(1) (i) Given error function of linear regression is

$$J(\omega) = \sum_i (y_i - \hat{y}_i)^2 \quad \text{---- eq (1)}$$

from the sum of squared differences

$J(\omega)$ can be written as:

$$J(\omega) = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \quad \text{---- eq (2)}$$

where y_i = actual value

$\hat{y}_i$ = predicted value

the partial derivative of $J(\omega)$ from eq (2) is

$$\frac{\partial J(\omega)}{\partial \omega_j} = -2 \sum_{i=1}^{n} (y_i - \hat{y}_i)\, x_{ij} \quad \text{---- eq (3)}$$

where x_{ij} is the j-th component vector of vector x_i $(\because \hat{y}_i = \omega \cdot x_i)$

Eq (3) can be written as:

$$\frac{\partial J}{\partial \omega_j} = -2 \sum_{i=1}^{n} (y_i - \hat{y}_i) \frac{\partial \hat{y}_i}{\partial \omega_j} \quad \text{---- eq (4)}$$

where $j = 0, 1, 2 \cdots p, n$

for $\frac{\partial \hat{y}_i}{\partial \omega_j} \Longrightarrow$

for $j=0$	for $j=1$	for $j=2$	for $j=p$
$\frac{\partial \hat{y}_i}{\partial w_0} = 1$	$\frac{\partial \hat{y}_i}{\partial w_1} = x_{i1}$	$\frac{\partial \hat{y}_i}{\partial w_2} = x_{i2}$	$\frac{\partial \hat{y}_i}{\partial w_p} = x_{ip}$

Using gradient descent eq (4) can be given as:

$$w_j = w_j + 2\alpha \sum_{i=1}^{n} (y_i - \hat{y}_i)\, x_{ij}$$

{α is the learning rate, $\alpha = 0.1$}

In a convex function any local minimum is also a global minimum. Therefore the given function is convex and gradient descent algorithm is guaranteed to give a global minimum as long as the learning rate is chosen appropriately.

(1) (iii) Given $\hat{y}_i = w_0 + w_1 x_1$

Partial derivative w.r.t w_0:

$$\frac{\partial J(w)}{\partial w_0} = -2 \sum_{i} (y_i - \hat{y}_i) \frac{\partial \hat{y}_i}{\partial w_0}$$

Substitute $\hat{y}_i = w_0 + w_1 x_1$

$$\frac{\partial J(w)}{\partial w_0} = -2 \sum_i (y_i - (w_0 + w_1 x_{1i}))$$

$$\frac{\partial J(w)}{\partial w_0} = -2 \sum_i (y_i - w_0 - w_1 x_{1i}) \quad \text{— eq ①}$$

Partial derivative w.r.t w_1:

$$\frac{\partial J(w)}{\partial w_1} = -2 \sum_i (y_i - \hat{y}_i) \frac{\partial \hat{y}_i}{\partial w_1}$$

Substitute $\hat{y}_i = w_0 + w_1 x_{1i}$

$$\frac{\partial J(w)}{\partial w_1} = -2 \sum_i (y_i - (w_0 + w_1 x_{1i})) \frac{\partial \hat{y}_i}{\partial w_1}$$

Since $\frac{\partial \hat{y}_i}{\partial w_1} = x_{1i}$ we get

$$\frac{\partial J(w)}{\partial w_1} = -2 \sum_i (y_i - w_0 - w_1 x_{1i}) x_{1i} \quad \text{— eq ②}$$

Eq ① and eq ② are the derivatives for the equation respectively

② (i) Given linear regression model $\hat{y} = 2 + 4 \cdot x_1$

Dataset $D = \{(1,2), (3,4)\}$

$x_1 = 2 \cdot 3$

Sub $x_1 = 2.3$ in the equation

$$\hat{y} = 2 + 4 \times 2.3$$

$$\hat{y} = 2 + 9.2$$

$$\hat{y} = 11.2$$

$\therefore$ for $x_1 = 2.3$, $\boxed{\hat{y} = 11.2}$

(2)(ii) Residual error (e_i) = actual target val -ve (y_i) - predicted value ($\hat{y}_i$)

$$e_i = y_i - \hat{y}_i$$

Given $D = \{(1,2), (3,4)\}$, $\hat{y} = 2 + 4 \cdot x_1$

for (1,2):

$$e_1 = y_1 - \hat{y}_1 = 2 - (2 + 4 \cdot 1) = 2 - 6 = -4$$

for (3,4):

$$e_2 = y_2 - \hat{y}_2 = 4 - (2 + 4 \cdot 3) = 4 - 14 = -10$$

Total sum of squared errors is sum of squared residuals

$$SSE = \sum_i e_i^2$$

$$SSE = (-4)^2 + (-10)^2 = 16 + 100 = 116$$

∴ the residuals for each point $e_1 = -4$,

$e_2 = -10$ and the SSE = 116

(2) (iii) Squaring of e_i values in error functi-

-on has several purposes:

* Handling positive and negative errors
* Emphasizing larger errors
* Mathematical simplicity

(4) Given $p(x;\omega) = \frac{1}{1+\exp(-\omega^T x)}$

x = input vector

ω = weight vector

- By defining $p_1(x;\omega)$ we get,

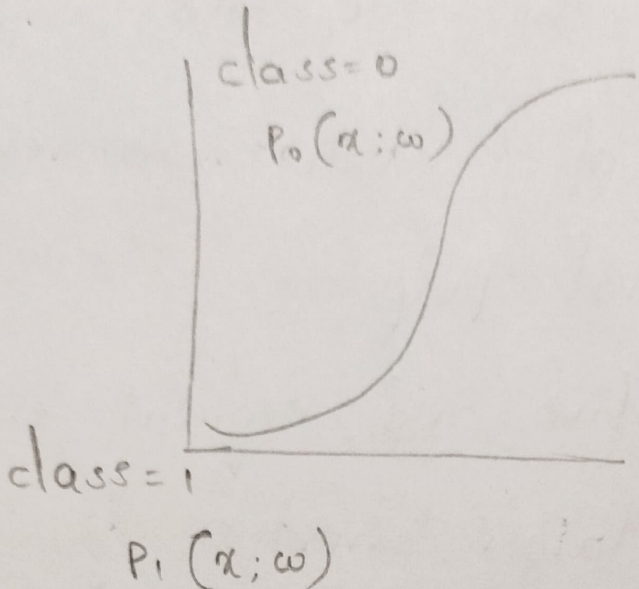
$$p_1(x;\omega) = P(x;\omega) = \frac{1}{1+\exp(-\omega^T x)}$$

this is the probability of the positive class (class 1)

- By defining $p_0(x;\omega)$ we get,

$$p_0(x;\omega) = 1 - p_1(x;\omega)$$

this is the probability of negative class. (class 0).



(6) In logistic regression the odds and logit are related to success probability (p)

* odds (O):

$$O = \frac{p}{1-p}$$

* logit (logit (p)):

$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right)$$

Given $p = 0.8$

odds:

$$O = \frac{0.8}{1-0.8} = \frac{0.8}{0.2} = 4$$

logit:

$$\text{logit}(0.8) = \log\left(\frac{0.8}{1-0.8}\right) = \log\left(\frac{0.8}{0.2}\right)$$

$\text{logit}(0.8) = \log_e(4)$

$\therefore$ Approximately $\text{logit}(0.8) \approx 1.386$

So the odds are 4 and the logit is approximately 1.386

(7) (i) In logistic regression the typical error or cost function used is the binary-cross entropy loss

The binary cross-entropy loss is defined as:

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

where

m — is the number of training examples

$h_\theta(x^{(i)})$ — is the predicted probability of the positive class for the i-th example

$y^{(i)}$ — is the actual value label (0 or 1) for the ith example

The partial derivative of the cost function with respect to each parameter θ_j is given by:

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

(7)(ii) The error function of logistic regression with sum of squared error can be given as:

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

where,

m - is the number of training examples

$h_\theta(x^{(i)})$ - is the predicted probability of the positive class for i-th exam-ple.

$y^{(i)}$ - is the actual label (0 or 1) for the ith example

(8) The cross-entropy loss for a binary classification problem is given as:

$$\text{Cross-entropy loss} = -(y \log(\hat{y}) + (1-y) \log(1-\hat{y}))$$

where y is the true label

$\hat{y}$ is the predicted probability

Given $y = 1$ and $\hat{y} = 0.2$

$$\text{Cross-entropy loss} = -(1 \cdot \log(0.2) + (1-1) \cdot \log(1-0.2))$$

$$= -(\log 0.2)$$

$$\approx -(-0.6989)$$

$$\approx 0.6989$$

$\therefore$ the value of cross-entropy ≈ 0.6989