Classification

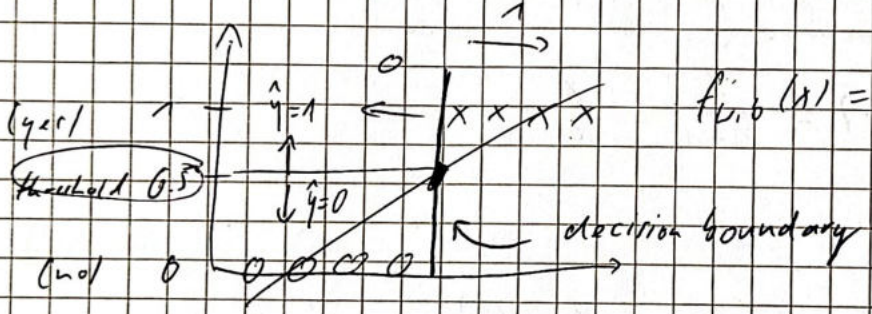
Question	Answer "y"
Is this email spam?	no / yes
Is the transaction fraudulent?	no / yes
Is the tumor malignant	no / yes (or false/true or 1/0)

→ y can only be one of two values

→ binary classification (class = category)

1 = positive class

0 = negative class



Here we want to predict categories!

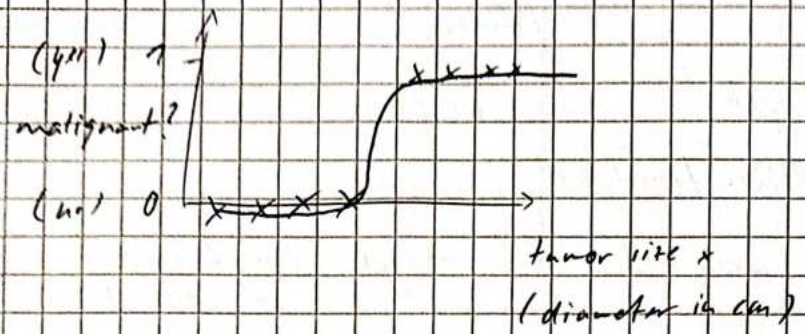
If $f_{w,b}(x) < 0.5 \rightarrow \hat{y} = 0$

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

But this is worse: misclassified

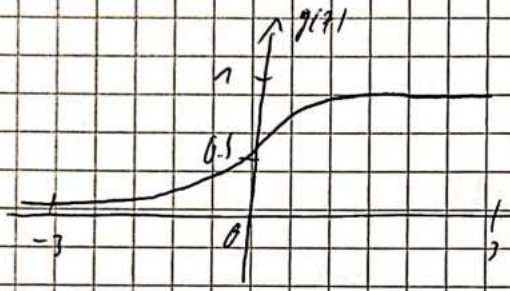
(18)

Logistic regression



Want output between 0 and 1!

Sigmoid Function / Logistic Function:



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

- $g(z=0) = \frac{1}{1+1} = \frac{1}{2}$

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

Model for logistic regression: $z := \vec{w} \cdot \vec{x} + b$

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

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

Interpretation of logistic regression output:

Example:

x = "tumor size"

$y = 0$ (not malignant)

$= 1$ (malignant)

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

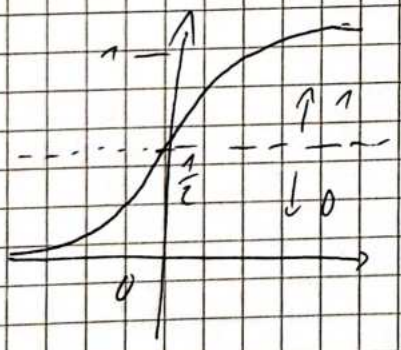
→ 70 % chance that y is 1

Notation:

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

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

Decision Boundary



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

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

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

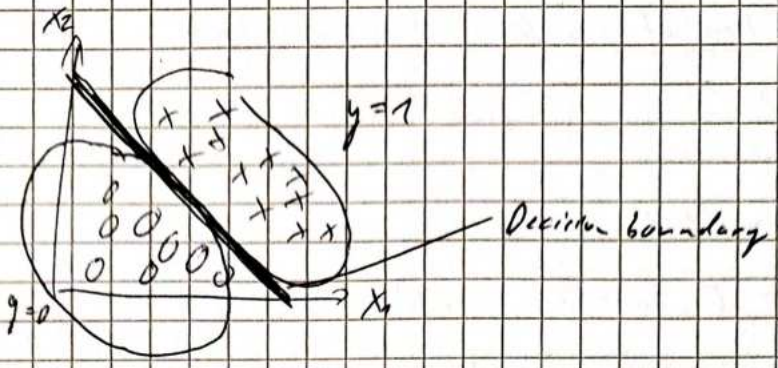
$g(z) \geq 0.5$

$z \geq 0$

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

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

(20)

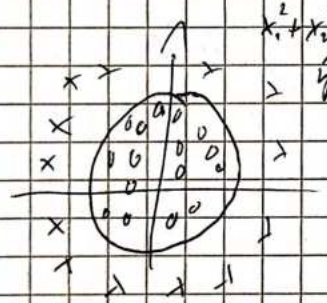


Decision boundary: $z = \vec{w} \cdot \vec{x} + b = 0$

$z = x_1 + x_2 - 3 = 0$

$\Leftrightarrow$ $x_1 + x_2 = 3$

Non linear Decision Boundaries:



$f_{w,b}(\vec{x}) = g(z) = g(\underbrace{w_1 x_1^2 + w_2 x_2^2 + b}_{=z})$

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

$x_1^2 + x_2^2 = 1$

$x_1^2 + x_2^2 = 1$

$\hat{y}=0$

Cost Function for Logistic Regression

Training set

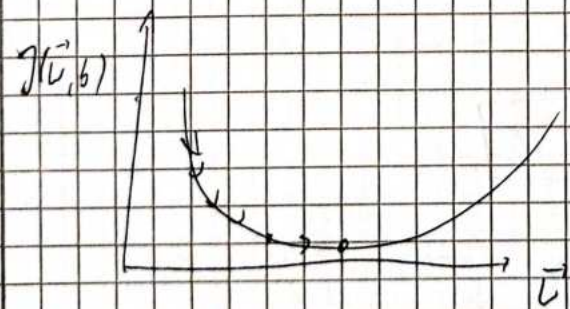
	x_1 tumor size (cm)	x_2 patient age	y malignant
$i=1$	10	52	1
	2	73	0
	5	55	0
	12	49	1
$i=m$	1	1	

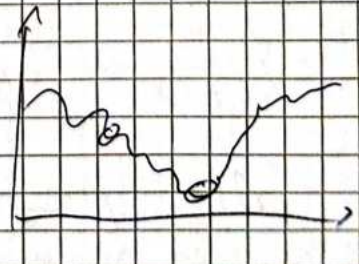
- $i = 1, \dots, m$ training examples
- $j = 1, \dots, n$ features
- target y is 0 or 1

↳ How to choose $\vec{w} = [w_1, w_2, \dots, w_n]$ and b?

Squared error cost

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





Linear regression

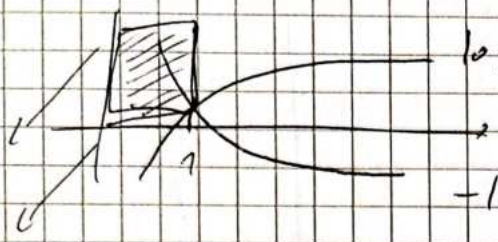
for $f_{\vec{w},b}(\vec{x}) = \vec{w} \cdot \vec{x} + b$

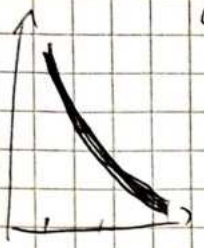
Logistic Regression

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

Logistic loss function

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



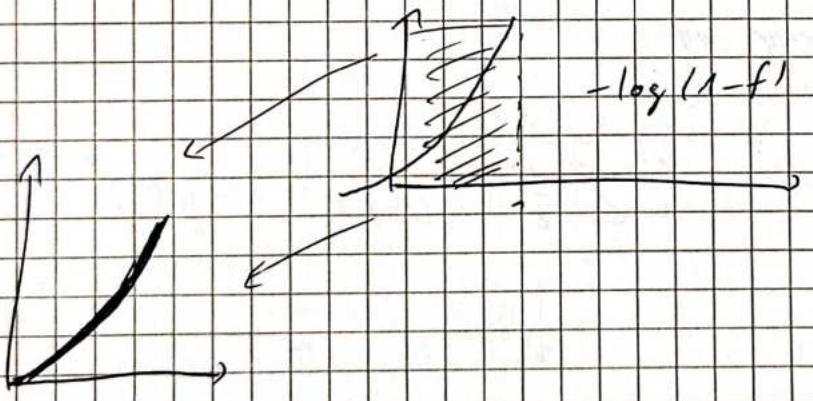


Loss is lowest when $f_{\vec{w},b}(\vec{x}^{(i)})$ predicts close to true label $y^{(i)}$.

