Machine Learning

By - Andrew NG

Tue 12 Jul

Week 3

Classification :

	answer "y"	
Is this spam	no	yes
tumour malignant	"	"

y can be 2 values

"binary classification"

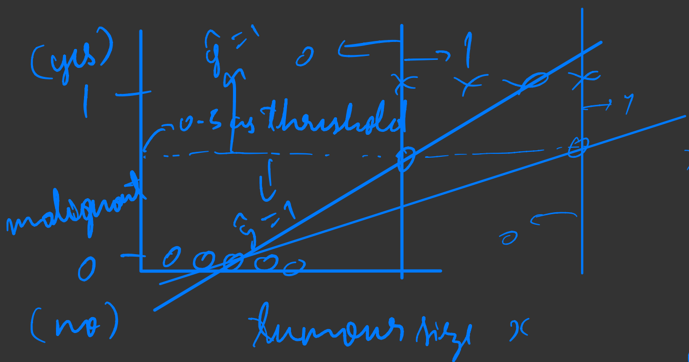
Class = category

absence
"negative class"

0 → false

1 → true

"positive class"
presence



$f_{w,b}(x) = wx + b$

× → added

↳ makes worse {misclassified}

↳ That's why linear regression is not best for classification.

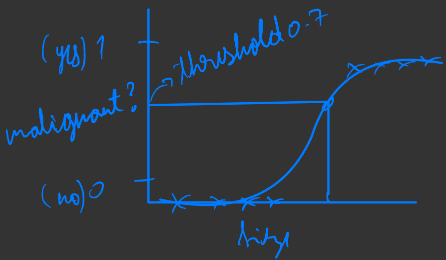
If $f_{w,b}(x) < 0.5 \rightarrow \hat{y} = 0$

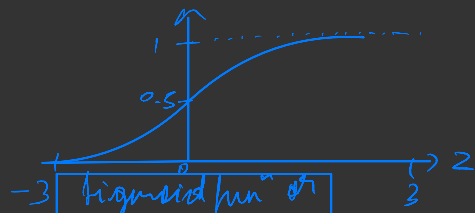
else $\hat{y} = 1$

next : Logistic regression

↳ classification

Logistic regression:





Sigmoid fun^n or logistic fun^n.

outputs b/w 0 & 1

$0 < g(z) < 1$

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

⇒ let z be big no. eg 100

⇒ $g(z) \approx \frac{1}{1}$

Similarly $z = -100$

$g(z) = \frac{1}{big} \approx 0$

$z = 0$ $g(z) = 0.5$

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

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

↓

z

↓

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

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

"Logistic regression"

Interpretation of logistic regression output

$$\boxed{f_{\vec{w},b}(\vec{x}) = \frac{1}{1+e^{-(\vec{w}\cdot\vec{x}+b)}}} \rightarrow \text{"Probability" that class is 1}$$

Eg x is "tumour size"

y is 0

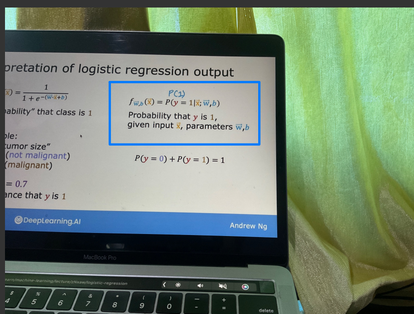
or 1

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

then 70% chance that y is 1

or 30% " " y is 0

$$\Rightarrow \boxed{P(y=0) + P(y=1) = 1}$$



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

0 or 1

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

yes $\hat{y}=1$ no $\hat{y}=0$

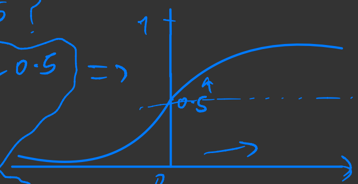
when is $f_{\vec{w},b}(\vec{x}) \geq 0.5$?

