

## Week 2-1 NN for LR

笔记本: DL 1 - NN and DL

创建时间: 2021/1/5 10:37

更新时间: 2021/1/7 01:09

## Binary Classification

64

64

Blue  
Green  
Red

255 134 93 22  
255 134 202 22  
255 231 42 22 4 30  
123 94 83 2 192 124  
34 44 187 92 34 142  
34 76 232 124 94  
67 83 194 202

64

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

$64 \times 64 \times 3 = 12288$

$n = n_x = 12288$

$X \rightarrow y$

## Notation

### Notation

$$(x, y) \quad x \in \mathbb{R}^{n_x}, y \in \{0, 1\}$$

$$m \text{ training examples: } \{(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \dots, (x^{(m)}, y^{(m)})\}$$

$$M = M_{\text{train}}$$

$$M_{\text{test}} = \# \text{ test examples.}$$

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

$X \in \mathbb{R}^{n_x \times m}$

$X \cdot \text{shape} = (n_x, m)$

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

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

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

# (different from ML)

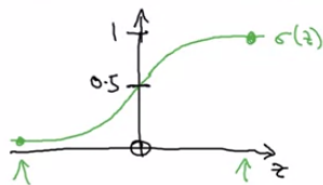
## LR

### Logistic Regression

Given  $x$ , want  $\hat{y} = P(y=1|x)$   
 $x \in \mathbb{R}^{n_x}$   $0 \leq \hat{y} \leq 1$

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

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



$$x_0 = 1, \quad x \in \mathbb{R}^{n_x+1}$$
$$\hat{y} = \sigma(\Theta^T x)$$

$$\Theta = \begin{bmatrix} \Theta_0 \\ \Theta_1 \\ \Theta_2 \\ \vdots \\ \Theta_{n_x} \end{bmatrix} \begin{matrix} \} b \leftarrow \\ \} w \leftarrow \end{matrix}$$

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

If  $z$  large  $\sigma(z) \approx \frac{1}{1+0} = 1$

If  $z$  large negative number

$$\sigma(z) = \frac{1}{1+e^{-z}} \approx \frac{1}{1+\text{Big num}} \approx 0$$

Andrew Ng

## Cost Func

### Logistic Regression cost function

$$\rightarrow \hat{y}^{(i)} = \sigma(w^T x^{(i)} + b), \text{ where } \sigma(z^{(i)}) = \frac{1}{1+e^{-z^{(i)}}} \quad z^{(i)} = w^T x^{(i)} + b$$

Given  $\{(x^{(1)}, y^{(1)}), \dots, (x^{(m)}, y^{(m)})\}$ , want  $\hat{y}^{(i)} \approx y^{(i)}$ .

$x^{(i)}$   
 $y^{(i)}$   
 $z^{(i)}$   $i$ -th example.

Loss (error) function:  $\mathcal{L}(\hat{y}, y) = \frac{1}{2} (\hat{y} - y)^2$

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

If  $y=1$ :  $\mathcal{L}(\hat{y}, y) = -\log \hat{y} \leftarrow$  Want  $\log \hat{y}$  large, want  $\hat{y}$  large.

If  $y=0$ :  $\mathcal{L}(\hat{y}, y) = -\log (1-\hat{y}) \leftarrow$  Want  $\log (1-\hat{y})$  large ... want  $\hat{y}$  small

$$\text{Cost function: } J(w, b) = \frac{1}{m} \sum_{i=1}^m \mathcal{L}(\hat{y}^{(i)}, y^{(i)}) = \frac{1}{m} \sum_{i=1}^m [y^{(i)} \log \hat{y}^{(i)} + (1-y^{(i)}) \log (1-\hat{y}^{(i)})]$$

What is the difference between the cost function and the loss function for logistic regression?

- ☐ The cost function computes the error for a single training example; the loss function is the average of the cost functions of the entire training set.
- ☐ They are different names for the same function.
- ☒ The loss function computes the error for a single training example; the cost function is the average of the loss functions of the entire training set.

