MACHINE LEARNING COURSE

PRESENTED BY ABDEL RAHMAN ALSABBAGH

LECTURE #4 – SAT - 20.5.2023

بِسْمِ ٱللَّهِ ٱلرَّحْمَٰنِ ٱلرَّحِيمِ

In the name of Allah, the most gracious, the most merciful, we start :)

Today's Quote

"You are responsible for the pursuit, not the outcome"

\- Bryant McGill

Introduction to Classification

- Logistic regression.
- Decision boundary.
- Cost function for logistic regression.
- Gradient descent implementation.
- Overfitting.
- Addressing overfitting.
- Cost function with regularization.
- Regularized gradient descent.

TECH REQUIREMENTS

Source: Machine Learning Specialization by Andrew Ng and Stanford Online.

Classification

Question	Answer "y"	
Is this email spam?	no	yes
Is the transaction fraudulent?	no	yes
Is the tumor malignant?	no	yes

$\boxed{y \text{ can only be one of two values}}$

false true

0 1

useful for classification

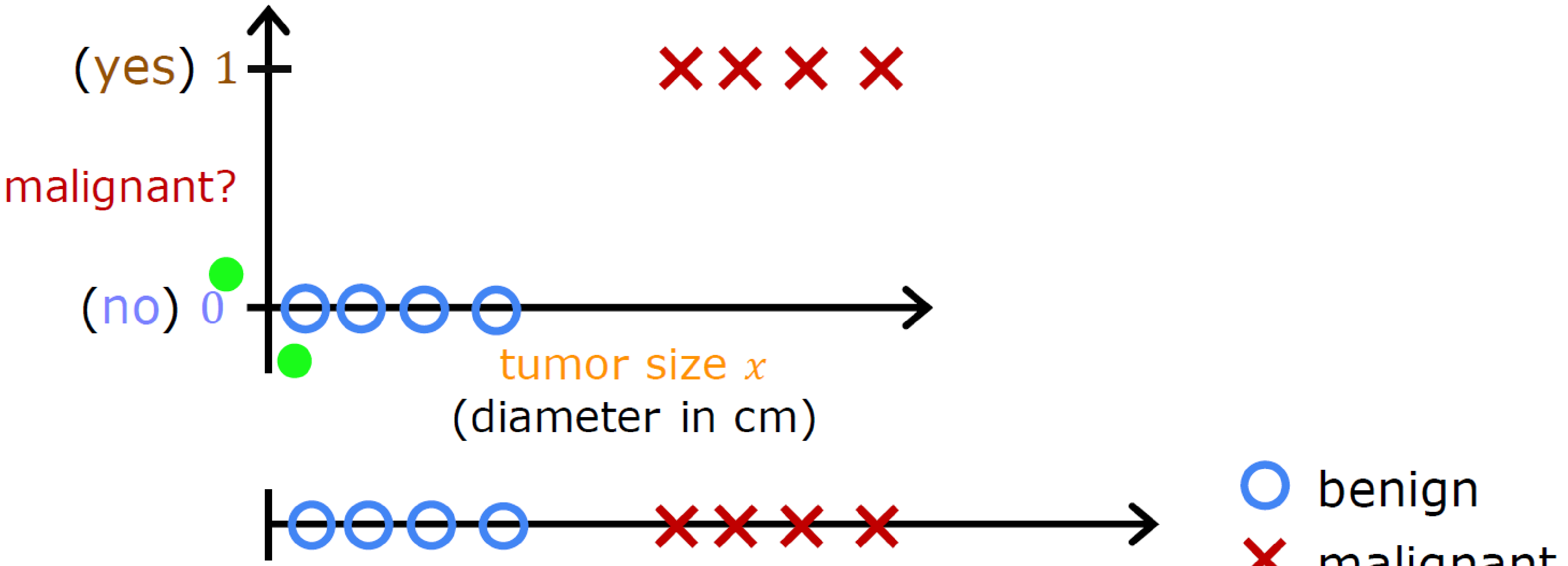
$\boxed{\text{"binary classification"}}$

