and Information Technology,
University of the Punjab, Lahore

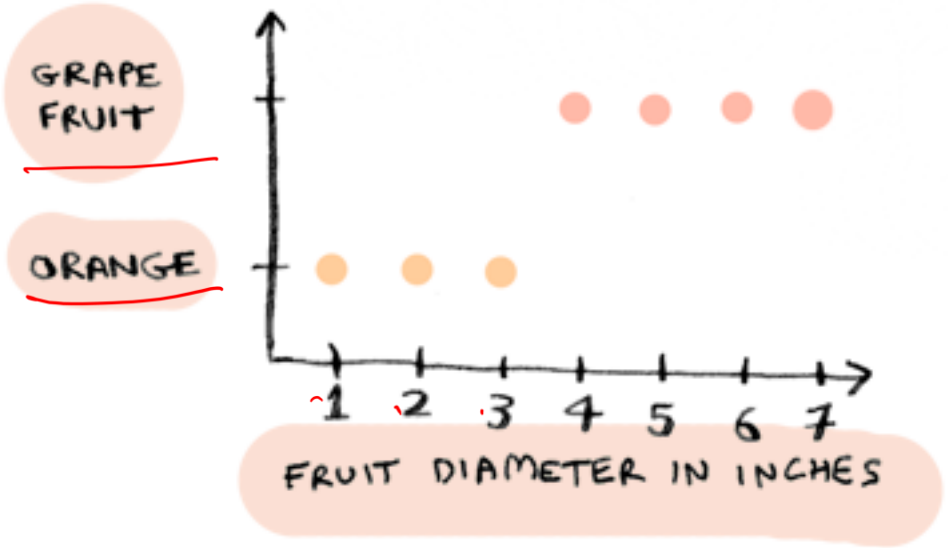
Logistic Regression

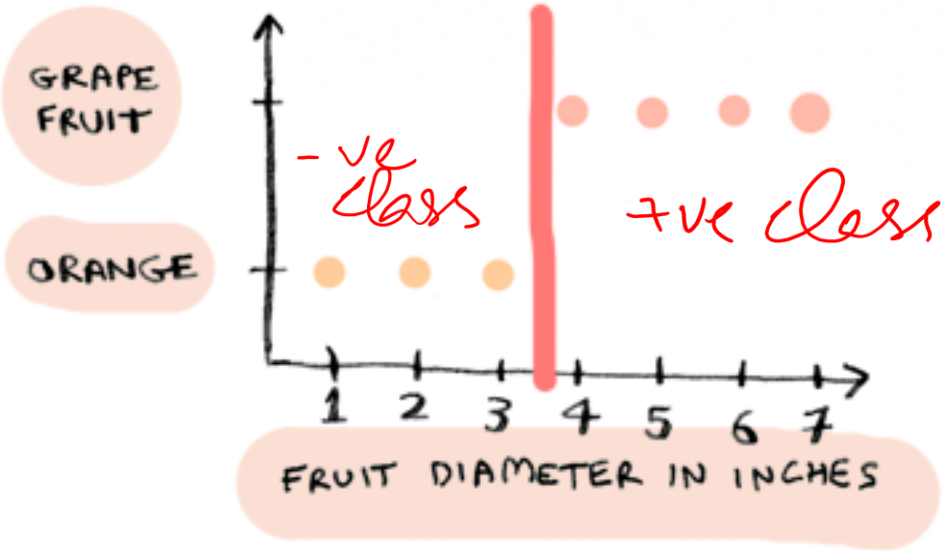
Source: https://www.coursera.org/learn/machine-learning

Logistic Regression

- Supervised Machine Learning algorithm
- Don't be confused by the name "Logistic Regression“.
- It is named that way for historical reasons and is actually an approach to classification problems, not regression problems
- Binary Classification Algorithm
- Instead of our output vector y being a continuous range of values, it will only be 0 or 1. $y \in \{0,1\}$
- Where 0 is usually taken as the "negative class" and 1 as the "positive class", but you are free to assign any representation to it.
- One method is to use linear regression and map all predictions greater than 0.5 as a 1 and all less than 0.5 as a 0.
- This method doesn't work well because classification is not actually a linear function.

Classification of Oranges and Grape Fruit

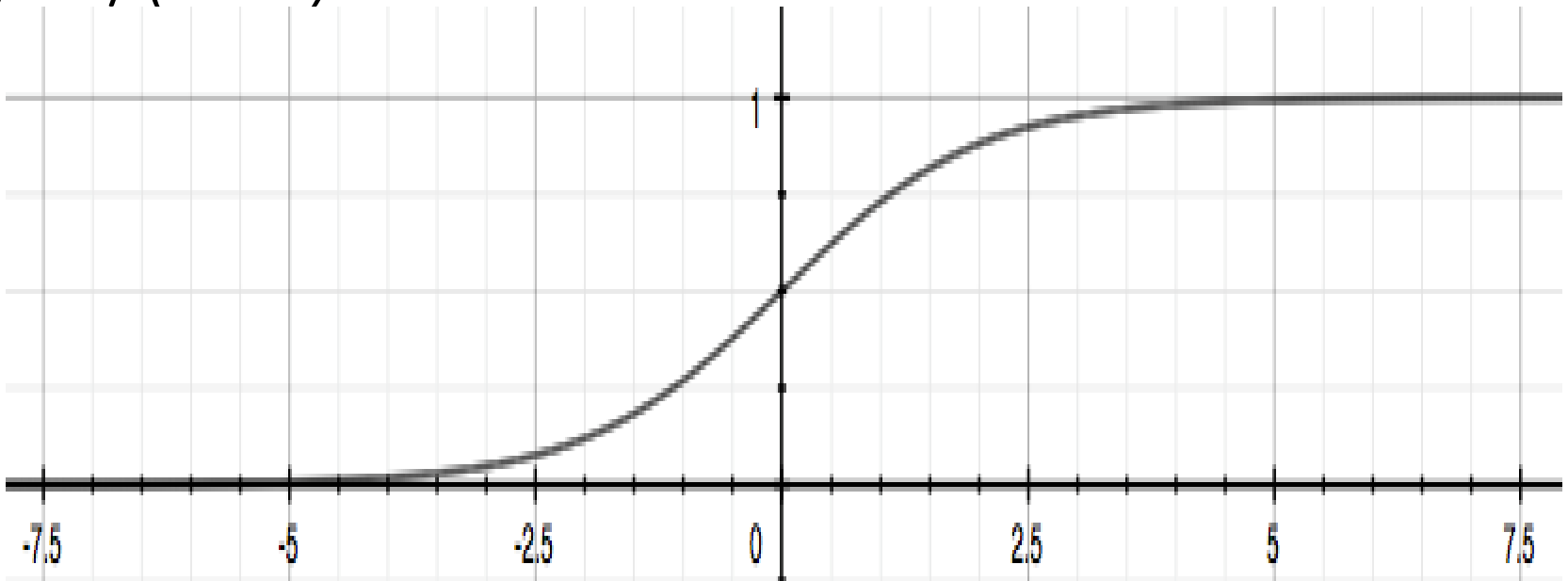




Adit.io

Logistic Regression

- Hypothesis function: $y' = h_\theta(x) = g(\theta_0 + \theta_1 x)$
- Our hypothesis should satisfy:
  - $0 \leq h\vartheta(x) \leq 1$
- Our new form uses the "Sigmoid Function" also called the "Logistic Function"
  - $h_\theta(x) = g(\theta^T x)$
  - $z = \theta^T x$
  - $g(z) = 1 / (1+e^{-z})$



z	sig(z)
-2	0.12
-1.5	0.18
-1	0.27
-0.5	0.38
0	0.50
0.5	0.62
1	0.73
1.5	0.82
2	0.88
2.5	0.92