$L(f(x^{(i)}), y^{(i)})$ if $\boxed{y^{(i)}=1}$

As $f \to 1$ then loss $\to 0$ ✓

As $f \to 0$ then loss $\to \infty$ ↯



$L(f(x^{(i)}), y^{(i)})$ if $\boxed{y^{(i)}=0}$

As $f \to 1$ then loss $\to \infty$ ↯

As $f \to 0$ then loss $\to 0$ ✓

Simplified Cost Function

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

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

($y^{(i)}$ multiplies the first term, $=1$; $(1-y^{(i)})$ multiplies the second term, $=0$)

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

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

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

$$L\left(f_{\vec{w},b}(\vec{x}^{(i)}), y^{(i)}\right) = -(1-0)\log\left(1-f(\vec{x})\right)$$

Cost Function:

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

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

This particular Cost Function is derived by the Maximum Likelihood

(24)

Gradient Descent Implementation

Find $\vec{\omega}, b$

Given new $\vec{x}$, output $f_{\vec{\omega},b}(\vec{x}) = \frac{1}{1+e^{-(\vec{\omega}\cdot\vec{x}+b)}}$

$$P(y=1 \mid \vec{x}; \vec{\omega}, b)$$

For the Cost-Function $J(\vec{\omega}, b)$

→

repeat {

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

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

}

with

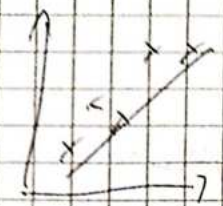
$$\frac{\partial}{\partial \omega_j} J(\vec{\omega}, b) = \frac{1}{m} \sum_{i=1}^{n} \left(f_{\vec{\omega},b}(\vec{x}^{(i)}) - y^{(i)}\right) x_j^{(i)}$$

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

→ Looks like linear regression, but f is different !!!

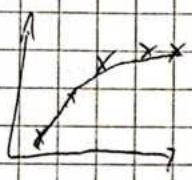
Here: $f = \frac{1}{1+e^{-(\vec{\omega}\vec{x}+b)}}$

The Problem of Overfitting



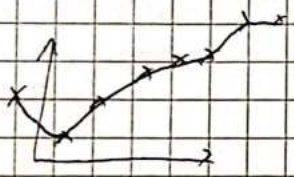
→ Underfit

- Does not fit the training set well
- High bias



- Fits training set pretty well

→ Generalization



→ Overfit

- Fits the training set extremely well
- High variance

Feature Selection

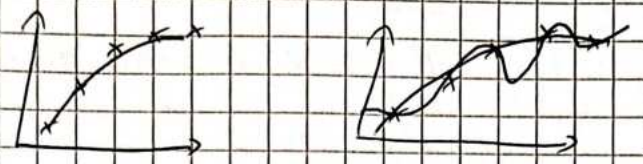
All features + insufficient data ⇒ Overfit

Selected features ⇒ Disadvantage

↳ useful features could be lost

(26) Regularization

Intuition:



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

$\approx 0 \quad \approx 0$

make w_3, w_4 really small

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

General:

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

λ = Regularization Parameter ($\lambda > 0$)

Goal:

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

Regularized Linear Regression:

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

Gradient Descent:

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

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

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

Regularized Logistic Regression

Cost: $J(\vec{w}, b) = -\frac{1}{m} \sum \left[y^{(i)} \log\left(f_{\vec{w},b}(\vec{x}^{(i)}) \right) + (1 - y^{(i)}) \log\left(1 - f_{\vec{w},b}(\vec{x}^{(i)}) \right) \right]$

$$+ \frac{1}{2m} \sum_{j=1}^{n} w_j^2$$

$$\min_{\vec{w}, b} J(\vec{w}, b)!$$

Gradient Descent:

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

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