class = category

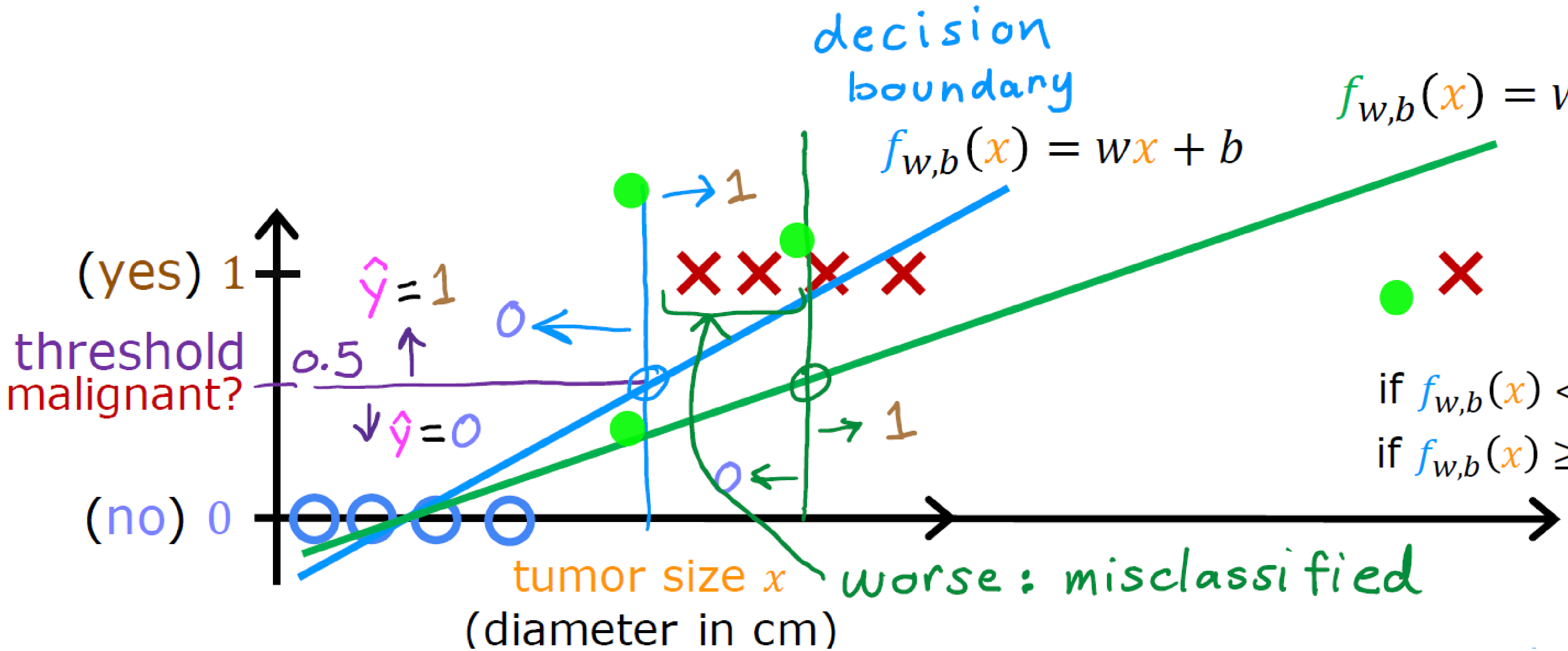
"negative class" ≠ "bad" absence

"positive class" ≠ "good" presence

Classification



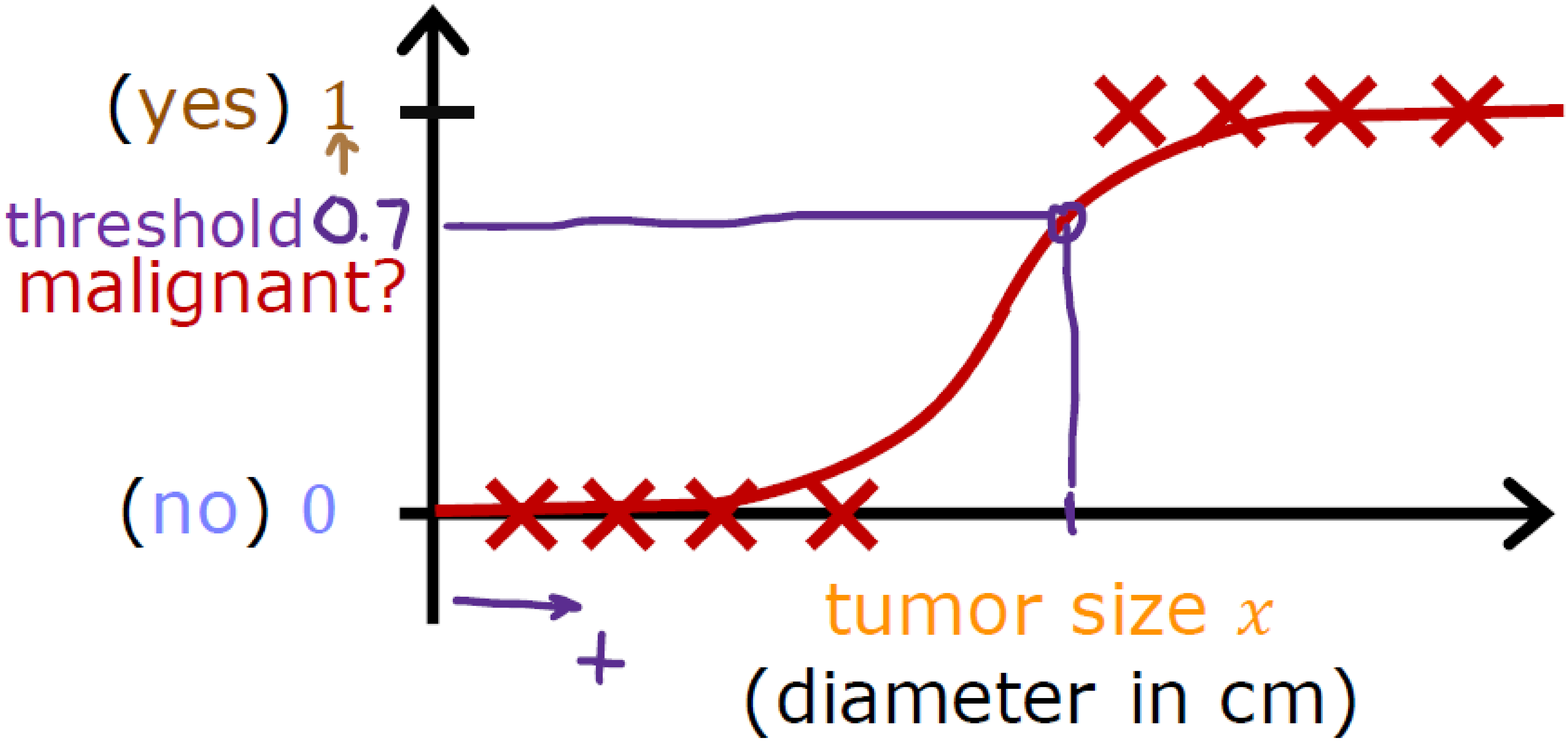
Classification



if $f_{w,b}(x) < 0.5 \rightarrow \hat{y} = 0$

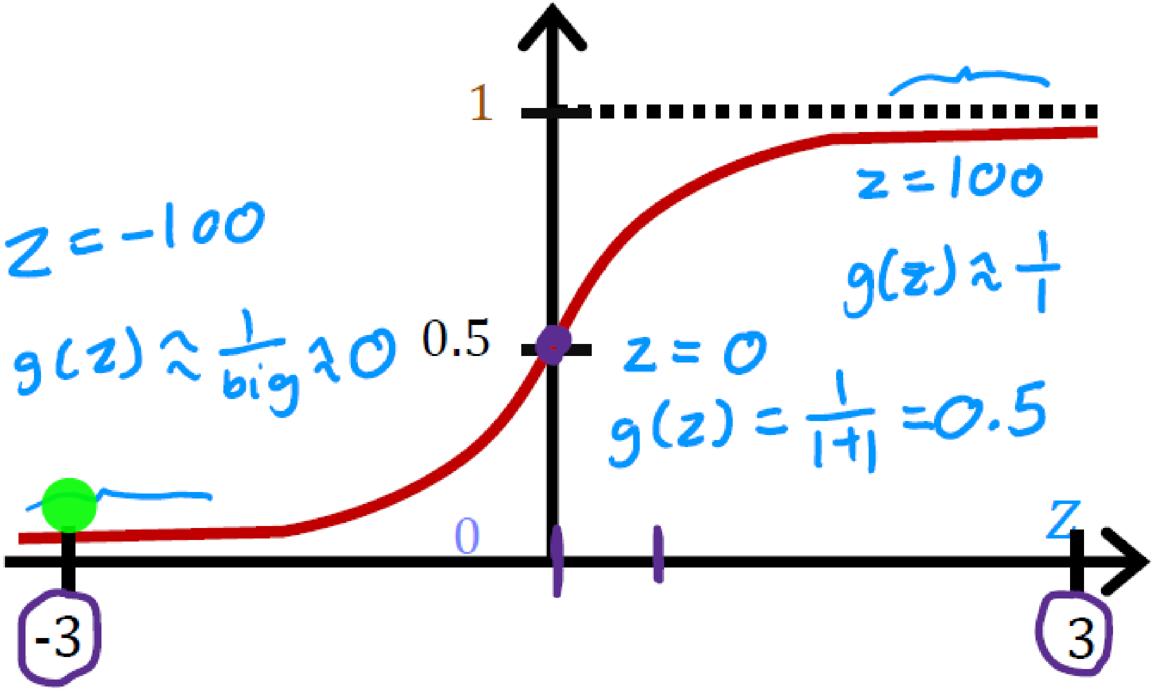
if $f_{w,b}(x) \geq 0.5 \rightarrow \hat{y} = 1$