Logistic Regression

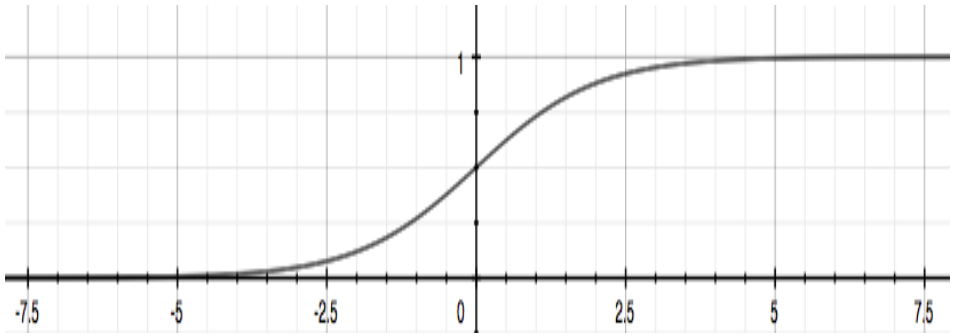
- We start with our old hypothesis (linear regression), except that we want to restrict the range to 0 and 1.
- This is accomplished by plugging $\vartheta^T x$ into the Logistic Function.
- $h\vartheta$ will give us the **probability** that our output is 1.
- For example, $h\vartheta(x)=0.7$ gives us the probability of 70% that our output is 1.
  - $h\vartheta(x) = P(y=1|x;\vartheta) = 1 - P(y=0|x;\vartheta)$
  - $P(y=0|x;\vartheta) + P(y=1|x;\vartheta) = 1$
- Our probability that our prediction is 0 is just the complement of our probability that it is 1 (e.g. if probability that it is 1 is 70%, then the probability that it is 0 is 30%).

Logistic Regression (Decision Boundary)

Source: https://www.coursera.org/learn/machine-learning

Logistic Regression (Decision Boundary)

- In order to get our discrete 0 or 1 classification, we can translate the output of the hypothesis function as follows:
  - $h\vartheta(x) \geq 0.5 \rightarrow y = 1$
  - $h\vartheta(x) < 0.5 \rightarrow y = 0$
- The way our logistic function g behaves is that when its input is greater than or equal to zero, its output is greater than or equal to 0.5:
  - $g(z) \geq 0.5$ *when* $z \geq 0$
- So if our input to g is $\vartheta^T X$, then that means:
  - $h\vartheta(x)=g(\vartheta^T x) \geq 0.5$ *when* $\vartheta^T x \geq 0$
- From these statements we can now say:
  - $\vartheta^T x \geq 0 \Rightarrow y = 1$
  - $\vartheta^T x < 0 \Rightarrow y = 0$



- The **decision boundary** is the line that separates the area where y = 0 and where y = 1. It is created by our hypothesis function.

Example

X	y
1	0
2	0
3	0
4	1
5	1

FRUIT DIAMETER IN INCHES

0 = NOT A GRAPEFRUIT
1 = GRAPEFRUIT

0.5

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

SIGMOID FUNCTION

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

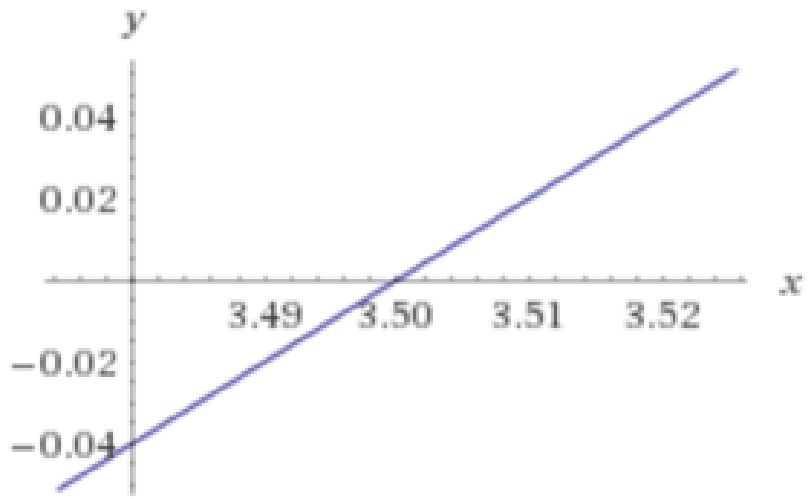
$g(z) < 0.5$
$g(z) \geq 0.5$

$h_\theta(x) = g(-7 + 2x)$

ACTUAL VALUE
1 = IS A GRAPEFRUIT

X	y	PREDICTED
1	0	0.6%
2	0	4.7%
3	0	26.9%
4	1	73.1%
5	1	95.2%

PERCENT CHANCE THAT THIS IS A GRAPEFRUIT



Logistic Regression (Decision Boundary)

- $\Theta = \begin{bmatrix} 5 \\ -1 \\ 0 \end{bmatrix} \begin{matrix} \vartheta_0 \\ \vartheta_1 \\ \vartheta_2 \end{matrix}$

y=1 *if* 5+ (−1) * x_1 + 0 * x_2 ≥ 0

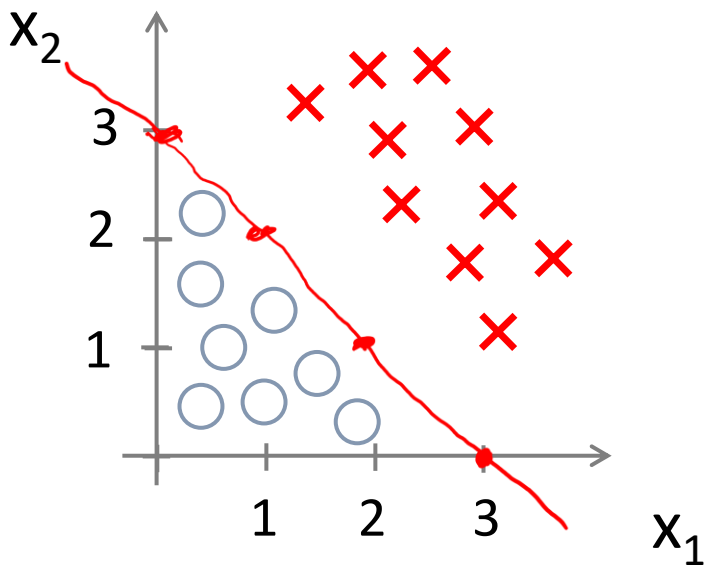
5− x_1 ≥ 0

−x_1 ≥ − 5

x_1 ≤ 5

- In this case, our decision boundary is a straight vertical line placed on the graph where x_1 = 5, and everything to the left of that denotes y = 1, while everything to the right denotes y = 0.
- Again, the input to the sigmoid function g(z) (e.g. $\vartheta^T X$) doesn't need to be linear, and could be a function that describes a circle (e.g. z = $\vartheta_0 + \vartheta_1 x_1^2 + \vartheta_2 x_2^2$) or any shape to fit our data.

