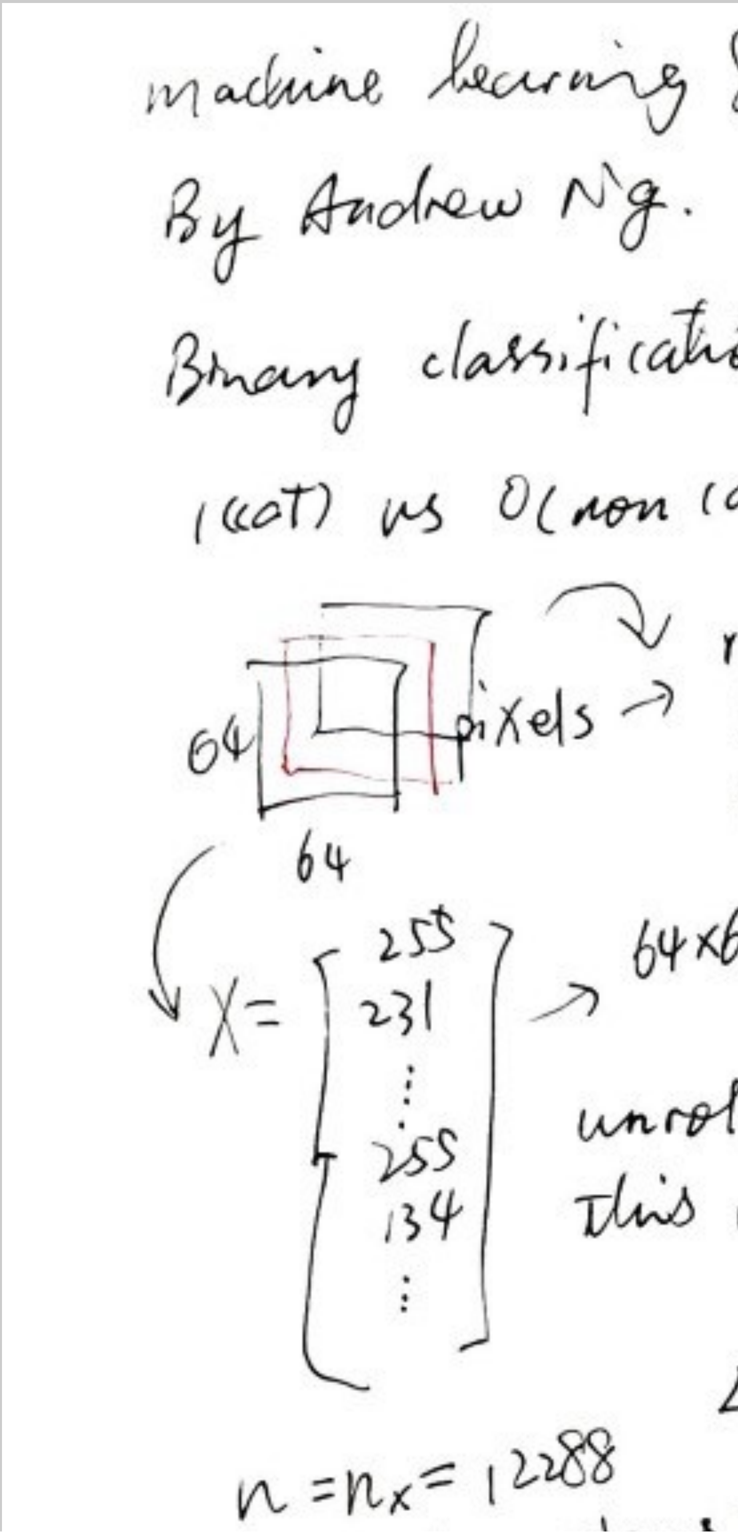
DN init1.pdf

Monday, August 12, 2019 3:12 PM



machine learning

By Andrew Ng.

Binary classificati

1(cat) vs 0(non

$$X = \begin{bmatrix} 255 \\ 231 \\ \vdots \\ 255 \\ 134 \\ \vdots \end{bmatrix} \rightarrow 64\times6$$

unrol

This

$n = n_x = 12288$

& neural networks.

ȣo →

at)

ed.
blue.
green

} 3 matrices

64x3

l all

values

each

$Y = [y^{(1)}\ y^{(2)} \dots\ y^{(m)}]$

$Y \in \mathbb{R}^{1 \times m}$.

$Y.\text{shape} = (1, m)$.

Logistic Regression :

Given x,

want $\boxed{\hat{y} = P(y=1 \mid x)}$

prediction.

if x is a pic.

you want $\hat{y}$ to tell

likely it is cat pi

dimensions.

(x, y) $x \in \mathbb{R}^{n_x}$.