Logistic Regression



Logistic Regression

Want outputs between 0 and 1



sigmoid function

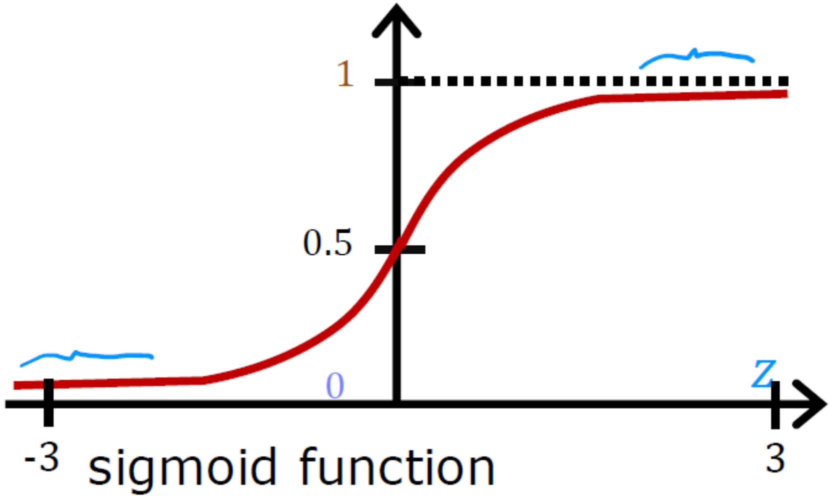
logistic function

outputs between 0 and 1

$$g(z) = \frac{1}{1+e^{-z}} \qquad 0 < g(z) < 1$$

Logistic Regression

Want outputs between 0 and 1



sigmoid function

logistic function

$$\boxed{\text{outputs between 0 and 1} \quad g(z) = \frac{1}{1+e^{-z}} \qquad 0 < g(z) < 1}$$

$$f_{\vec{w},b}(\vec{x})$$

$$\boxed{\vec{w} \cdot \vec{x} + b}$$

$$\downarrow$$

$$z$$

$$\downarrow$$

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

$$f_{\vec{w},b}(\vec{x}) = g(\underbrace{\vec{w} \cdot \vec{x} + b}_{z}) = \frac{1}{1 + e^{-(\vec{w}\cdot\vec{x}+b)}}$$

"logistic regression"

$e \approx 2.7$

Logistic Regression

$$f_{\vec{w},b}(\vec{x}) = \frac{1}{1 + e^{-(\vec{w}\cdot\vec{x}+b)}}$$

"probability" that class is 1

Example:

x is "tumor size"
y is 0 (not malignant)
or 1 (malignant)

$$f_{\vec{w},b}(\vec{x}) = 0.7$$

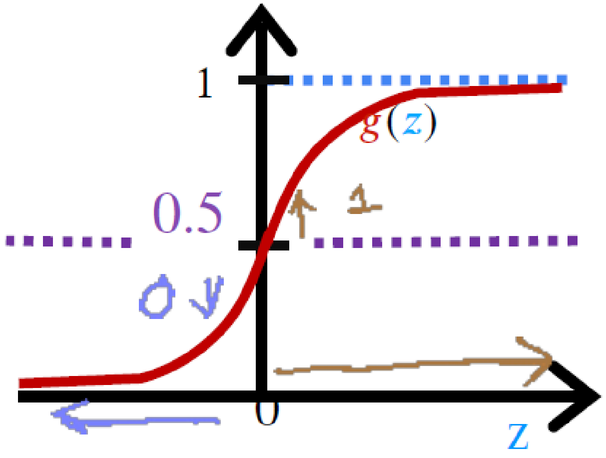
70% chance that y is 1

$$f_{\vec{w},b}(\vec{x}) = P(y = 1|\vec{x};\vec{w},b)$$

Probability that y is 1,
given input $\vec{x}$, parameters $\vec{w},b$

$$P(y=0) + P(y=1) = 1$$

Decision Boundary



Is $f_{\vec{w},b}(\vec{x}) \geq 0.5$?

Yes: $\hat{y} = 1$ No: $\hat{y} = 0$

When is $f_{\vec{w},b}(\vec{x}) \geq 0.5$?

$g(z) \geq 0.5$

$z \geq 0$ $z < 0$

$\vec{w} \cdot \vec{x} + b \geq 0$ $\vec{w} \cdot \vec{x} + b < 0$

$\hat{y} = 1$ $\hat{y} = 0$

Decision Boundary

$$f_{\vec{w},b}(\vec{x}) = g(z) = g(\overbrace{\underset{1}{w_1}x_1 + \underset{1}{w_2}x_2 + \underset{-3}{b}})$$