Decision Boundary



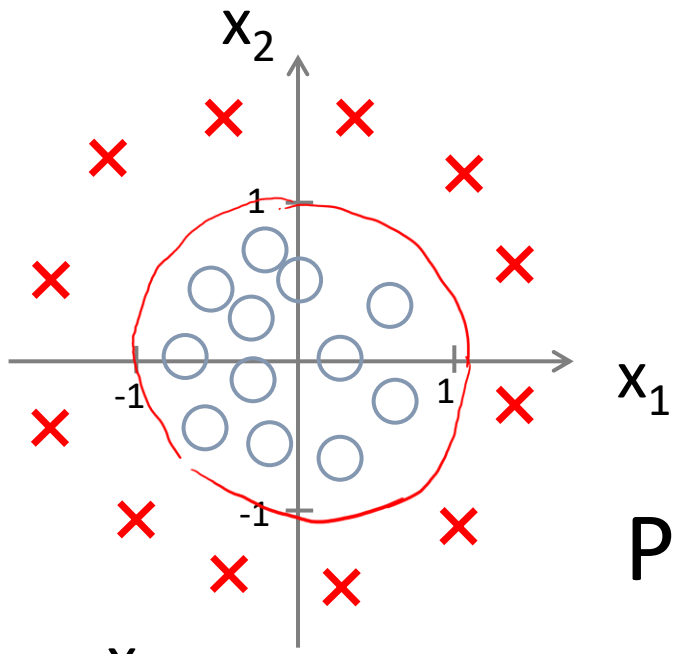
$$\Theta = \begin{bmatrix} -3 \\ 1 \\ 1 \end{bmatrix}$$

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

Predict "$y = 1$" if

$$-3 + x_1 + x_2 \geq 0$$

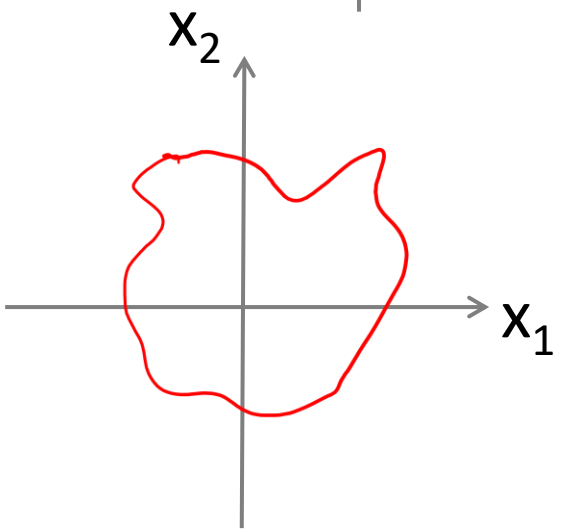
Non-linear decision boundaries



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

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

$x_1^2 + x_2^2 \geq 1$



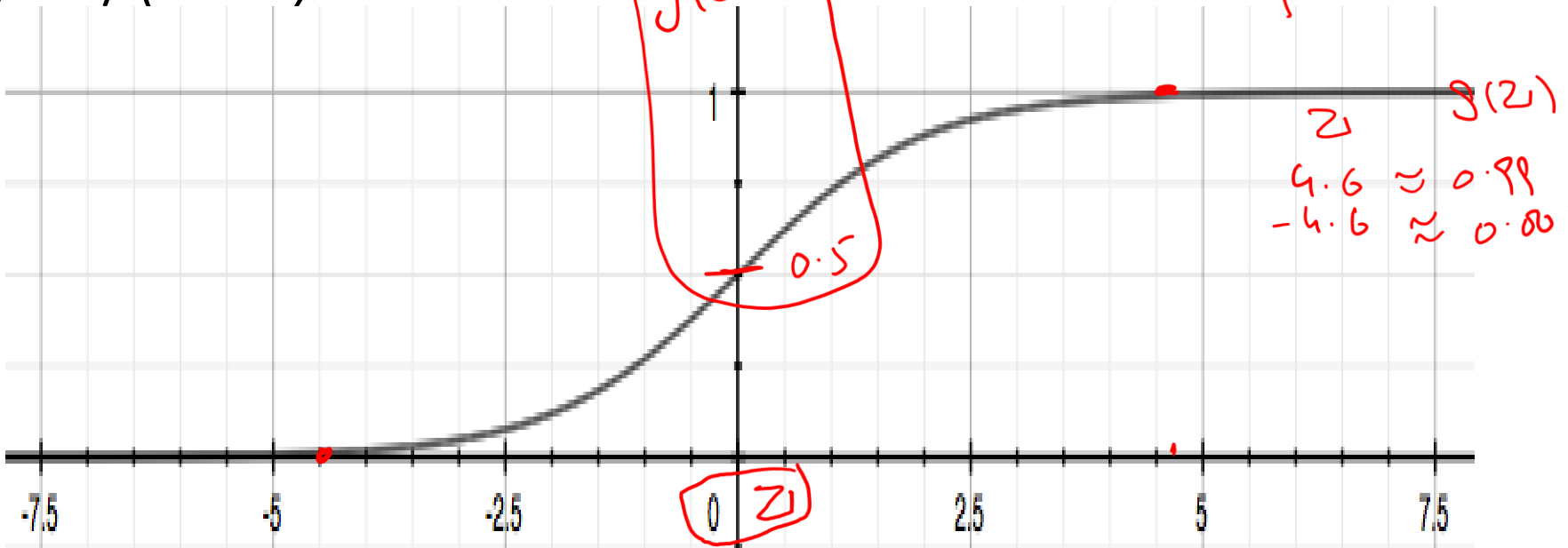
$$h_\theta(x) = g(\theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_1^2 + \theta_4 x_1^2 x_2 + \theta_5 x_1^2 x_2^2 + \theta_6 x_1^3 x_2 + \ldots)$$

Logistic Regression

Source: https://www.coursera.org/learn/machine-learning

Logistic Regression - Recap

- Hypothesis function: $y' = h_\theta(x) = g(\theta_0 + \theta_1 x)$
- Our hypothesis should satisfy:
  - $0 \le h\vartheta(x) \le 1$
- Our new form uses the "Sigmoid Function" also called the "Logistic Function"
  - $h_\theta(x) = g(\theta^T x)$
  - $z = \theta^T x$
  - $g(z) = 1 / (1+e^{-z})$



z	sig(z)
-2	0.12
-1.5	0.18
-1	0.27
-0.5	0.38
0	0.50
0.5	0.62
1	0.73
1.5	0.82
2	0.88
2.5	0.92

Logistic Regression (Cost Function)

Source: https://www.coursera.org/learn/machine-learning

Logistic Regression (Cost Function)

- The more our hypothesis is off from y, the larger the cost function output. If our hypothesis is equal to y, then our cost is 0:

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

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

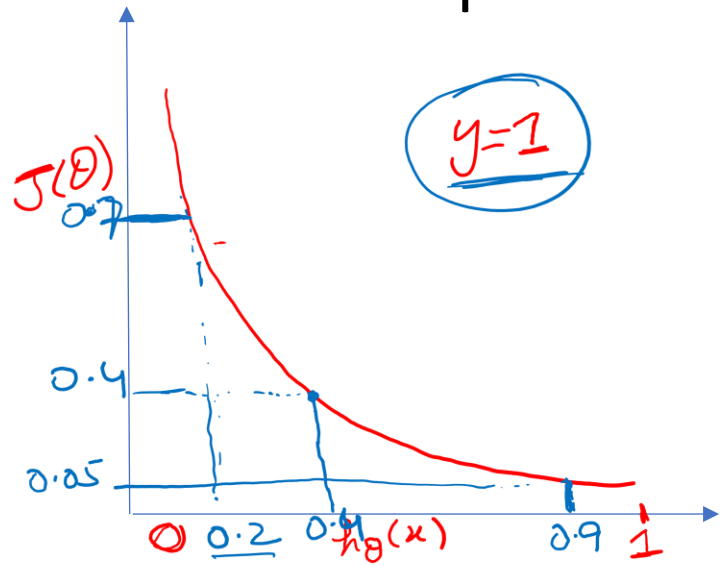
$$\text{Cost}(h_\theta(x), y) = -\log(1 - h_\theta(x)) \quad \text{if } y = 0$$

