Linear Regression

Single feature

(x) hours	(y) marks

m, c ?

$y = \theta_1 x + \theta_0$ → bias

weight θ_1

θ_1, θ_0 ?

$J(\theta)$ Loss = $\frac{1}{m}\sum_{i=1}^{m}(\hat{y}-y)^2$ (MSE)

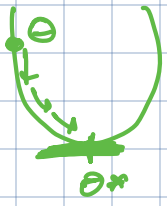
$h_\theta(x) = \theta_1 x + \theta_0$

→ Random θ_0, θ_1

→ How good θ_0, θ_1 is ?

→ Update θ_0, θ_1

GRADIENT DESCENT



$\theta = \theta - \eta \frac{\partial J(\theta)}{\partial \theta}$ $\nabla J(\theta)$

$\theta_0 = \theta_0 - \eta \cdot \frac{2}{m}\sum_{i=1}^{m}(\theta_1 x + \theta_0 - y)$

$\theta_0 = \theta_1 - \eta \cdot \frac{2}{m}\sum_{i=1}^{m}(\theta_1 x + \theta_0 - y)x^i$

Multiple features

$$\begin{bmatrix} — & x^1 & — \\ — & x^2 & — \end{bmatrix} \rightarrow \begin{bmatrix} x_1^1 & x_2^1 & x_3^1 & \cdots & x_n^1 \\ x_1^2 & x_2^2 & x_3^2 & \cdots & x_n^2 \end{bmatrix}$$

x^1

x^m

$x_1^m \quad x_2^m \quad x_3^m \quad \dots \quad x_n^m$

x_1	x_2	x_3	x_4	y

→ Regression

$$h_\theta(x) = y = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3 + \theta_4 x_4$$

θ_0 → bias

$\theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3 + \theta_4 x_4$ → weights

$$h_\theta(x) = \theta_0 + \sum_{i=1}^{n} \theta_i x_i$$

$$h_\theta(x) = \theta_0 x_0 + \sum_{i=1}^{n} \theta_i x_i \qquad x_0 = 1$$

$$h_\theta(x) = \sum_{i=0}^{n} \theta_i x_i \longrightarrow \text{feature 1}$$

$\theta_0 x_0 + \theta_1 x_1$

1

$\theta_0 + \theta_1 x_1$

} single feature

x_0	x_1	x_2	x_3	x_4	y
1					

$$\theta = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \theta_1 \\ \vdots \\ \theta_n \end{bmatrix} \qquad x = \begin{bmatrix} 1 \\ x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

$\theta^T x$

$$[\theta_0\ \theta_1\ \theta_2\ \cdots\ \theta_n] \begin{bmatrix} 1 \\ x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \theta_0 + \theta_1 x_1 + \theta_2 x_2 \cdots \theta_n x_n$$

$$\hat{y} = h_\theta(x) = \sum_{i=0}^{n} \theta_i x_i = \theta^T x$$

→ Hypothesis
→ Error / loss
→ Gradient

Loss fn:

MSE $$J(\theta) = \frac{1}{m} \sum_{i=1}^{m} \left(\hat{y}^{(i)} - y^{(i)}\right)^2$$

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

TASK: $\frac{\partial J(\theta)}{\partial \theta} \overset{?}{=} 0$

$\nabla_\theta J(\theta)$

$$\frac{\partial}{\partial \theta_j} J(\theta) = \frac{\partial}{\partial \theta_j} \frac{1}{m} \sum_{i=1}^{m} \left(\hat{y}^{(i)} - y^{(i)}\right)^2$$

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

$$= \frac{1}{m} \sum_{i=1}^{m} \frac{\partial}{\partial \theta_j} \left(\theta_0 x_0^{(i)} + \theta_1 x_1^{(i)} + \cdots \theta_j x_j^{(i)} + \cdots \theta_n x_n^{(i)} - y^{(i)}\right)^2$$

$$= \frac{1}{m} \sum_{i=1}^{m} 2\left(\theta_0 x_0^{(i)} + \theta_1 x_1^{(i)} + \cdots \theta_n x_n^{(i)} - y^{(i)}\right) x_j^{(i)}$$

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

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

Final Gradient Update Rule

$$\theta_j = \theta_j - \eta \cdot \frac{1}{m} \sum_{i=1}^{m} \left(\hat{y}^{(i)} - y^{(i)}\right) x_j^{(i)}$$

$$\hat{y}^{(i)} = h_\theta(x^{(i)}) = \sum_{i=0}^{n} \theta_i x_i = \theta^T \cdot x$$