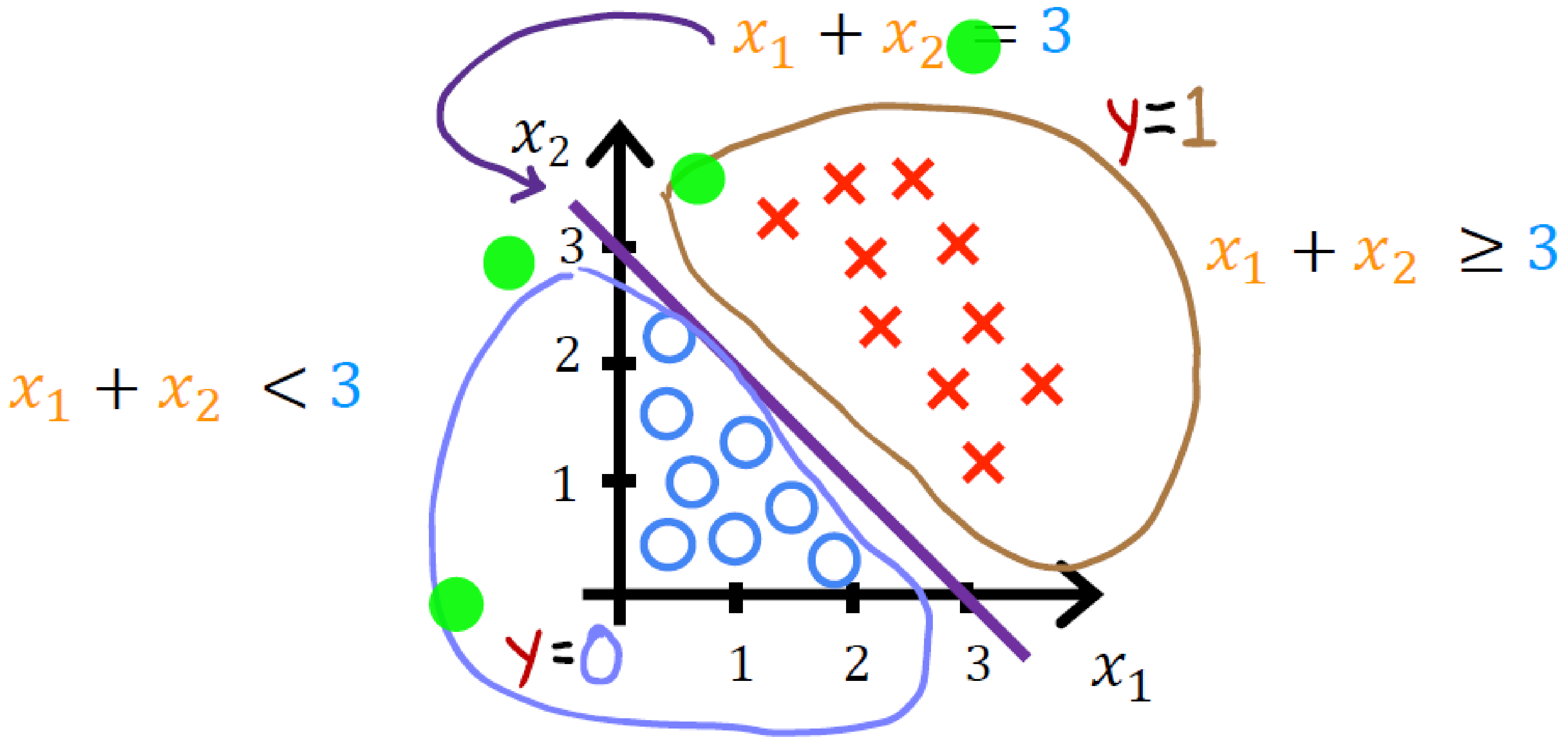
Decision boundary

$$z = \vec{w} \cdot \vec{x} + b = 0$$

$$z = x_1 + x_2 - 3 = 0$$

$$x_1 + x_2 = 3$$

Decision Boundary



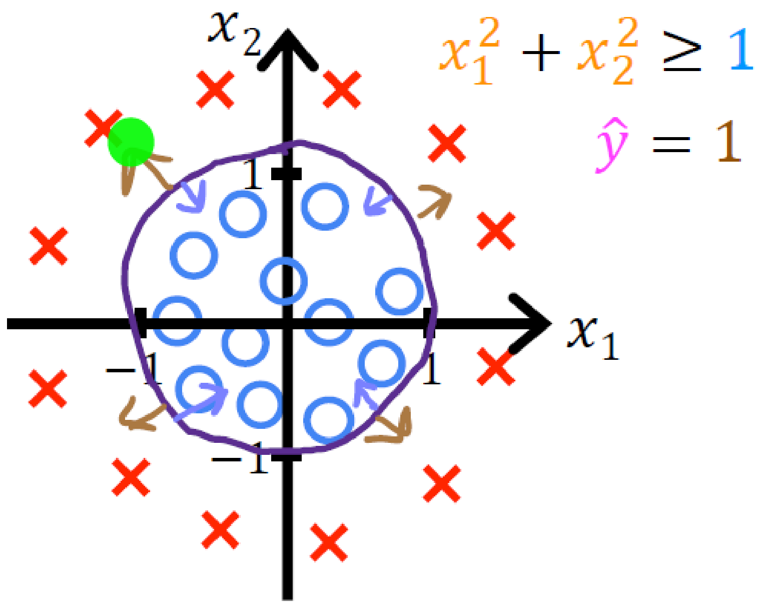
Non-linear Decision Boundaries

$$\overbrace{\underset{1}{w_1} x_1^2 + \underset{1}{w_2} x_2^2 + \underset{-1}{b}}^{z}$$

decision boundary

$$z = x_1^2 + x_2^2 - 1 = 0$$

$$x_1^2 + x_2^2 = 1$$

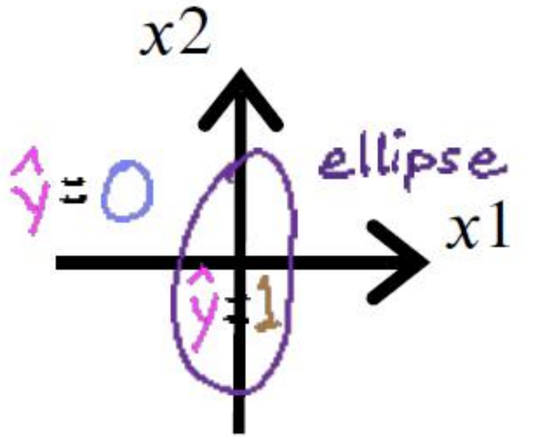


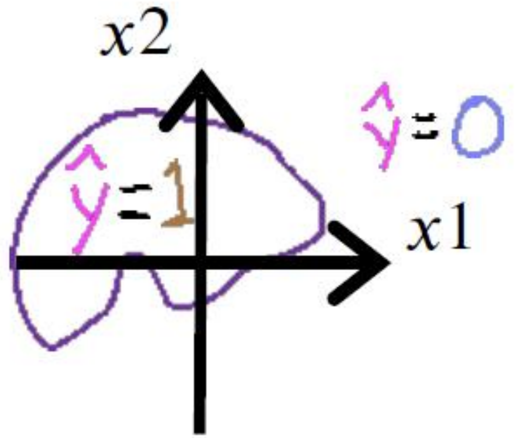
$x_1^2 + x_2^2 < 1$

$\hat{y} = 0$

Non-linear Decision Boundaries

$$f_{\vec{w},b}(\vec{x}) = g(z) = g(w_1x_1 + w_2x_2$$
$$+ w_3x_1^2 + w_4x_1x_2 + w_5x_2^2$$
$$+ w_6x_1^3 + \cdots + b)$$