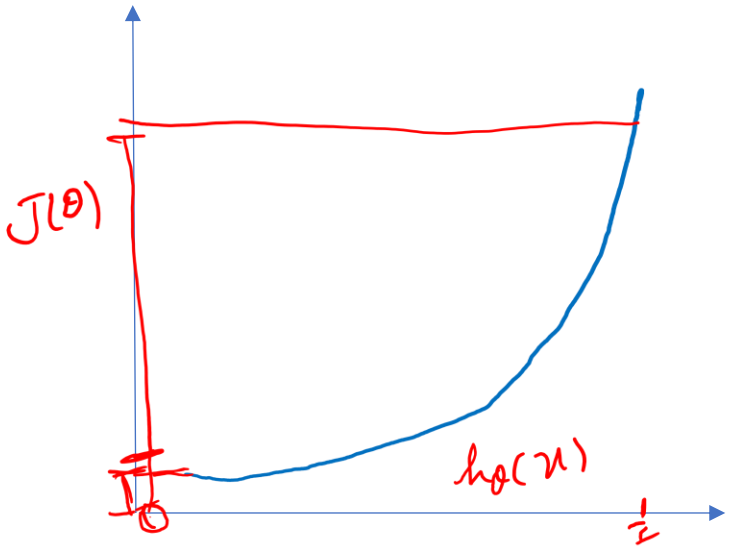
$$\text{Cost}(h_\theta(x), y) = 0 \text{ if } h_\theta(x) = y$$

$$\text{Cost}(h_\theta(x), y) \to \infty \text{ if } y = 0 \text{ and } h_\theta(x) \to 1$$

$$\text{Cost}(h_\theta(x), y) \to \infty \text{ if } y = 1 \text{ and } h_\theta(x) \to 0$$

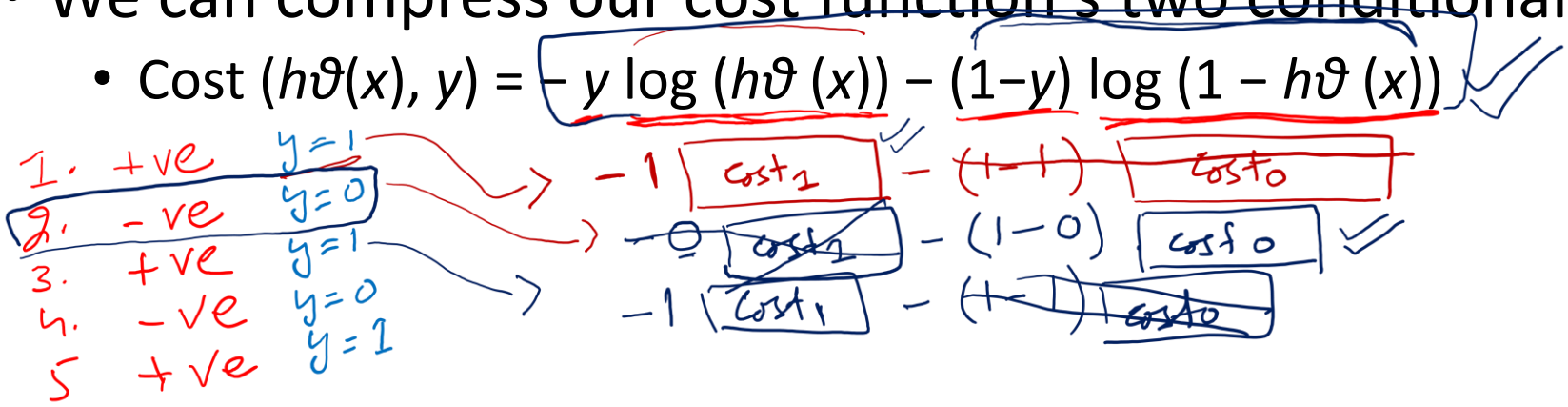
x	-log($h_\vartheta(x)$)	-log(1-$h_\vartheta(x)$)
0.1	1.00	0.05
0.2	0.70	0.10
0.3	0.52	0.15
0.4	0.40	0.22
0.5	0.30	0.30
0.6	0.22	0.40
0.7	0.15	0.52
0.8	0.10	0.70
0.9	0.05	1.00





Logistic Regression (Simplified Cost Function)

- We can compress our cost function's two conditional cases into one case:
  - Cost ($h\vartheta(x)$, y) = $-y \log(h\vartheta(x)) - (1-y) \log(1 - h\vartheta(x))$



- Notice that when y is equal to 1, then the second term $-(1-y) \log(1 - h\vartheta(x))$ will be zero and will not affect the result. If y is equal to 0, then the first term $-y \log(h\vartheta(x))$ will be zero and will not affect the result.
- We can fully write out our entire cost function as follows:
  - $J(\theta) = \frac{-1}{m} \sum_{i=1}^{m} [y^i \log(y'^i) + (1 - y^i) \log(1 - y'^i)]$

Logistic Regression (Gradient Descent)

Source: https://www.coursera.org/learn/machine-learning

Logistic Regression (Gradient Descent)

- *Repeat*{

$$\vartheta_j := \vartheta_j - \alpha \frac{\partial}{\partial \vartheta_j} J(\vartheta)$$

}

learning rate

-ve, +ve

$y'^{(i)} = h_\theta(x^{(i)})$

(i) Hyp ✓
(ii) cost func ✓
(iii) optimization func.

$$y'^{(i)} = h_\vartheta(x^{(i)}) = \sigma(\vartheta_0 + \vartheta_1 x^{(i)})$$

$\theta^T x$

$$\sigma(z) = \frac{1}{1 + e^{-z}}$$

$\left(\frac{1}{1+e^{-z}}\right)\left(1 - \frac{1}{1+e^{-z}}\right)$

$$J(\theta) = \frac{-1}{m}\sum_{i=1}^{m}[y^i \log(y'^i) + (1 - y^i)\log(1 - y'^i)]$$

$$\sigma(z)' = \sigma(z)\left(1 - \sigma(z)\right)$$

0.001, 0.01, 0.1

$h_\theta(x^{(i)}) = \sigma(\theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3)$

Repeat {

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

ϑ_2

}

Sigmoid Derivation

$$\sigma(x) = \frac{1}{1 + e^{-x}}$$

$$\frac{d\sigma(x)}{dx} = \sigma(x)'$$

$$\frac{d\sigma(x)}{dx} = \frac{d}{dx}\left(\frac{1}{1 + e^{-x}}\right)$$

$$= \frac{d}{dx}(1 + e^{-x})^{-1}$$

$$= -1 * (1 + e^{-x})^{-2} \frac{d}{dx}(1 + e^{-x})$$

$$= \frac{-1}{(1 + e^{-x})^2} \frac{d}{dx}(1 + e^{-x})$$

$$= \frac{-1}{(1 + e^{-x})^2}\left(\frac{d}{dx}(1) + \frac{d}{dx}(e^{-x})\right)$$

$$= \frac{-1}{(1 + e^{-x})^2}\left(0 + e^{-x}\frac{d}{dx}(-x)\right)$$

$$= \frac{-1}{(1 + e^{-x})^2}(0 + -e^{-x})$$

$$= \frac{e^{-x}}{(1 + e^{-x})^2}$$

$$= \left(\frac{1}{1 + e^{-x}}\right)\left(\frac{e^{-x}}{1 + e^{-x}}\right)$$

$$= \sigma(x)\left(\frac{e^{-x}}{1 + e^{-x}}\right)$$

$$= \sigma(x)\left(\frac{1 - 1 + e^{-x}}{1 + e^{-x}}\right)$$

$$= \sigma(x)\left(\frac{1 + e^{-x} - 1}{1 + e^{-x}}\right)$$

$$= \sigma(x)\left(\frac{1 + e^{-x}}{1 + e^{-x}} - \frac{1}{1 + e^{-x}}\right)$$

$$= \sigma(x)\left(1 - \sigma(x)\right)$$

https://medium.com/analytics-vidhya/derivative-of-log-loss-function-for-logistic-regression-9b832f025c2d