$[\theta_0 \ \theta_1 \ \ldots \ \theta_n]$ Randomly

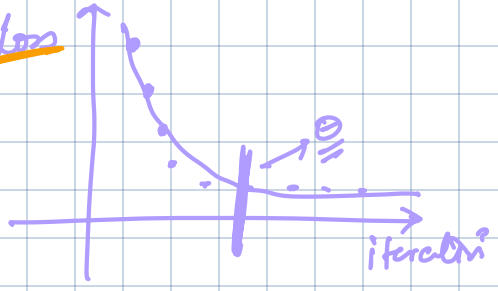
do
{

loss fxn (MSE) → how good ur θ's are

update θ

}
while (converge)

① 2000 times



Interpolation

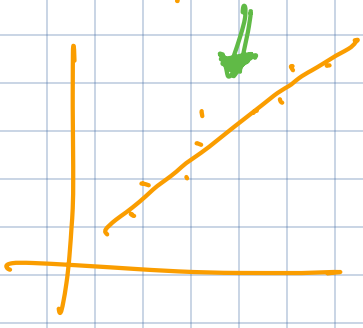
OVERFITTING

- not generalizable for test data points
- has info only about training data points.

ML Algo

- Supervised (y given)
- Unsupervised
- Reinforcement

Supervised:
- Regression (y continuous) (Linear Regression)
- Classification (y discrete) (Logistic Regression)



Logistic Regression

↳ Classification Algo

Eg:

→ wt, ht → Dog, cat

→ Spam or not Spam

→ Animal → Cat / Dog

} Binary Classification

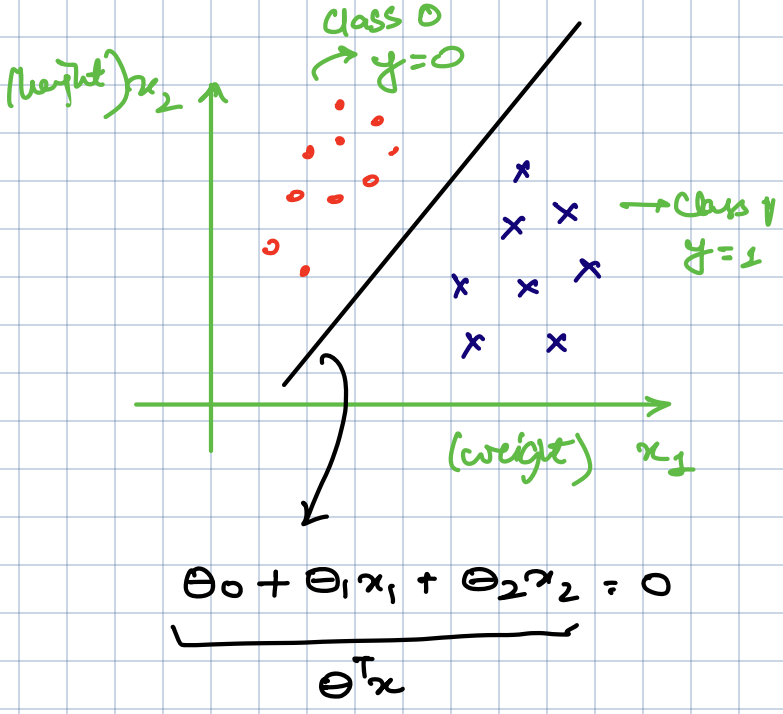
Training Data:

$\{x^{(i)}, y^{(i)}\}$

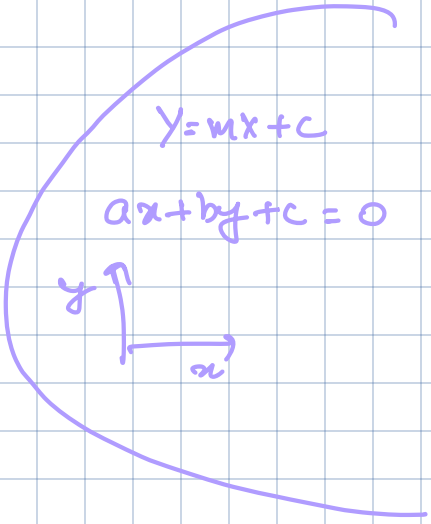
$x \in R^n$ → x has n features & all features are real nos.

Classification:
y should be a discrete value.