Cost Function for Logistic Regression

Squared error cost

cost

$$J(\vec{w},b) = \frac{1}{m}\sum_{i=1}^{m} \frac{1}{2}\left(f_{\vec{w},b}\left(\vec{x}^{(i)}\right) - y^{(i)}\right)^2$$

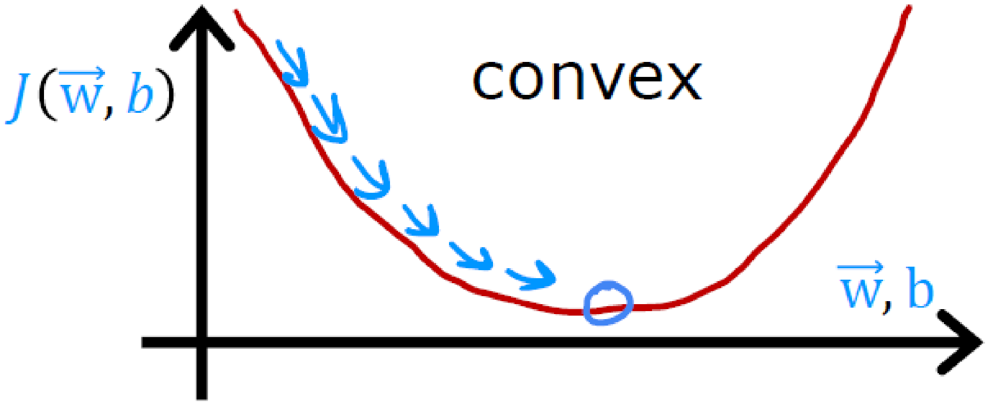
average of training set

loss $\boxed{L\left(f_{\vec{w},b}\left(\vec{x}^{(i)}\right), y^{(i)}\right)}$

Cost Function for Logistic Regression

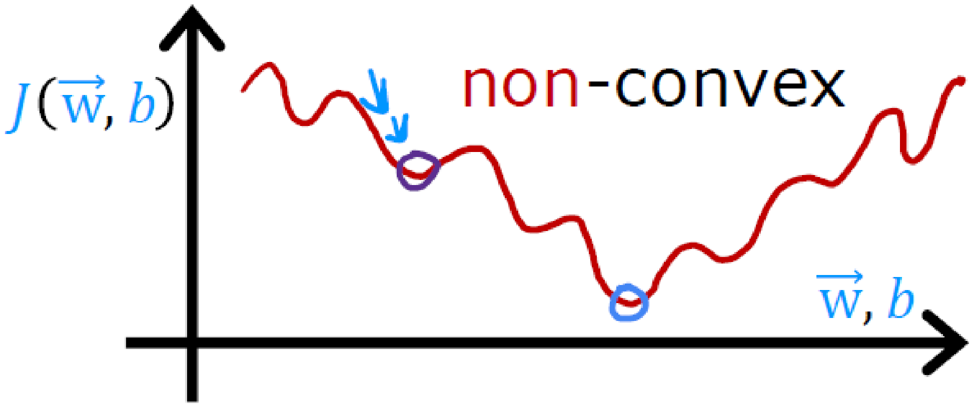
linear regression

$$f_{\vec{w},b}(\vec{x}) = \vec{w} \cdot \vec{x} + b$$



logistic regression

$$f_{\vec{w},b}(\vec{x}) = \frac{1}{1 + e^{-(\vec{w} \cdot \vec{x} + b)}}$$



Cost Function for Logistic Regression

Logistic loss function

$$L\left(f_{\vec{w},b}\left(\vec{x}^{(i)}\right), y^{(i)}\right) = \begin{cases} -\log\left(f_{\vec{w},b}\left(\vec{x}^{(i)}\right)\right) & \text{if } y^{(i)} = 1 \\ -\log\left(1 - f_{\vec{w},b}\left(\vec{x}^{(i)}\right)\right) & \text{if } y^{(i)} = 0 \end{cases}$$

Cost Function for Logistic Regression

cost

$$J(\vec{w}, b) = \frac{1}{m} \sum_{i=1}^{m} L\underbrace{\left(f_{\vec{w},b}\left(\vec{x}^{(i)}\right), y^{(i)}\right)}_{\text{loss}}$$

$$= \begin{cases} -\log\left(f_{\vec{w},b}\left(\vec{x}^{(i)}\right)\right) & \text{if } y^{(i)} = 1 \\ -\log\left(1 - f_{\vec{w},b}\left(\vec{x}^{(i)}\right)\right) & \text{if } y^{(i)} = 0 \end{cases}$$

convex

↳ can reach a global minimum

find w, b that minimize cost J

Cost Function for Logistic Regression

cost

$$J(\vec{w}, b) = \frac{1}{m} \sum_{i=1}^{m} L\underbrace{\left(f_{\vec{w},b}\left(\vec{x}^{(i)}\right), y^{(i)}\right)}_{\text{loss}}$$

$$= \begin{cases} -\log\left(f_{\vec{w},b}\left(\vec{x}^{(i)}\right)\right) & \text{if } y^{(i)} = 1 \\ -\log\left(1 - f_{\vec{w},b}\left(\vec{x}^{(i)}\right)\right) & \text{if } y^{(i)} = 0 \end{cases}$$

convex

↳ can reach a global minimum

find w, b that minimize cost J

Simplified Cost Function for Logistic Regression

loss

$$L\left(f_{\vec{\mathrm{w}},b}\left(\vec{\mathrm{x}}^{(i)}\right), y^{(i)}\right) = -\,y^{(i)}\log\left(f_{\vec{\mathrm{w}},b}\left(\vec{\mathrm{x}}^{(i)}\right)\right) - \left(1 - y^{(i)}\right)\log\left(1 - f_{\vec{\mathrm{w}},b}\left(\vec{\mathrm{x}}^{(i)}\right)\right)$$