Binary Cross Entropy

Cost function:

$$J(\theta) = \frac{-1}{m}\sum_{i=1}^{m}[y^i \log(y'^i) + (1 - y^i)\log(1 - y'^i)]$$

$$\frac{\partial J(\theta)}{\partial \theta_j} = \frac{\partial}{\partial \theta_j}\frac{-1}{m}\sum_{i=1}^{m}[y^i \log(y'^i) + (1 - y^i)\log(1 - y'^i)]$$

$$= \frac{-1}{m}\sum_{i=1}^{m}[y^i \frac{\partial}{\partial \theta_j}\log(y'^i) + (1 - y^i)\frac{\partial}{\partial \theta_j}\log(1 - y'^i)]$$

$$= \frac{-1}{m}\sum_{i=1}^{m}[\frac{y^i \frac{\partial}{\partial \theta_j} y'^i}{y'^i} + \frac{(1 - y^i)\frac{\partial}{\partial \theta_j}(1 - y'^i)}{1 - y'^i}]$$

$$= \frac{-1}{m}\sum_{i=1}^{m}[\frac{y^i \frac{\partial}{\partial \theta_j} y'^i}{y'^i} + \frac{(1 - y^i)\frac{\partial}{\partial \theta_j}(1 - y'^i)}{1 - y'^i}]$$

$$= \frac{-1}{m}\sum_{i=1}^{m}[\frac{y^i\ (y'^i)(1 - y'^i)\frac{\partial}{\partial \theta_j}\theta^T x}{y'^i} + \frac{-(1 - y^i)\ (y'^i)(1 - y'^i)\frac{\partial}{\partial \theta_j}\theta^T x}{1 - y'^i}]$$

Objective:

$$\min_{\vartheta_0, \vartheta_1} J(\vartheta_0, \vartheta_1)$$

Hypothesis Function:

$$y'^{(i)} = h_\vartheta(x^{(i)}) = \sigma(\vartheta_0 + \vartheta_1 x^{(i)})$$

$$= \frac{-1}{m}\sum_{i=1}^{m}[\frac{y^i\ (y'^i)(1 - y'^i)\frac{\partial}{\partial \theta_j}\theta^T x}{y'^i} - \frac{(1 - y^i)\ (y'^i)(1 - y'^i)\frac{\partial}{\partial \theta_j}\theta^T x}{1 - y'^i}]$$

$$= \frac{-1}{m}\sum_{i=1}^{m}[y^i(1 - y'^i)x_j^i - (1 - y^i)y'^i x_j^i]$$

$$= \frac{-1}{m}\sum_{i=1}^{m}[y^i(1 - y'^i) - (1 - y^i)y'^i]x_j^i$$

$$= \frac{-1}{m}\sum_{i=1}^{m}[y^i - y^i y'^i - y'^i + y^i y'^i]x_j^i$$

$$= \frac{-1}{m}\sum_{i=1}^{m}[y^i - y'^i]x_j^i$$

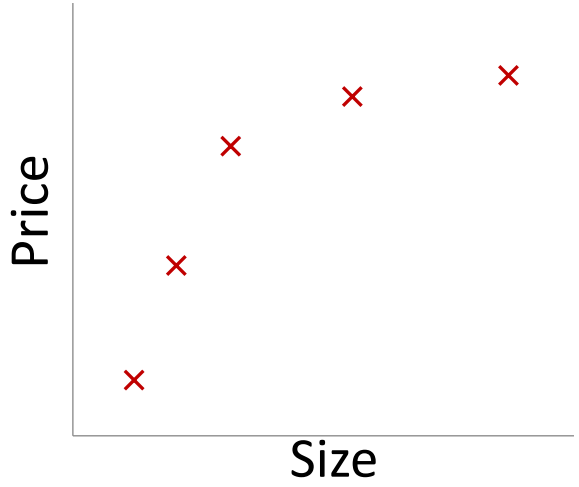
$$= \frac{1}{m}\sum_{i=1}^{m}[y'^i - y^i]x_j^i$$

Regularization

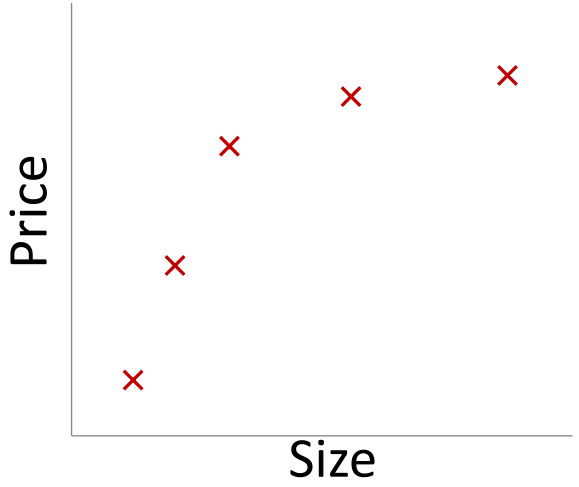
The problem of overfitting

Source: https://www.coursera.org/learn/machine-learning

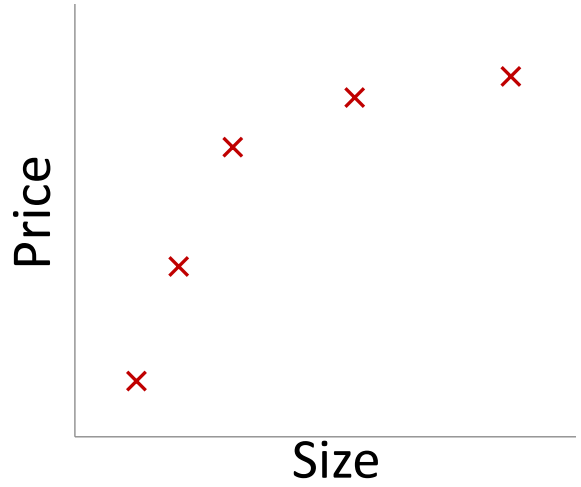
Example: Linear regression (plot prices)