$y \in \{0, 1\}$ → 0: Cat, 1: Dog



x: cat : 1
o : dog : 0

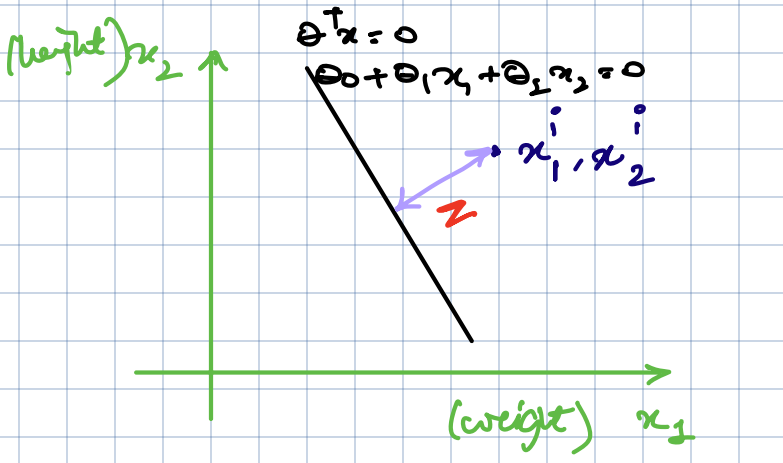


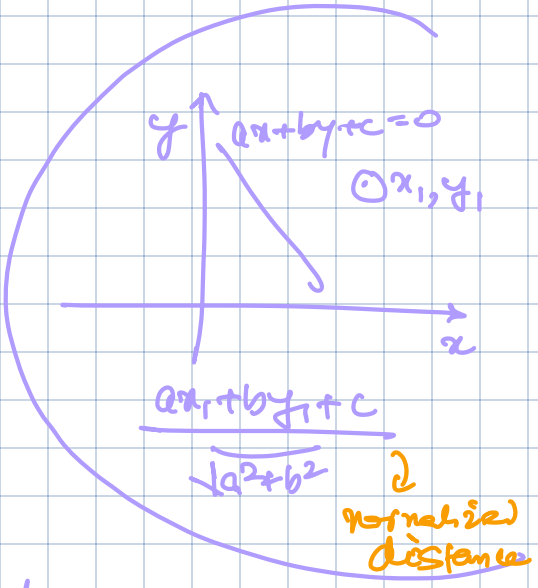
$\theta_0 + \theta_1 x_1 + \theta_2 x_2 = 0$

$\theta^T x$

$\theta = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \end{bmatrix}$ $x = \begin{bmatrix} x_0 \\ x_1 \\ x_2 \end{bmatrix}$ $\theta^T x = [\theta_0\ \theta_1\ \theta_2] \begin{bmatrix} 1 \\ x_1 \\ x_2 \end{bmatrix}$ $x_0 = 1$

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



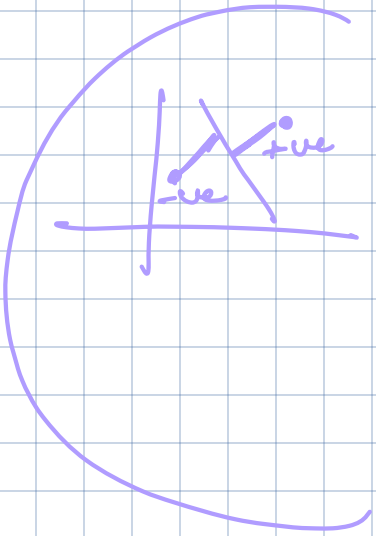


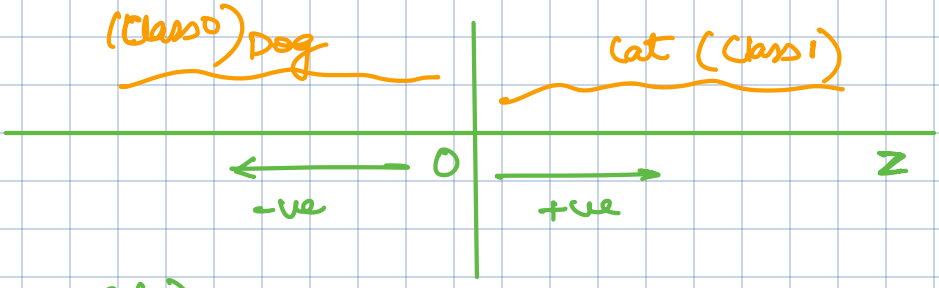
$\theta_0 + \theta_1 x_1' + \theta_2 x_2'$ → unnormalized distance b/w x^i & line $\theta_0 + \theta_1 x_1 + \theta_2 x_2$