Gradient Descent for Logistic Regression

repeat {

$j=1...n$

$$w_j = w_j - \alpha \frac{\partial}{\partial w_j} J(\vec{w}, b)$$

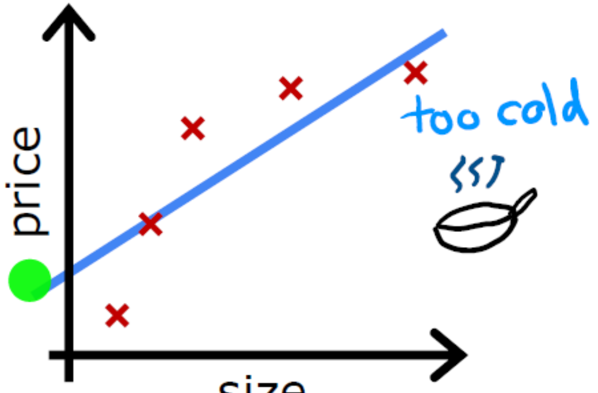
$$b = b - \alpha \frac{\partial}{\partial b} J(\vec{w}, b)$$

} simultaneous updates

$$\frac{\partial}{\partial w_j} J(\vec{w}, b) = \frac{1}{m} \sum_{i=1}^{m} \left(f_{\vec{w},b}\left(\vec{x}^{(i)}\right) - y^{(i)}\right) x_j^{(i)}$$

$$\frac{\partial}{\partial b} J(\vec{w}, b) = \frac{1}{m} \sum_{i=1}^{m} \left(f_{\vec{w},b}\left(\vec{x}^{(i)}\right) - y^{(i)}\right)$$

The Problem of Overfitting

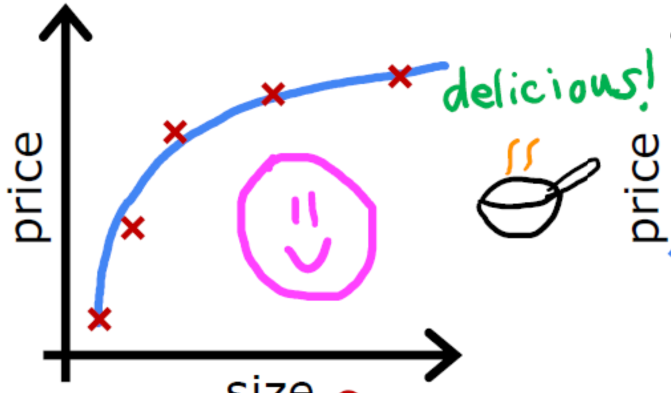


$w_1 x + b$

underfit

- Does not fit the training set well

high bias

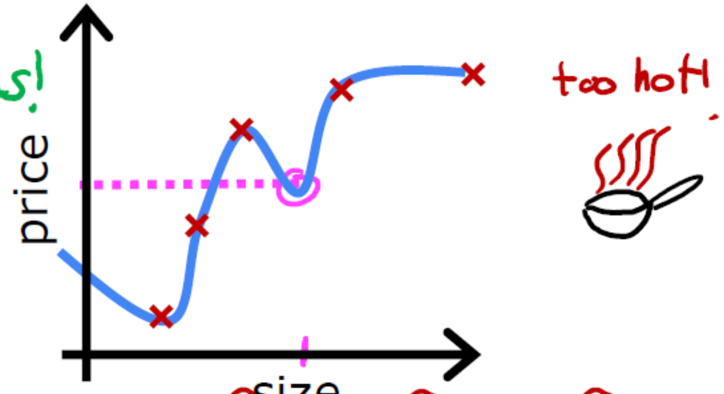


$w_1 x + w_2 x^2 + b$

just right

- Fits training set pretty well

generalization



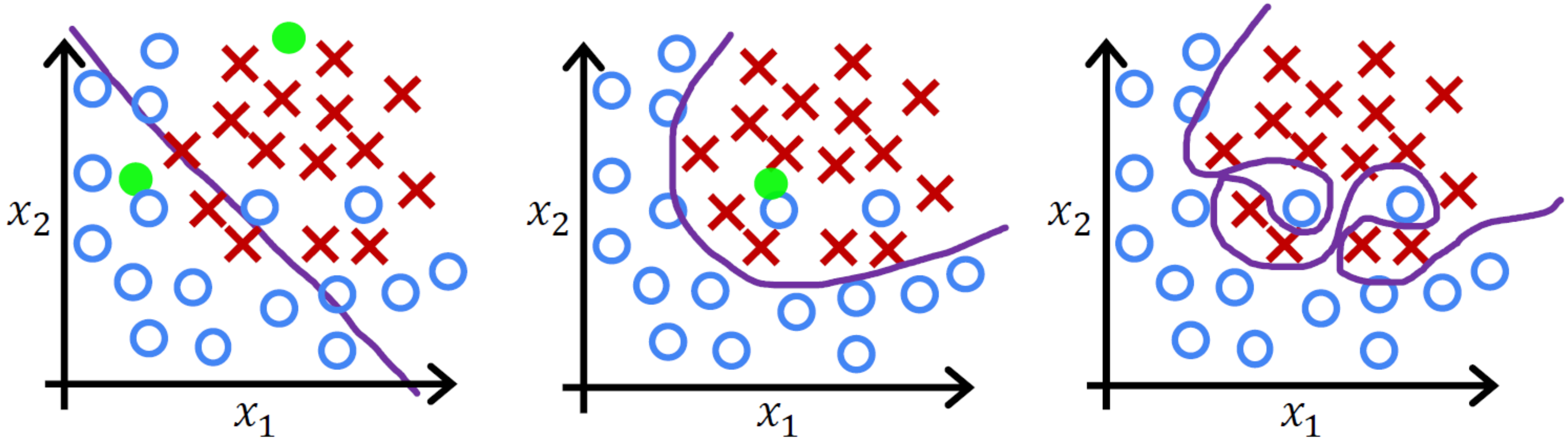
$w_1 x + w_2 x^2 + w_3 x^3 + w_4 x^4 + b$