$\theta_0 + \theta_1 x$



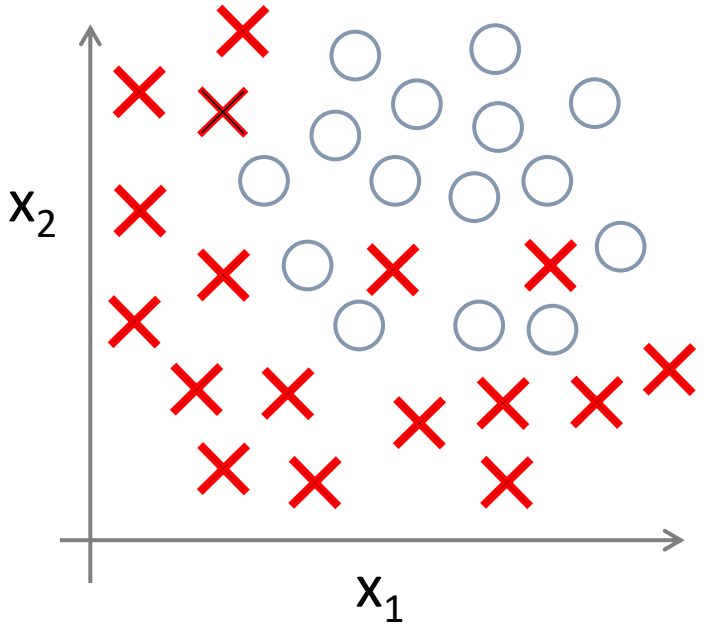
$\theta_0 + \theta_1 x + \theta_2 x^2$



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

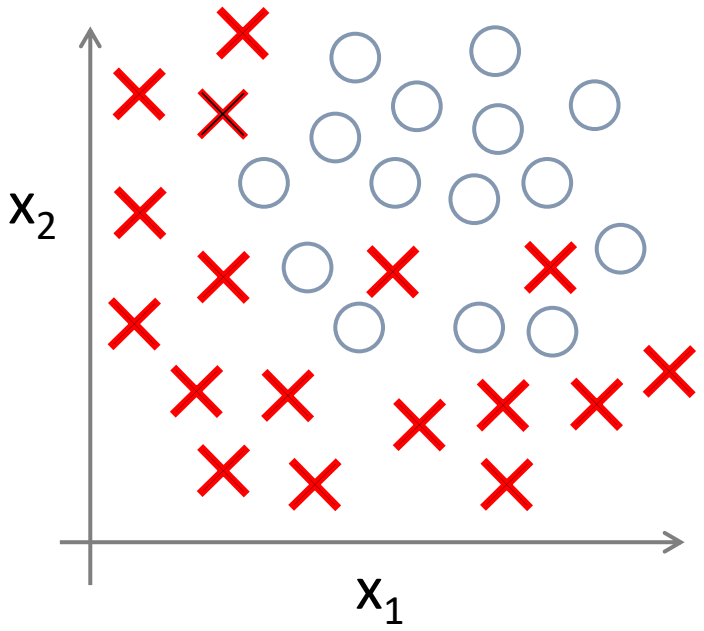
Overfitting: If we have too many features, the learned hypothesis may fit the training set very well ($J(\theta) = \frac{1}{2m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)})^2 \approx 0$), but fail to generalize to new examples (predict prices on new examples).

Example: Logistic regression

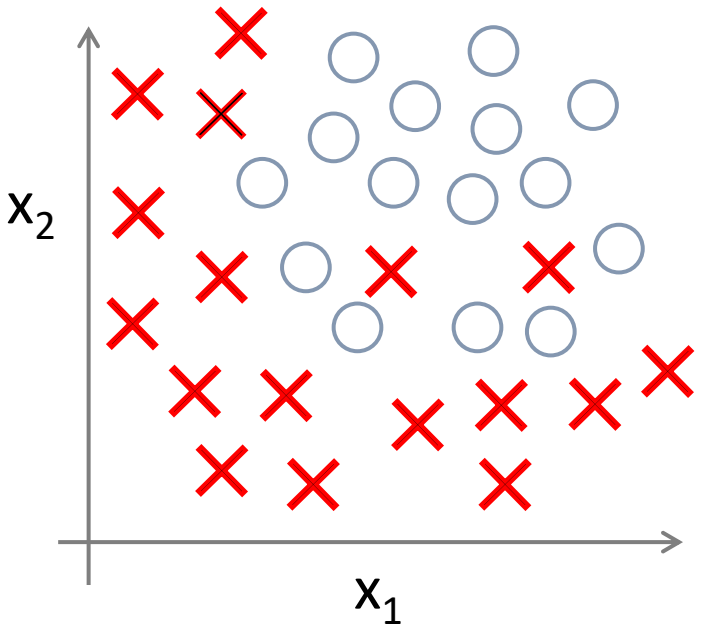


$h_\theta(x) = g(\theta_0 + \theta_1 x_1 + \theta_2 x_2)$

(g = sigmoid function)



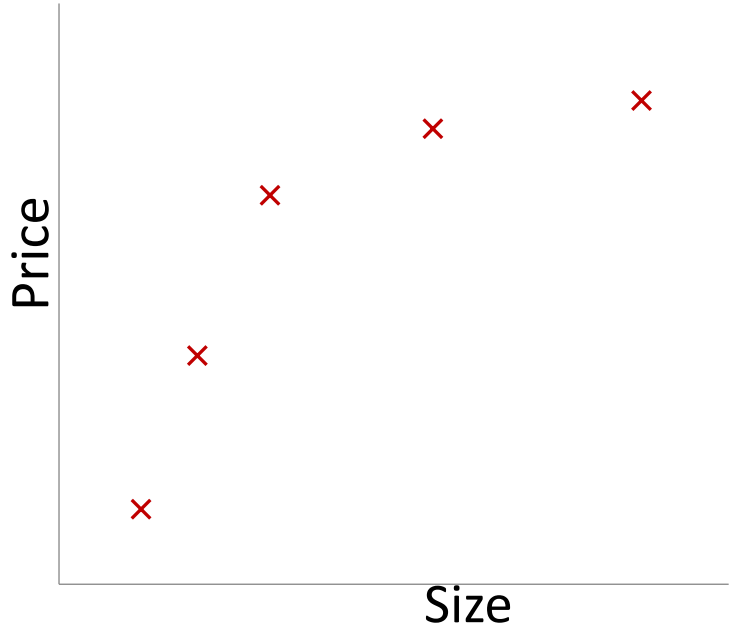
$g(\theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_1^2 + \theta_4 x_2^2 + \theta_5 x_1 x_2)$



$g(\theta_0 + \theta_1 x_1 + \theta_2 x_1^2 + \theta_3 x_1^2 x_2 + \theta_4 x_1^2 x_2^2 + \theta_5 x_1^2 x_2^3 + \theta_6 x_1^3 x_2 + \dots)$

Addressing overfitting:

$x_1 =$ size of house
$x_2 =$ no. of bedrooms
$x_3 =$ no. of floors
$x_4 =$ age of house
$x_5 =$ average income in neighborhood
$x_6 =$ kitchen size
$\vdots$
x_{100}



Addressing overfitting:

Options:

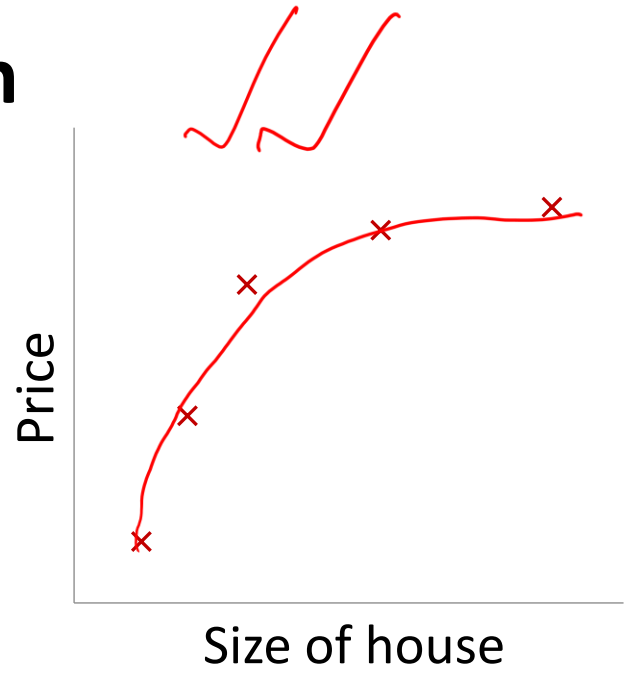
1. Reduce number of features.
   - Feature selection
     - Manually select which features to keep
     - Feature selection algorithm
2. Regularization.
   - Keep all the features, but reduce magnitude/values of parameters θ_j.
   - Works well when we have a lot of features, each of which contributes a bit to predicting y .

Regularization

Cost Function

Source: https://www.coursera.org/learn/machine-learning

Intuition



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

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

Suppose we penalize and make θ_3, θ_4 really small.

$$\min_{\theta} \frac{1}{2m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)})^2 + \lambda \theta_1 + \lambda \theta_2 + \lambda \theta_3 \cdots \lambda \theta_n$$

$$+ \lambda \sum_{j=1}^{n} \theta_j^2$$

Regularization.