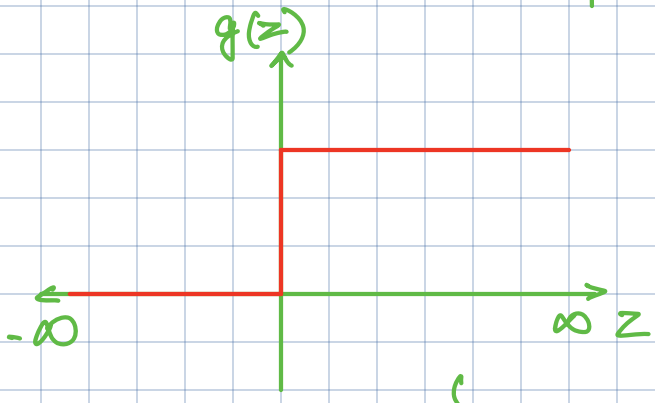
$$[\theta_0 \ \theta_1 \ \theta_2] \begin{bmatrix} 1 \\ x_1^{(i)} \\ x_2 \end{bmatrix} = \theta^T x^{(i)}$$

θ^T x^i

z



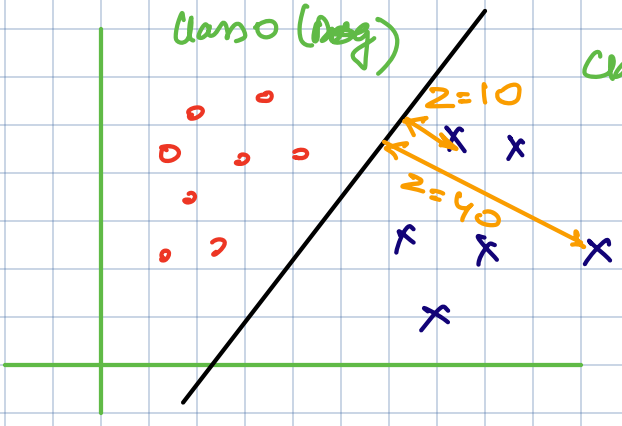




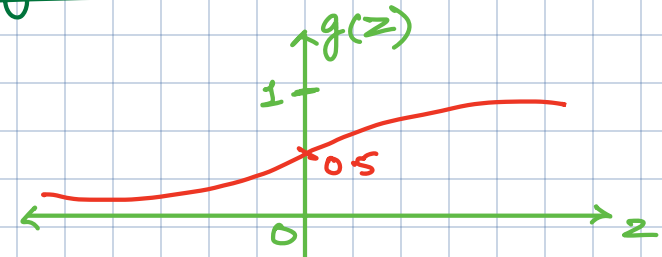
$g(z) = 1$ if $z \geq 0$

$g(z) = 0$ if $z < 0$

PROBLEM? Not able to distinguish b/w points which are close to line & which are far.



Sigmoid function



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

sigmoid fn

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

$z = 0$ $\quad g(z) = \frac{1}{1 + \frac{1}{e^{0}}} = \frac{1}{2} = 0.5$ ($e^{0} = 1$)

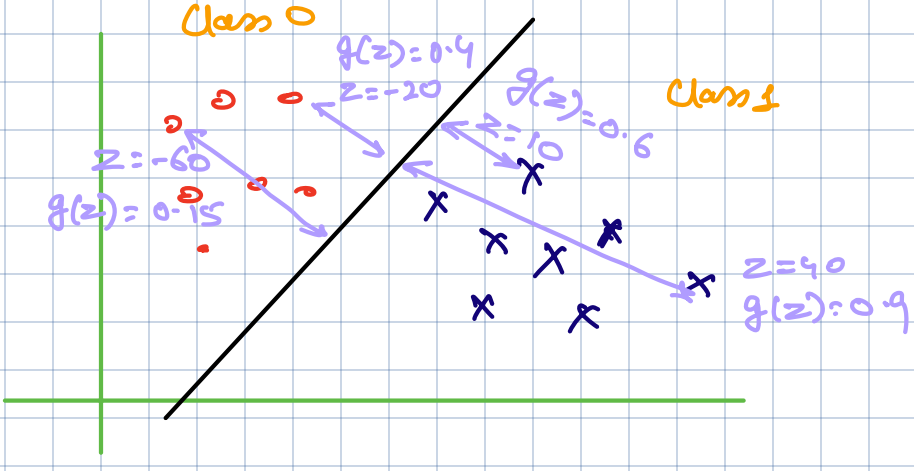
$z = -\infty$ $\quad g(z) = \frac{1}{1 + \frac{1}{e^{-\infty}}} = \frac{1}{1 + e^{\infty}} = 0$