overfit

- Fits the training set extremely well

high variance

The Problem of Overfitting



$$z = w_1 x_1 + w_2 x_2 + b$$

$$f_{\vec{w},b}(\vec{x}) = g(z)$$

g is the sigmoid function

underfit   high bias

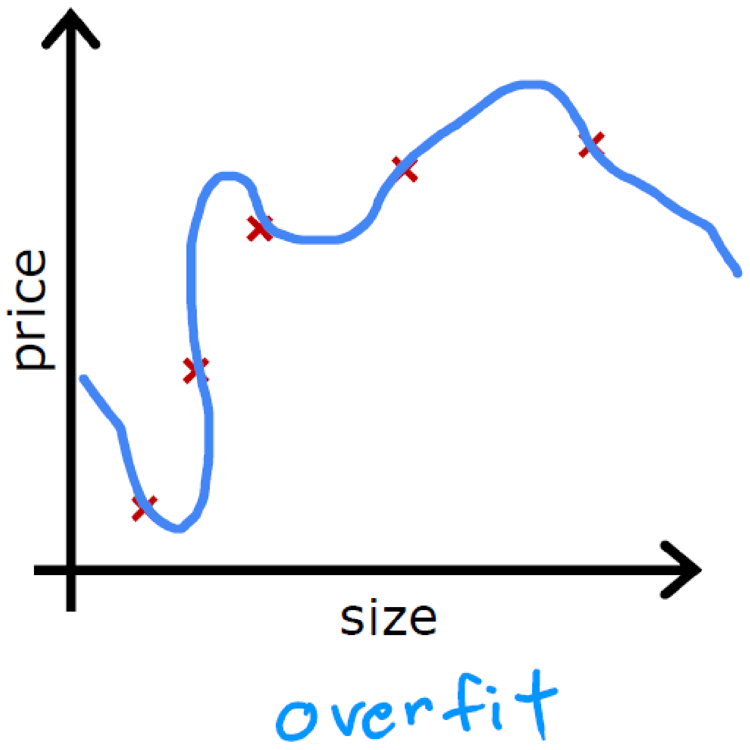
$$z = w_1 x_1 + w_2 x_2 + w_3 x_1^2 + w_4 x_2^2 + w_5 x_1 x_2 + b$$

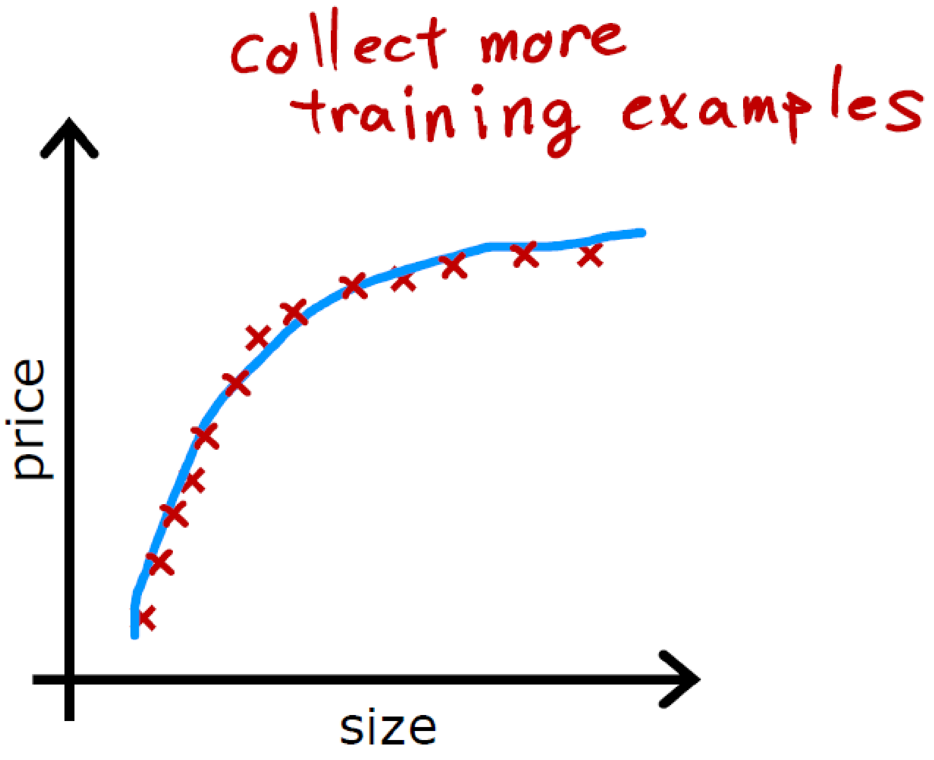
just right

$$z = w_1 x_1 + w_2 x_2 + w_3 x_1^2 x_2 + w_4 x_1^2 x_2^2 + w_5 x_1^2 x_2^3 + w_6 x_1^3 x_2 + \cdots + b$$

overfit

Addressing Overfitting





Addressing Overfitting

size	bedrooms	floors	age	avg income	...	distance to coffee shop	price
x_1	x_2	x_3	x_4	x_5		x_{100}	y