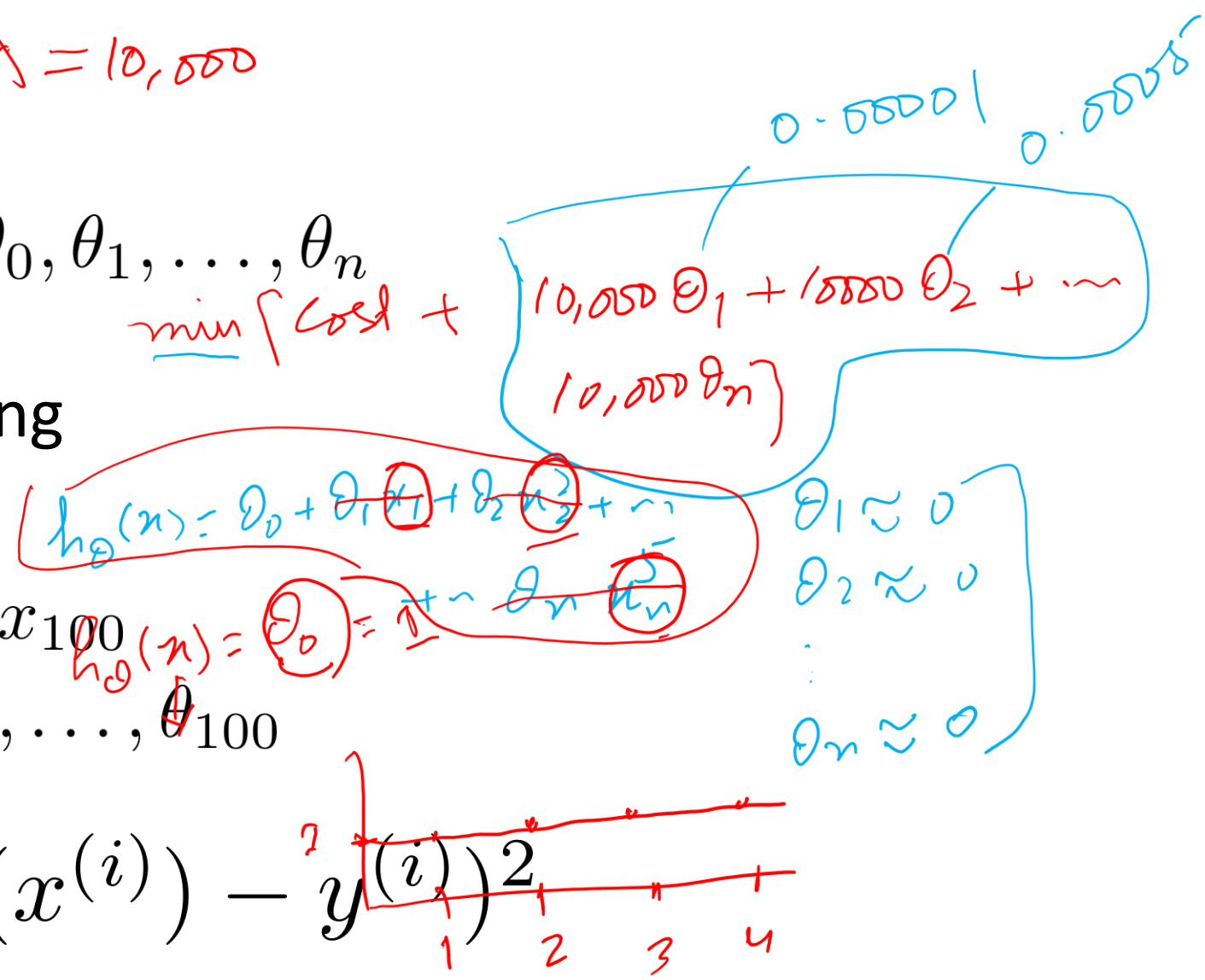
Small values for parameters $\theta_0, \theta_1, \ldots, \theta_n$

- "Simpler" hypothesis
- Less prone to overfitting

Plot Price:

- Features: $x_1, x_2, \ldots, x_{100}$
- Parameters: $\theta_0, \theta_1, \theta_2, \ldots, \theta_{100}$

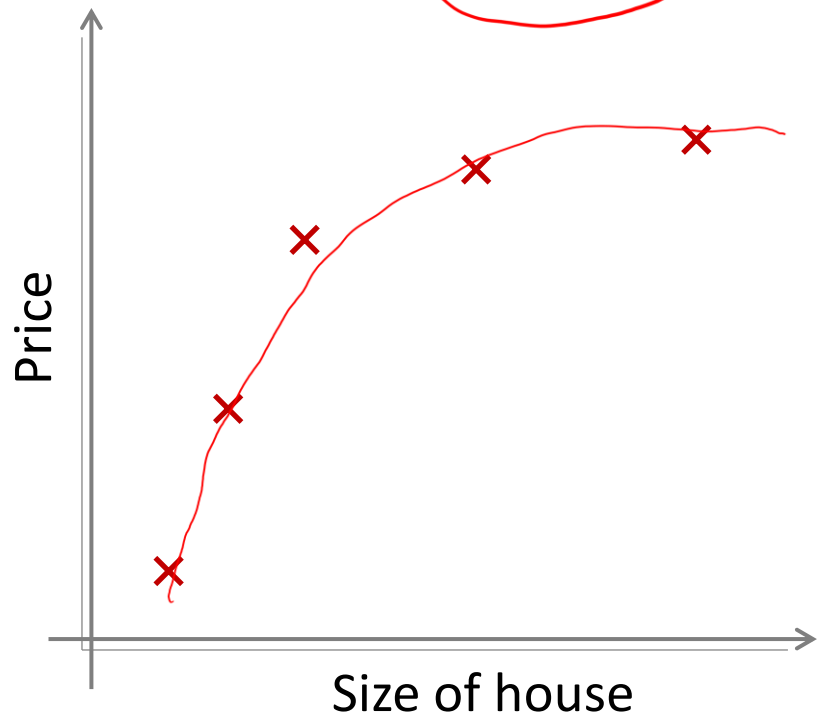
$$J(\theta) = \frac{1}{2m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)})^2$$



Regularization.

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

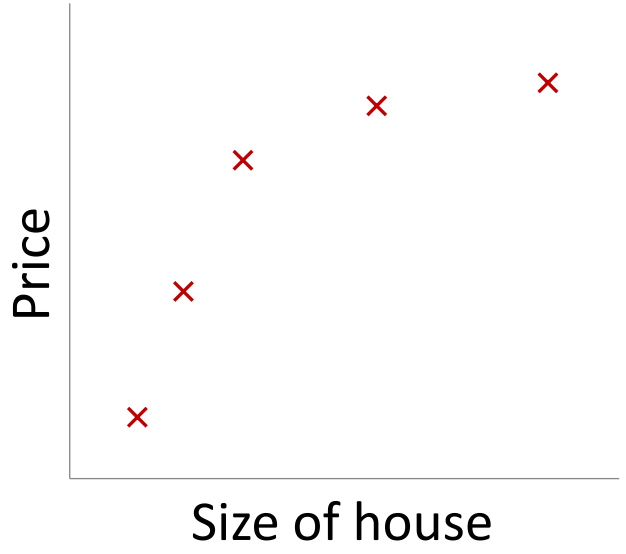
$$\min_\theta J(\theta)$$



In regularized linear regression, we choose θ to minimize

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

What if λ is set to an extremely large value (perhaps for too large for our problem, say $\lambda = 10^{10}$)?