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

$$g(z) = \frac{1}{1 + e^{-z}} \quad \text{where} \quad z = \theta^T x$$

value b/w 0 & 1

Probability / Confidence with which you can say the point belongs to class 1



$g(z) = 0.4$

40% sure points belongs to class 1

60% sure point belongs to class 0

point lies on the line $g(z) = 0.5$

50% sure belongs to Class 0
Class 1

$$h_\theta(x) = g(z) = g(\theta^T x) = \frac{1}{1+e^{-\theta^T x}}$$ Prob. that test point ∈ Class 1

$$\hat{y} = 1 \quad \text{if} \quad h_\theta(x) \geq 0.5$$
$$= 0 \quad \text{if} \quad h_\theta(x) < 0.5$$

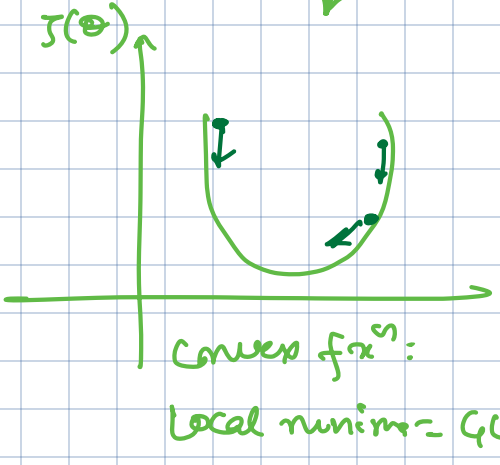
Loss fx^n:

$$MSE = \frac{1}{m}\sum_{i=1}^{m} (\hat{y}^{(i)} - y^{(i)})^2$$

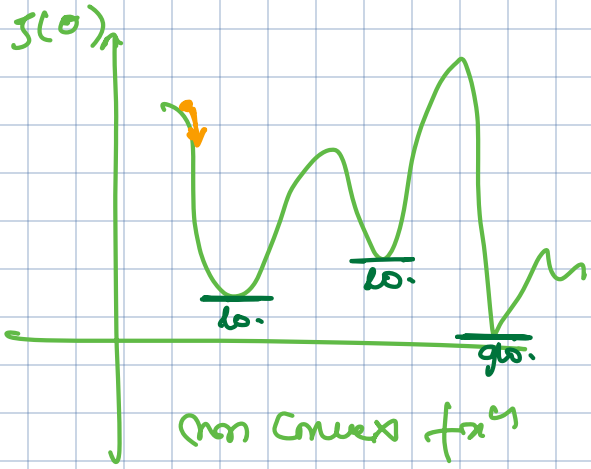
→ Hypothesis
→ Error
→ Gradient

Linear regression: $h_\theta(x) = \theta^T x = \theta_0 + \theta_1 x_1 + \cdots \theta_n x_n$

Logistic regression: $\frac{1}{1+e^{-\theta^T x}}$



Convex fx^n:
Local minima = Global minima



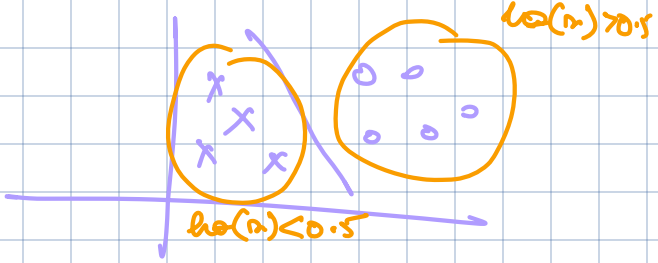
Non Convex fx^n

Binary Cross Entropy

$$h_\theta(x) = \frac{1}{1+e^{-\theta^T x}}$$ Prob. test point belongs to Class 1

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

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



0.6 60% class 1
40% class 0

Probability mass function

$$P(y \mid x;\theta) = [h_\theta(x)]^{y} [1 - h_\theta(x)]^{1-y}$$

Bernoulli Distribution

$y=1$ → $h_\theta(x)$; $y=0$ → $1 - h_\theta(x)$

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

Likelihood

$$P(y^{(1)} y^{(2)} \dots y^{(m)} \mid x^{(1)} x^{(2)} \dots x^{(m)};\theta)$$

$$= P(y^{(1)} \mid x^{(1)};\theta) \cdot P(y^{(2)} \mid x^{(2)};\theta) \cdot \dots \cdot P(y^{(m)} \mid x^{(m)};\theta)$$

$$= \prod_{i=1}^{m} P(y^{(i)} \mid x^{(i)};\theta)$$

→ likelihood of the data y.

$$LL(\theta) = \prod_{i=1}^{m} [h_\theta(x)]^{y} [1 - h_\theta(x)]^{1-y}$$

Loss minimize
LL maximize