all features

+

insufficient data

↓

overfit

selected features

size
bedrooms
age
just right

feature selection

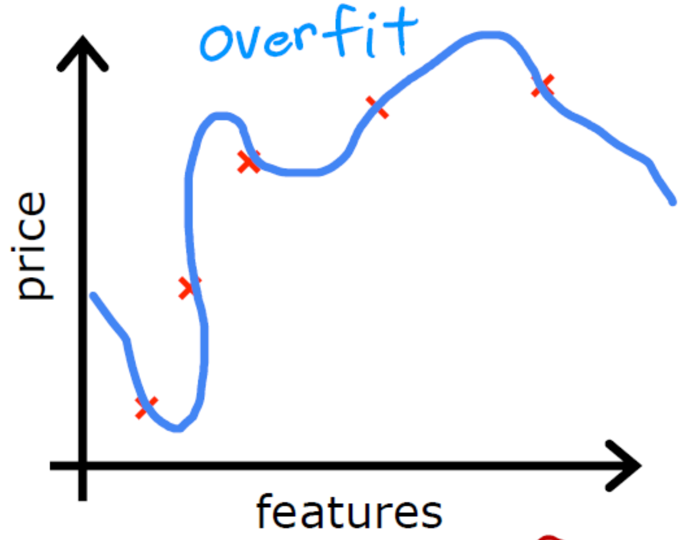
disadvantage

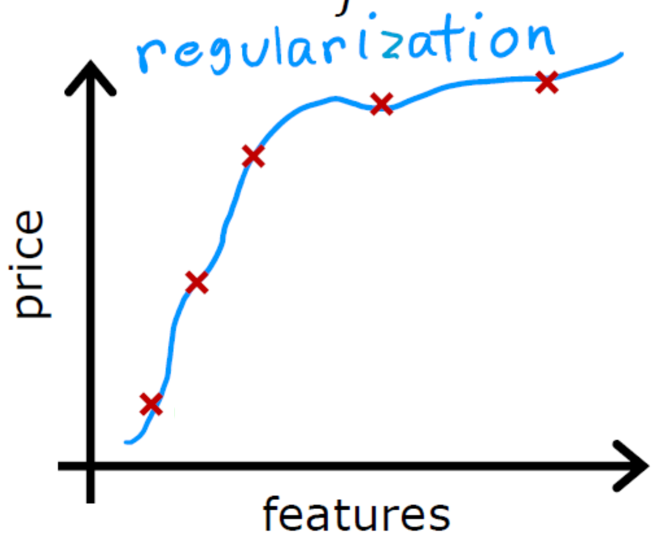
↓

useful features could be lost

Addressing Overfitting

Reduce the size of parameters w_j





$$f(x) = 28x - 385x^2 + 39x^3 - \underbrace{\cancel{174}x^4}_{0} + 100$$

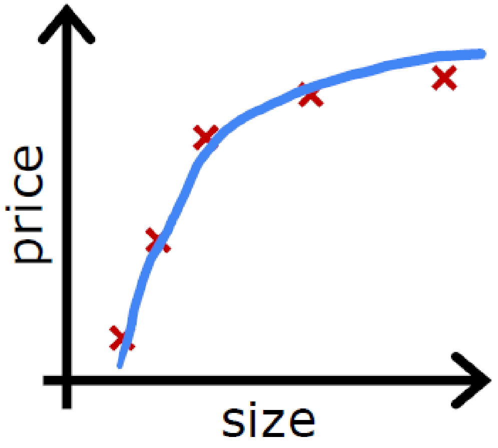
eliminate feature

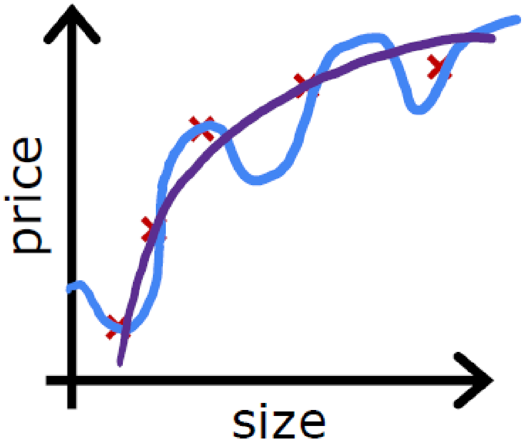
large values for w_j

$$f(x) = 13x - 0.23x^2 + \underbrace{0.000014}x^3 - \underbrace{0.0001}\,x^4 + 10$$

small values for w_j

Cost Function with Regularization





$w_1 x + w_2 x^2 + b$

$w_1 x + w_2 x^2 + \underbrace{\cancel{w_3 x^3}}_{\approx 0} + \underbrace{\cancel{w_4 x^4}}_{\approx 0} + b$

make w_3, w_4 really small (≈ 0)

$$\min_{\vec{w},b} \frac{1}{2m}\sum_{i=1}^{m}\left(f_{\vec{w},b}(\vec{x}^{(i)}) - y^{(i)}\right)^2 + \cancel{1000\underbrace{w_3^2}_{0.001}} + \cancel{1000\underbrace{w_4^2}_{0.002}}$$

Cost Function with Regularization

Regularization

small values $w_1, w_2, \cdots, w_n, b$

simpler model $w_3 \approx 0$

less likely to overfit $w_4 \approx 0$

size	bedrooms	floors	age	avg income	...	distance to coffee shop	price
x_1	x_2	x_3	x_4	x_5		x_{100}	y

n features

$n = 100$

$w_1, w_1, w_2, \cdots, w_{100}, b$

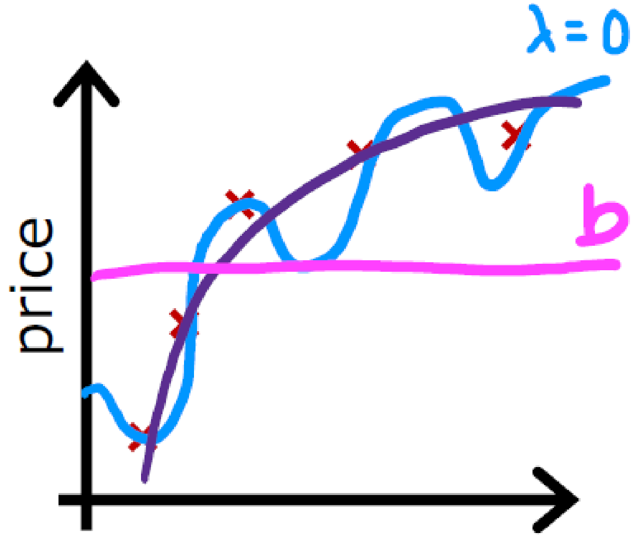
Cost Function with Regularization

$$J(\vec{\mathrm{w}}, b) = \frac{1}{2m}\left[\sum_{i=1}^{m}\left(f_{\vec{\mathrm{w}},b}\left(\vec{\mathrm{x}}^{(i)}\right) - y^{(i)}\right)^2 + \overbrace{\frac{\underbrace{\lambda}_{\text{"lambda"}}}{2m}\sum_{j=1}^{n} w_j^2}^{\text{regularization term}} + \overbrace{\frac{\lambda}{2m} b^2}^{\text{can include or exclude } b}\right]$$

"lambda"

regularization parameter $\lambda > 0$

Cost Function with Regularization



fit data

λ balances both goals

Keep w_j small

choose $\lambda = 10^{10}$

$$f_{\vec{\mathrm{w}},b}(\vec{\mathrm{x}}) = \underbrace{\cancel{w_1 x}}_{\approx 0} + \underbrace{\cancel{w_2 x^2}}_{\approx 0} + \underbrace{\cancel{w_3 x^3}}_{\approx 0} + \underbrace{\cancel{w_4 x^4}}_{\approx 0} + b$$

$f(x) = b$

Choose λ

Gradient Descent Implementation

repeat {

$$w_j = w_j - \alpha \left[\frac{1}{m} \sum_{i=1}^{m} \left[\left(f_{\vec{\mathrm{w}},b}\left(\vec{\mathrm{x}}^{(i)}\right) - y^{(i)} \right) x_j^{(i)} \right] + \frac{\lambda}{m} w_j \right]$$

$$b = b - \alpha \frac{1}{m} \sum_{i=1}^{m} \left(f_{\vec{\mathrm{w}},b}\left(\vec{\mathrm{x}}^{(i)}\right) - y^{(i)} \right)$$

} simultaneous update $j = 1 \ldots n$

THANK YOU

NEXT LECTURE WILL BE IN-PERSON ON MON, 22.5.2023, IN SHAA ALLAH!

SABBAGH@IEEE.ORG

CONNECT ON LINKEDIN