$g(z) \geq 0.5 \Rightarrow$

when $z \geq 0$

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

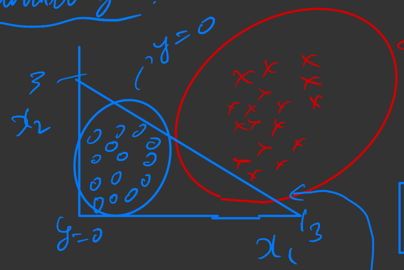
$\hat{y}=1$



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

$\hookrightarrow \hat{y}=0$

Decision Boundary :



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

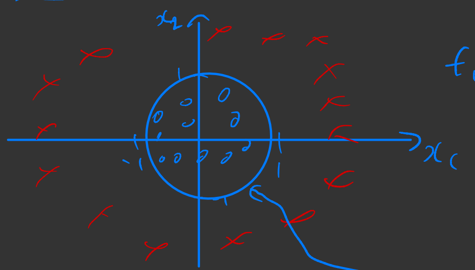
$$\boxed{z = \vec{w}\cdot\vec{x} + b = 0}$$

Decision boundary

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

$$\Rightarrow x_1 + x_2 = 3$$

Non-linear Decision boundary :



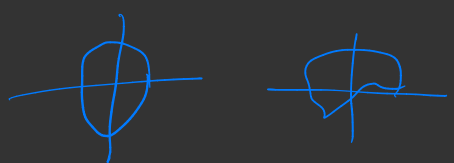
$$f_{\vec{w},b}(\vec{x}) = g(z) = g(\overbrace{\underset{1}{w_1}x_1^2 + \underset{1}{w_2}x_2^2 + \underset{-1}{b}}^{z})$$

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

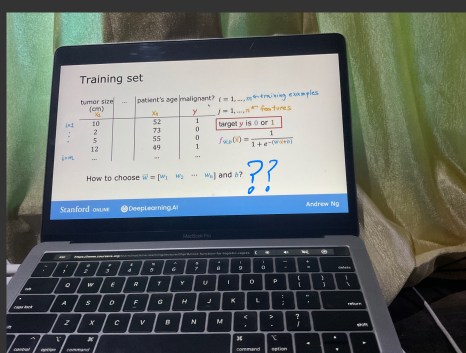
$$x_1^2 + x_2^2 = 1 \to \text{circle}$$

Let $z = w_1x_1 + w_2x_2 + w_3x_1^2 + w_4x_1x_2 + w_5x_2^2 + w_6x_1^3 + \dots + b$

Ellipse



.... so on



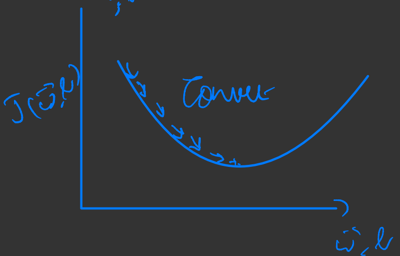
Squared error cost:

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

→ Loss → $L(f_{\vec{w},b}(x^{(i)}), y^{(i)})$

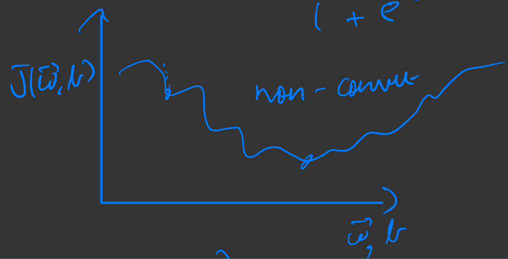
Linear regression:

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

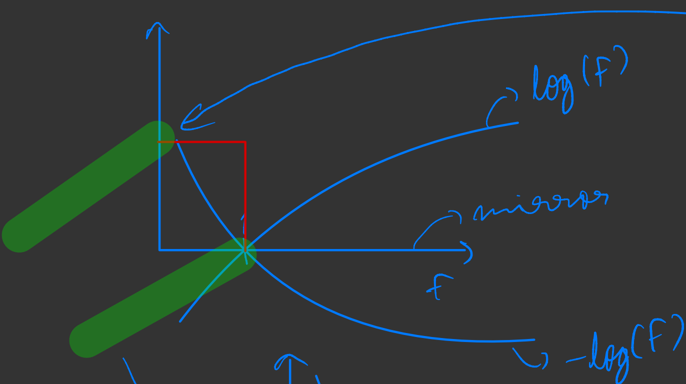


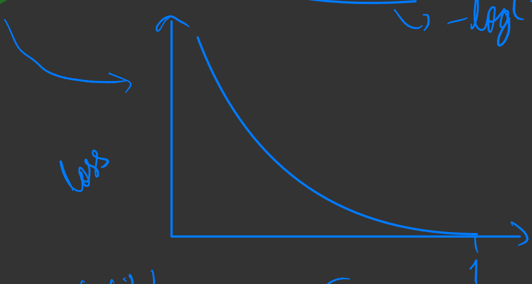
Logistic regression:

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



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



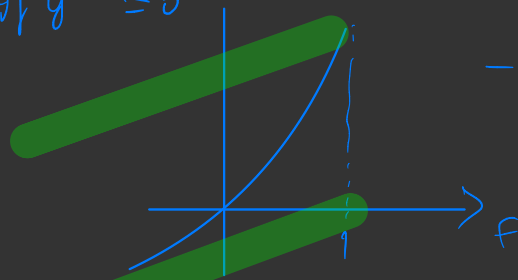


as $f_{\vec{w},b}(x^{(i)}) \to 0$
then loss $\to \infty$:(

If $y^{(i)} = 1$

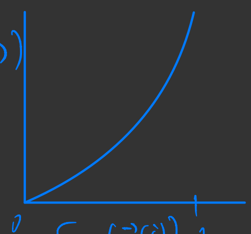
as $f_{\vec{w},b}(\vec{x}^{(i)}) \to 1$
then loss $\to 0$:)

If $y^{(i)} = 0$



$-\log(1-f)$

$L(f_{\vec{w},b}(\vec{x}^{(i)}, y^{(i)})$ if $y^{(i)}=0$



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

as $f_{\vec{w},b}(\vec{x}^{(i)}) \to 0$ then loss $\to 0$:)

as $f_{\vec{w},b}(\vec{x}^{(i)}) \to 1$ then loss $\to \infty$:(

$\Rightarrow$ Cost for logistic regression

is $J(\vec{w}, b) = \frac{1}{m} \sum_{i=1}^{m} L(f_{\vec{w},b}(\vec{x}^{(i)}), y^{(i)})$

can reach global min.

convex

Loss

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

find w, b to minimize cost J.

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

$$\boxed{J_{\vec{w},b} = \frac{1}{m} \sum_{i=1}^{m} L(f_{\vec{w},b}(\vec{x}^{(i)}), y^{(i)})}$$

$$J_{\vec{w},b} = -\frac{1}{m}\sum_{i=1}^{m}\left[y^{(i)}\log\left(f_{\vec{w},b}(\vec{x}^{(i)})\right) + (1-y^{(i)})\cdot\log\left(1-f_{\vec{w},b}(\vec{x}^{(i)})\right)\right]$$

Cost function

maximum likelihood.