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

Regularization

Regularized linear regression

Source: https://www.coursera.org/learn/machine-learning

Regularized linear regression

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

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

$$\theta_0 = \theta_0 - \alpha\frac{1}{m}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})1$$

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

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

$$\theta_j = (1 - \alpha\frac{\lambda}{m})\theta_j$$

Gradient descent

Repeat $\{$

$$\theta_0 := \theta_0 - \alpha \frac{1}{m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)}) x_0^{(i)}$$

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

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

$\}$

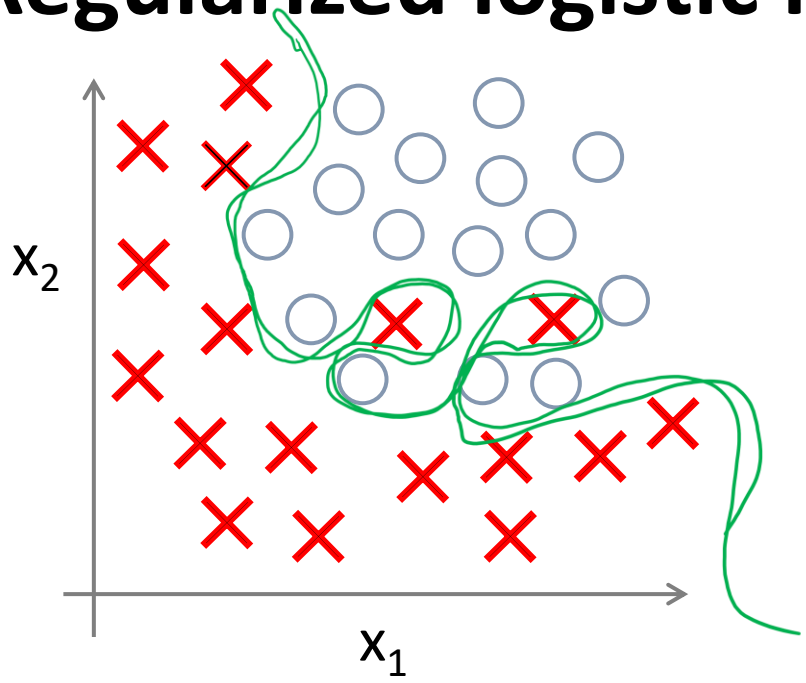
$$\theta_j := \theta_j(1 - \alpha \frac{\lambda}{m}) - \alpha \frac{1}{m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)}) x_j^{(i)}$$

Regularization

Regularized logistic regression

Source: https://www.coursera.org/learn/machine-learning

Regularized logistic regression.



$$h_\theta(x) = g(\theta_0 + \theta_1 x_1 + \theta_2 x_1^2 + \theta_3 x_1^2 x_2 + \theta_4 x_1^2 x_2^2 + \theta_5 x_1^2 x_2^3 + \dots)$$

Cost function:

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

Gradient descent

Repeat $\{$

$$\theta_0 := \theta_0 - \alpha \frac{1}{m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)}) x_0^{(i)}$$

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

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

$\}$

$\theta_0 + \theta_1 x_1 + \theta_2 x_2$

$g(\theta_0 + \theta_1 x_1 + \theta_2 x_2)$

$\frac{1}{1 + e^{-(\theta_0 + \theta_1 x_1 + \theta_2 x_2)}}$

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

Sigmoid Derivation

$$\sigma(x) = \frac{1}{1 + e^{-x}}$$

$$\frac{d\sigma(x)}{dx} = \sigma(x)'$$

$$\frac{d\sigma(x)}{dx} = \frac{d}{dx}\left(\frac{1}{1 + e^{-x}}\right)$$

$$= \frac{d}{dx}(1 + e^{-x})^{-1}$$

$$= -1 * (1 + e^{-x})^{-2}\frac{d}{dx}(1 + e^{-x})$$

$$= \frac{-1}{(1 + e^{-x})^2}\frac{d}{dx}(1 + e^{-x})$$

$$= \frac{-1}{(1 + e^{-x})^2}\left(\frac{d}{dx}(1) + \frac{d}{dx}(e^{-x})\right)$$

$$= \frac{-1}{(1 + e^{-x})^2}\left(0 + e^{-x}\frac{d}{dx}(-x)\right)$$

$$= \frac{-1}{(1 + e^{-x})^2}(0 + -e^{-x})$$

$$= \frac{e^{-x}}{(1 + e^{-x})^2}$$

$$= \left(\frac{1}{1 + e^{-x}}\right)\left(\frac{e^{-x}}{1 + e^{-x}}\right)$$

$$= \sigma(x)\left(\frac{e^{-x}}{1 + e^{-x}}\right)$$

$$= \sigma(x)\left(\frac{1 - 1 + e^{-x}}{1 + e^{-x}}\right)$$

$$= \sigma(x)\left(\frac{1 + e^{-x} - 1}{1 + e^{-x}}\right)$$

$$= \sigma(x)\left(\frac{1 + e^{-x}}{1 + e^{-x}} - \frac{1}{1 + e^{-x}}\right)$$

$$= \sigma(x)\left(1 - \sigma(x)\right)$$

https://medium.com/analytics-vidhya/derivative-of-log-loss-function-for-logistic-regression-9b832f025c2d

Binary Cross Entropy

Cost function:

$$J(\theta) = \frac{-1}{m}\sum_{i=1}^{m}[y^i \log(y'^i) + (1 - y^i)\log(1 - y'^i)]$$

$$\frac{\partial J(\theta)}{\partial \theta_j} = \frac{\partial}{\partial \theta_j}\frac{-1}{m}\sum_{i=1}^{m}[y^i \log(y'^i) + (1 - y^i)\log(1 - y'^i)]$$

$$= \frac{-1}{m}\sum_{i=1}^{m}[y^i \frac{\partial}{\partial \theta_j}\log(y'^i) + (1 - y^i)\frac{\partial}{\partial \theta_j}\log(1 - y'^i)]$$

$$= \frac{-1}{m}\sum_{i=1}^{m}[\frac{y^i \frac{\partial}{\partial \theta_j} y'^i}{y'^i} + \frac{(1 - y^i)\frac{\partial}{\partial \theta_j}(1 - y'^i)}{1 - y'^i}]$$

$$= \frac{-1}{m}\sum_{i=1}^{m}[\frac{y^i \frac{\partial}{\partial \theta_j} y'^i}{y'^i} + \frac{(1 - y^i)\frac{\partial}{\partial \theta_j}(1 - y'^i)}{1 - y'^i}]$$

$$= \frac{-1}{m}\sum_{i=1}^{m}[\frac{y^i\,(y'^i)(1 - y'^i)\frac{\partial}{\partial \theta_j}\theta^T x}{y'^i} + \frac{-(1 - y^i)\,(y'^i)(1 - y'^i)\frac{\partial}{\partial \theta_j}\theta^T x}{1 - y'^i}]$$

Objective:

$$\min_{\vartheta_0,\vartheta_1} J(\vartheta_0,\vartheta_1)$$

Hypothesis Function:

$$y'^{(i)} = h_\vartheta(x^{(i)}) = \sigma(\vartheta_0 + \vartheta_1 x^{(i)})$$

$$= \frac{-1}{m}\sum_{i=1}^{m}[\frac{y^i\,(y'^i)(1 - y'^i)\frac{\partial}{\partial \theta_j}\theta^T x}{y'^i} - \frac{(1 - y^i)\,(y'^i)(1 - y'^i)\frac{\partial}{\partial \theta_j}\theta^T x}{1 - y'^i}]$$

$$= \frac{-1}{m}\sum_{i=1}^{m}[y^i(1 - y'^i)x_j^i - (1 - y^i)y'^i x_j^i]$$

$$= \frac{-1}{m}\sum_{i=1}^{m}[y^i(1 - y'^i) - (1 - y^i)y'^i]x_j^i$$

$$= \frac{-1}{m}\sum_{i=1}^{m}[y^i - y^i y'^i - y'^i + y^i y'^i]x_j^i$$

$$= \frac{-1}{m}\sum_{i=1}^{m}[y^i - y'^i]x_j^i$$

$$= \frac{1}{m}\sum_{i=1}^{m}[y'^i - y^i]x_j^i$$

$h_\theta(x^{(i)})$

Reading - Homework

- Resource R1
- Book B1: 3.2, Chapter 5 (5.1 to 5.7)
- Book B2: Chapter 3 and 4