✓ Correct

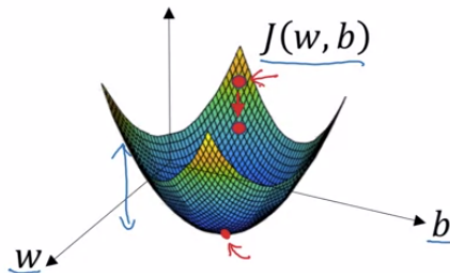
## GD

### Gradient Descent

Recap:  $\hat{y} = \sigma(w^T x + b)$ ,  $\sigma(z) = \frac{1}{1+e^{-z}}$  ←

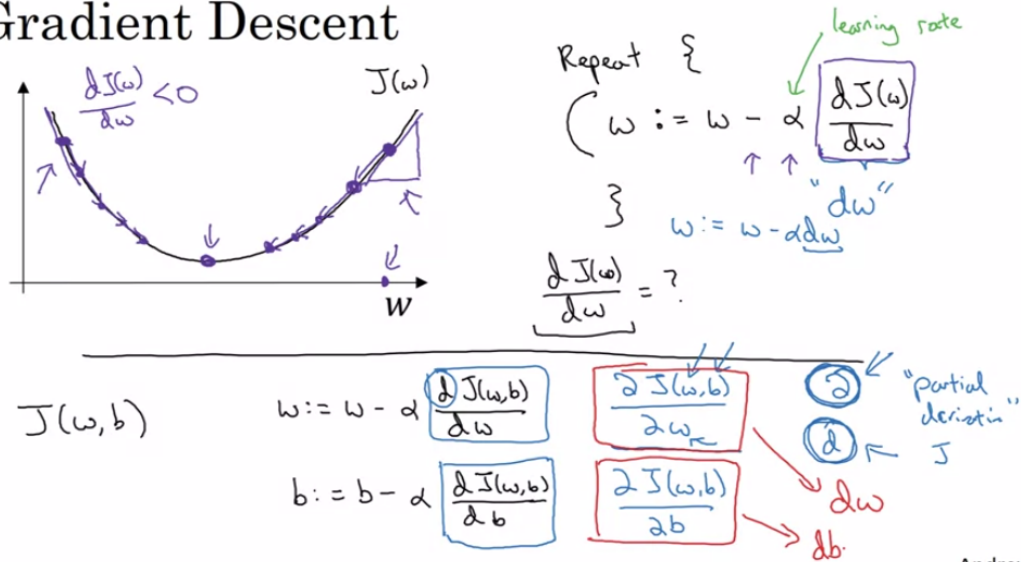
$$J(w, b) = \frac{1}{m} \sum_{i=1}^m \mathcal{L}(\hat{y}^{(i)}, y^{(i)}) = -\frac{1}{m} \sum_{i=1}^m y^{(i)} \log \hat{y}^{(i)} + (1 - y^{(i)}) \log(1 - \hat{y}^{(i)})$$