Gradient Descent implementation:

Find $\vec{w}, b$

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

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

$j = 1 \ldots n$ features

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

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

} simultaneous updates.

repeat {

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

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

} simultaneous updates

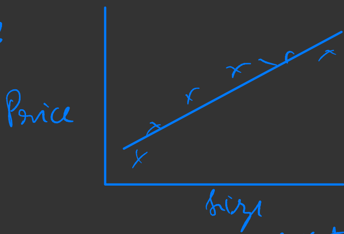
Here $f_{\vec{w},b}(\vec{x}^{(i)}) = \frac{1}{1+e^{-(\vec{w}\cdot\vec{x}+b)}}$

Same concept: learning curve

- Vectorized implementation
- Feature scalling

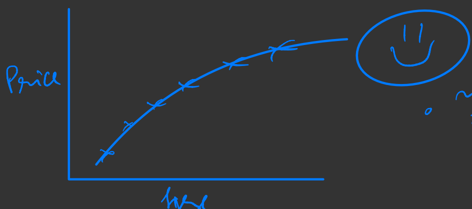
Overfitting {avoided by regularization}

Linear regression eg:



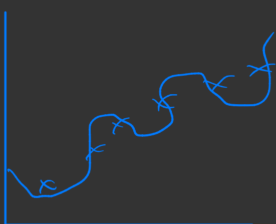
$w_1 x + b$

- Underfit or high bias {Doesn't fit the training set well}



$w_1 x + w_2 x^2 + b$

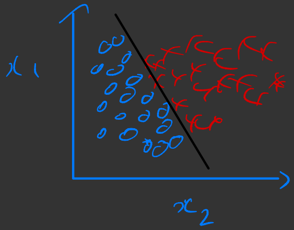
- Fits training set pretty well
  generalization



- Fits the training set extremely well
  overfit or high variance.

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

In Classification:

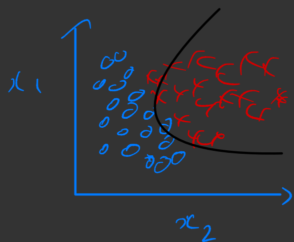


$$z = w_1 x_1 + w_2 x_2 + b$$

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

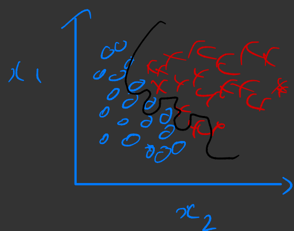
g is sigmoid fun^n.

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!

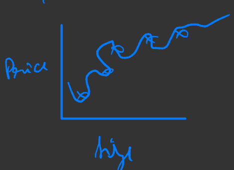


$$z = w_1 x_1 + w_2 x_2 + w_3 x_1^2 x_2 + w_4 x_1^2 x_2^2 + w_5 x_1^2 x_2^3 + \dots + b$$

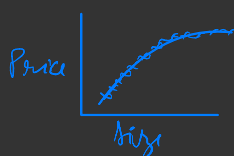
Overfitting or high variance

Addressing overfitting:

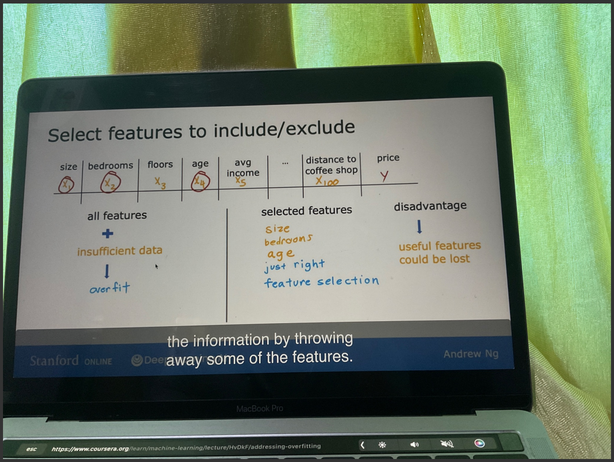
1) Collect more training examples



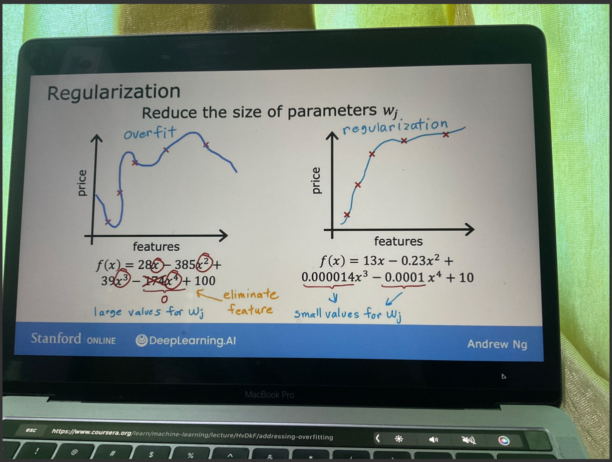
overfit



2) Select features to include/Exclude – Feature selection.



3) Regularization: i.e reduce the size of parameters w_j



Cost function with regularization:

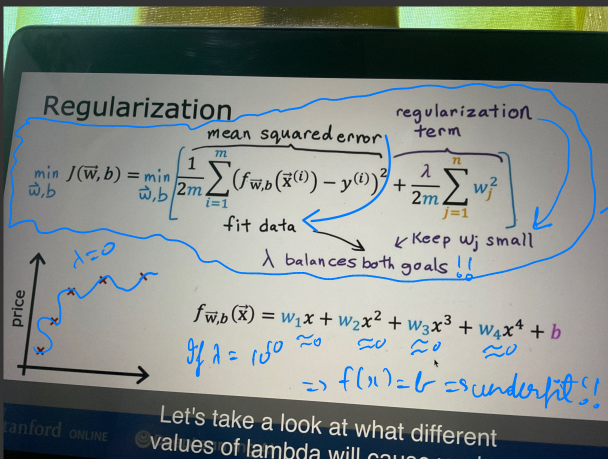
$$\min . \frac{1}{2m}\sum_{i=1}^{m}\left(f_{\vec{w},b}(\vec{x}^{(i)})-y^{(i)}\right)^2$$

w_j is reduced from $j=1$ to n.

regularization parameter

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

"lambda" (λ)
↳ regularization parameter $\lambda > 0$



→ regularized cost fun^n.

$\therefore$ Choose λ wisely!
as $\lambda\uparrow$ $w_j\downarrow$

$$\min_{\vec{w},b} J(\vec{w},b)=\min_{\vec{w},b}\left[\frac{1}{2m}\sum_{i=1}^{m}\left(f_{\vec{w},b}(\vec{x}^{(i)})-y^{(i)}\right)^2+\frac{\lambda}{2m}\sum_{j=1}^{n} w_j^2\right]$$

Gradient Descent

repeat {

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

↳ no $\sum_{j=1}^{n}$

$j=1,\ldots,n$

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

don't have to regularize b

} simultaneous update

Implement gradient descent:

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

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

} simultaneous update. $j = 1 \ldots n$

$$w_j = \underbrace{w_j\left(1 - \alpha\frac{\lambda}{m}\right)}_{\text{shrink } w_j} - \underbrace{\alpha\frac{1}{m}\sum_{i=1}^{m}\left(f_{\vec{w},b}(\vec{x}^{(i)}) - y^{(i)}\right)x_j^{(i)}}_{\text{usual update}}$$

let $\alpha = 0.01, \lambda = 1, m = 50$

$\Rightarrow 1 - \alpha\frac{\lambda}{m} = 0.9998$

Regularized logistic regression:

$$\min_{\vec{w},b} J(\vec{w},b) = -\frac{1}{m}\sum_{i=1}^{m}\left[y^{(i)}\log\left(f_{\vec{w},b}(\vec{x}^{(i)})\right) + (1-y^{(i)})\log\left(1-f_{\vec{w},b}(\vec{x}^{(i)})\right)\right] + \frac{\lambda}{2m}\sum_{j=1}^{n}w_j^2$$

$\min_{\vec{w},b} J(\vec{w},b) \rightarrow w_j \downarrow$