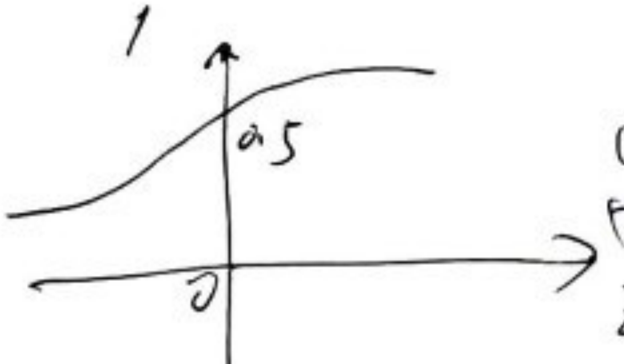
$y \in \{0,1$

m traing exampl

$\{(x^{(1)}, y^{(1)}) \ldots\ldots (x^{(m)},$

M train, M test

define. $X = \begin{bmatrix} | & | \\ x^{(1)} & x^{(2)} \\ | & | \end{bmatrix}$



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

add dumb feature

$x_0 = 1.$ $x \in \mathbb{R}^{n_x+1}$

}.

s.

$y^{(m)}\}$.

m examples.

$\ldots x^{(m)} \Big] \Big\} n_x$..

pic has n_x pixels in total.

you now

$\boxed{X \in \mathbb{R}^{n_x}}$

Paras: $w \in \mathbb{R}^{n_x}, b \in \mathbb{R}$.

Output:

$$\hat{y} = \boxed{w^T X + b}$$

X for binary

linear regression

In order to regulate $\hat{y}$ value between 0 & 1 to express probability. use. sigmoid function $\sigma(w^T x + b)$

---

onvex

r linear.

Square error makes gradient descent not working well for Logistic Regression.

Instead. we use: → Loss func.

$$\boxed{\mathcal{L}(\hat{y}, y) = -(y \log \hat{y} + (1-y)\log(1-\hat{y}))}$$

(If $y = 1$: $\mathcal{L}(\hat{y}, y) = -\log \hat{y}$ | essentially same.

* want $\mathcal{L} \downarrow$, $\hat{y} \to \uparrow$

$$\hat{y} = \sigma(\theta^T x)$$

$$\theta = \left[\begin{array}{l} \left.\theta_0\right\} \rightarrow b. \\ \left.\begin{array}{l}\theta_1 \\ \vdots \\ \theta_{n_x}\end{array}\right\} w \end{array}\right]$$

in pro

easie

Seperat

w . s

only theor

Want to find $\boxed{w \,\&\, b}$

fit $\hat{y}^{(i)} \approx y^{(i)}$

$$\hat{y}^{(i)} = \sigma(w^T x^{(i)} + b)$$

$$z^{(i)} = w^T x^{(i)} + b.$$

Loss function:

square error: $L(\hat{y}, y$

optima

co

strong

Reason for loss func

If $y=0$: $\mathcal{L}(\hat{y}, y) = -\log(1-\hat{y})$

ctive

want $\mathcal{L}\downarrow$, $(1-\hat{y})\uparrow \rightarrow \hat{y}$ small

→ overall cost func.

to

b &

o its

itical.

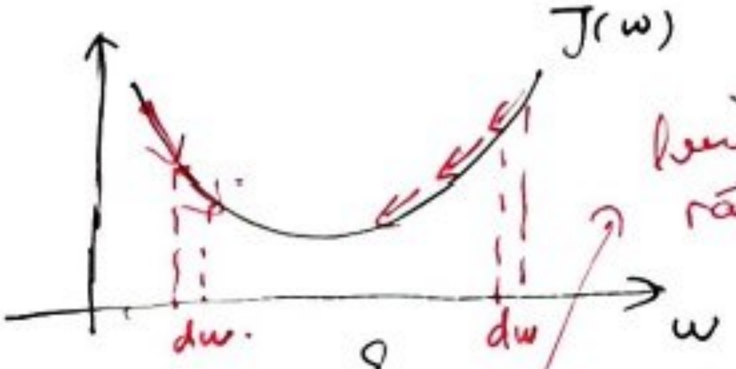
Cost func:

$$J(w, b) = \frac{1}{m}\sum_{i=1}^{m} \mathcal{L}(\hat{y}^{(i)}, y^{(i)})$$

to

$$= -\frac{1}{m}\left[\sum_{i=1}^{m} y^{(i)}\log\hat{y}^{(i)} + (1-y^{(i)})\log(1-\hat{y}^{(i)})\right]$$

Gradient Descent:



$) = \frac{1}{2}(\hat{y}-y)^2$

lling not

nvex

in fluence

Repeat {

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

learning rate

"dw"

}

Initialisation on both sides is fine.