Want to find  $w, b$  that minimize  $J(w, b)$



Andrew Ng

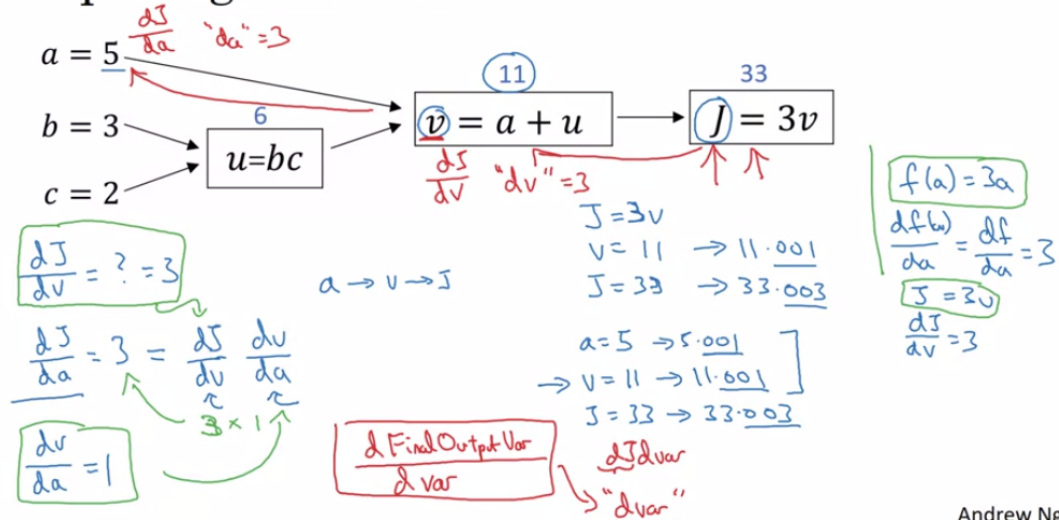
# Gradient Descent



Andrew Ng

# Computation Graph for derivatives

## Computing derivatives



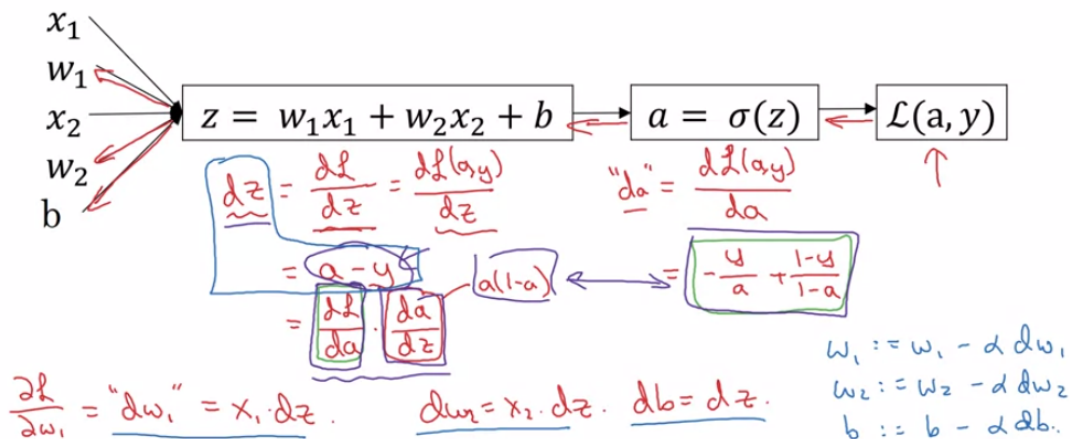
Andrew Ng

right to left

# Logistic regression recap

- $z = w^T x + b$
- $\hat{y} = a = \sigma(z)$
- $\mathcal{L}(a, y) = -(y \log(a) + (1 - y) \log(1 - a))$

## Logistic regression derivatives



## Logistic regression on $m$ examples

$$J = 0; \underline{dw_1} = 0; \underline{dw_2} = 0; \underline{db} = 0$$

For  $i = 1$  to  $m$

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

$$a^{(i)} = \sigma(z^{(i)})$$

$$J += -[y^{(i)} \log a^{(i)} + (1 - y^{(i)}) \log(1 - a^{(i)})]$$

$$\underline{dz^{(i)}} = a^{(i)} - y^{(i)}$$

$$dw_1 += x_1^{(i)} \underline{dz^{(i)}} \quad \uparrow n=2$$

$$dw_2 += x_2^{(i)} \underline{dz^{(i)}} \quad \downarrow$$

$$db += \underline{dz^{(i)}}$$

$$J /= m \leftarrow$$

$$\underline{dw_1} /= m; \underline{dw_2} /= m; \underline{db} /= m. \leftarrow$$

$$\underline{dw_1} = \frac{\partial J}{\partial w_1}$$

$$w_1 := w_1 - \alpha \underline{dw_1}$$

$$w_2 := w_2 - \alpha \underline{dw_2}$$

$$b := b - \alpha \underline{db}$$

# Logistic regression cost function

$$\begin{aligned} \rightarrow & \text{If } y = 1: p(y|x) = \hat{y} \\ \rightarrow & \text{If } y = 0: p(y|x) = 1 - \hat{y} \end{aligned} \quad \left. \vphantom{\begin{aligned} \rightarrow & \text{If } y = 1: p(y|x) = \hat{y} \\ \rightarrow & \text{If } y = 0: p(y|x) = 1 - \hat{y} \end{aligned}} \right\} p(y|x)$$

$$p(y|x) = \hat{y}^y (1-\hat{y})^{(1-y)} \quad \leftarrow$$

$$\text{If } y=1: p(y|x) = \hat{y} \underbrace{(1-\hat{y})^0}_{=1}$$

$$\text{If } y=0: p(y|x) = \hat{y}^0 (1-\hat{y})^{(1-0)} = 1 \times (1-\hat{y}) = 1-\hat{y}$$

$$\begin{aligned} \uparrow \log p(y|x) &= \log \hat{y}^y (1-\hat{y})^{(1-y)} = y \log \hat{y} + (1-y) \log (1-\hat{y}) \\ &= -\mathcal{L}(\hat{y}, y) \downarrow \end{aligned}$$

Andrew Ng

## maximum likelihood method

Cost on  $m$  examples

$$\log p(\text{labels in training set}) = \log \prod_{i=1}^m p(y^{(i)} | x^{(i)}) \quad \leftarrow$$

$$\begin{aligned} \log p(\dots) &= \sum_{i=1}^m \log p(y^{(i)} | x^{(i)}) \\ &= \sum_{i=1}^m -\mathcal{L}(\hat{y}^{(i)}, y^{(i)}) \\ &= -\sum_{i=1}^m \mathcal{L}(\hat{y}^{(i)}, y^{(i)}) \end{aligned}$$

maximum likelihood  
estimator  $\nwarrow$

$$\text{Cost: } \underbrace{J(w, b)}_{\text{(minimize)}} = \frac{1}{m} \sum_{i=1}^m \mathcal{L}(\hat{y}^{(i)}, y^{(i)})$$