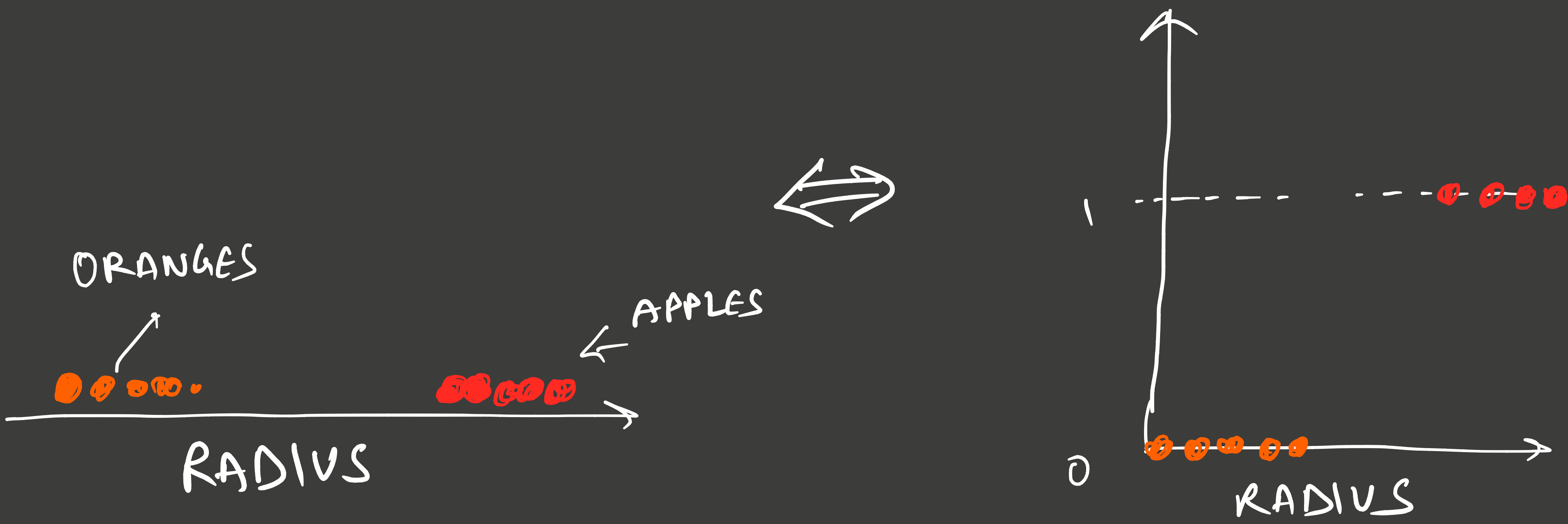
LOGISTIC REGRESSION

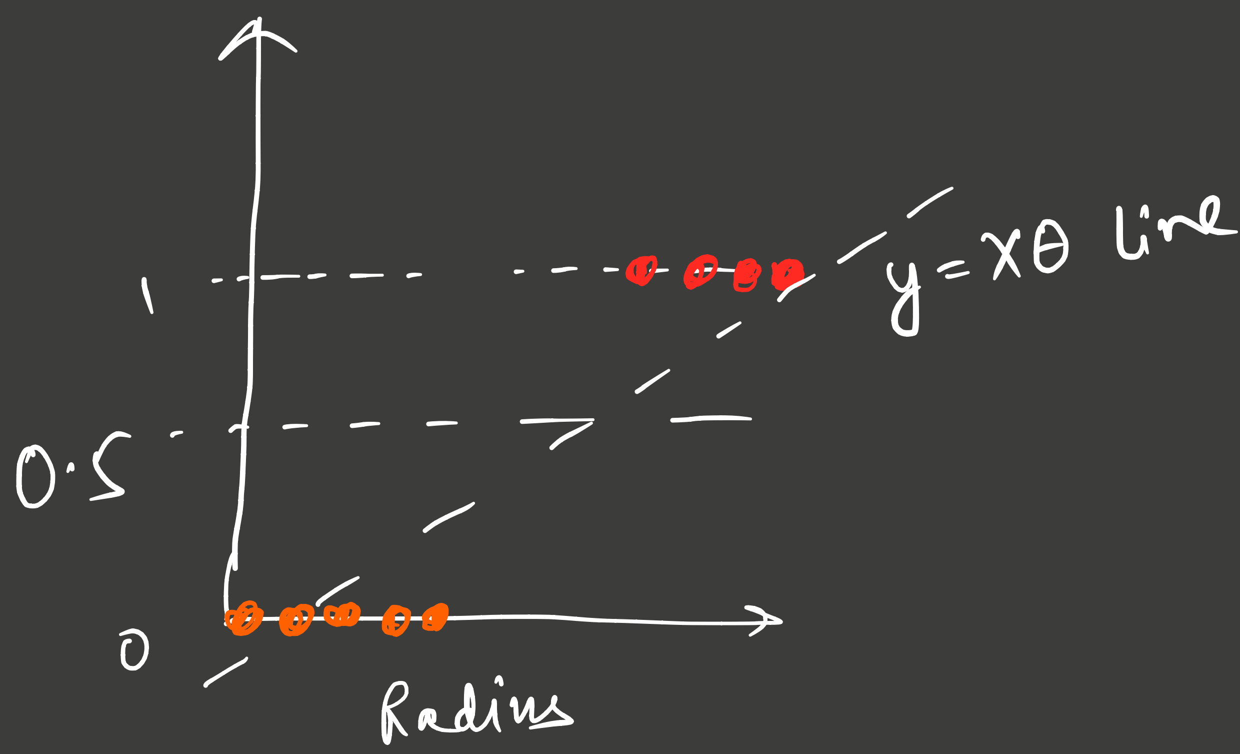
* Classification Technique



AIM:

Probability (Apple | Radius)? or $\boxed{\underset{\text{more generally}}{p(1|x)}}$

IDEA : JUST USE LINEAR REGRESSION



$$P(\text{FRUIT} = \text{APPLE} \mid \text{RADIUS}) = \theta_0 + \theta_1 * \text{RADIUS}$$

Generally

$$P(y=1 \mid x) = X\theta$$

PREDICTION

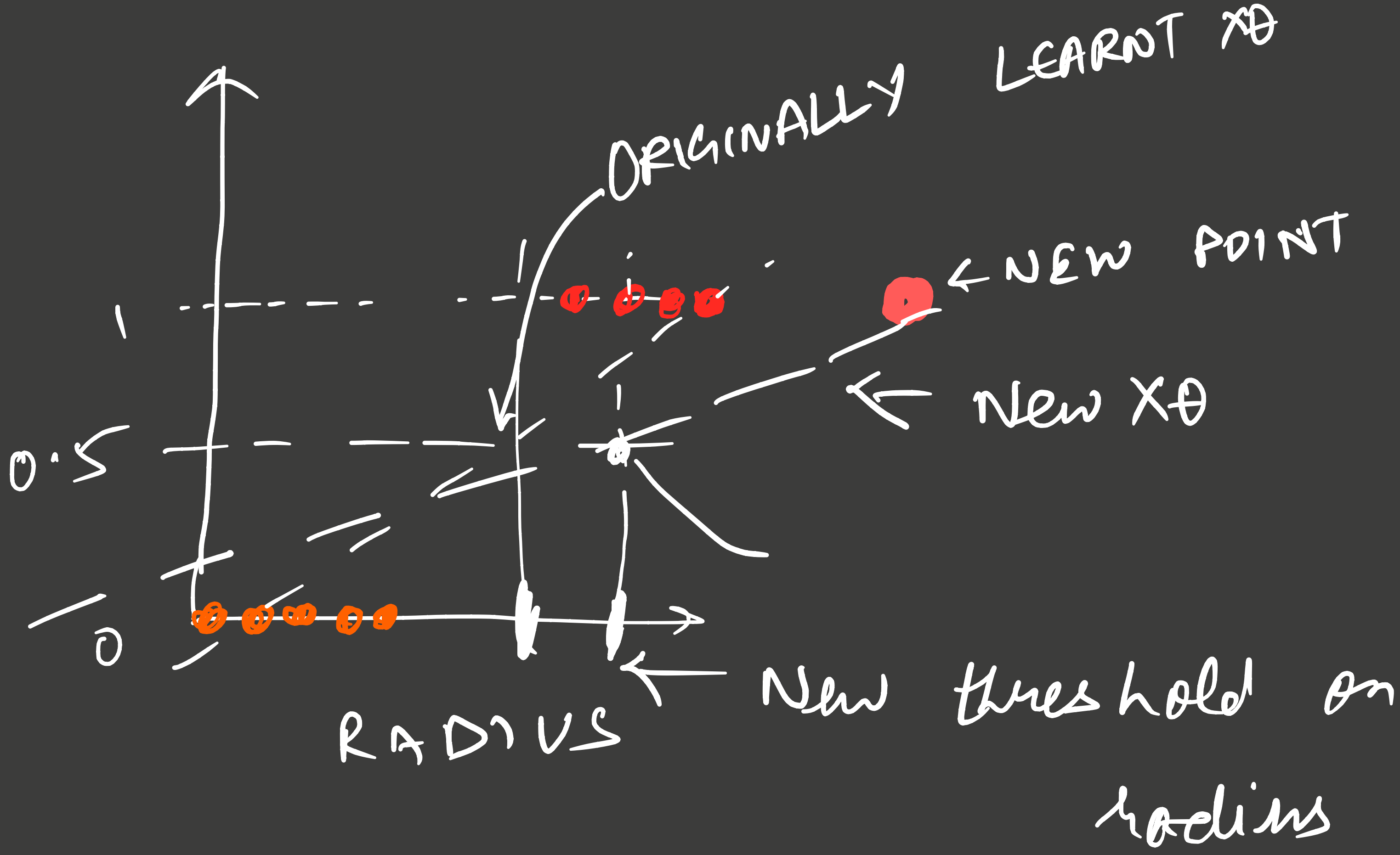
If $\theta_0 + \theta_1 * \text{Radius} > 0.5 \longrightarrow$ APPLE

else $\rightarrow$ ORANGE

PROBLEM

1. Range of $X\theta$ is $(-\infty, \infty)$

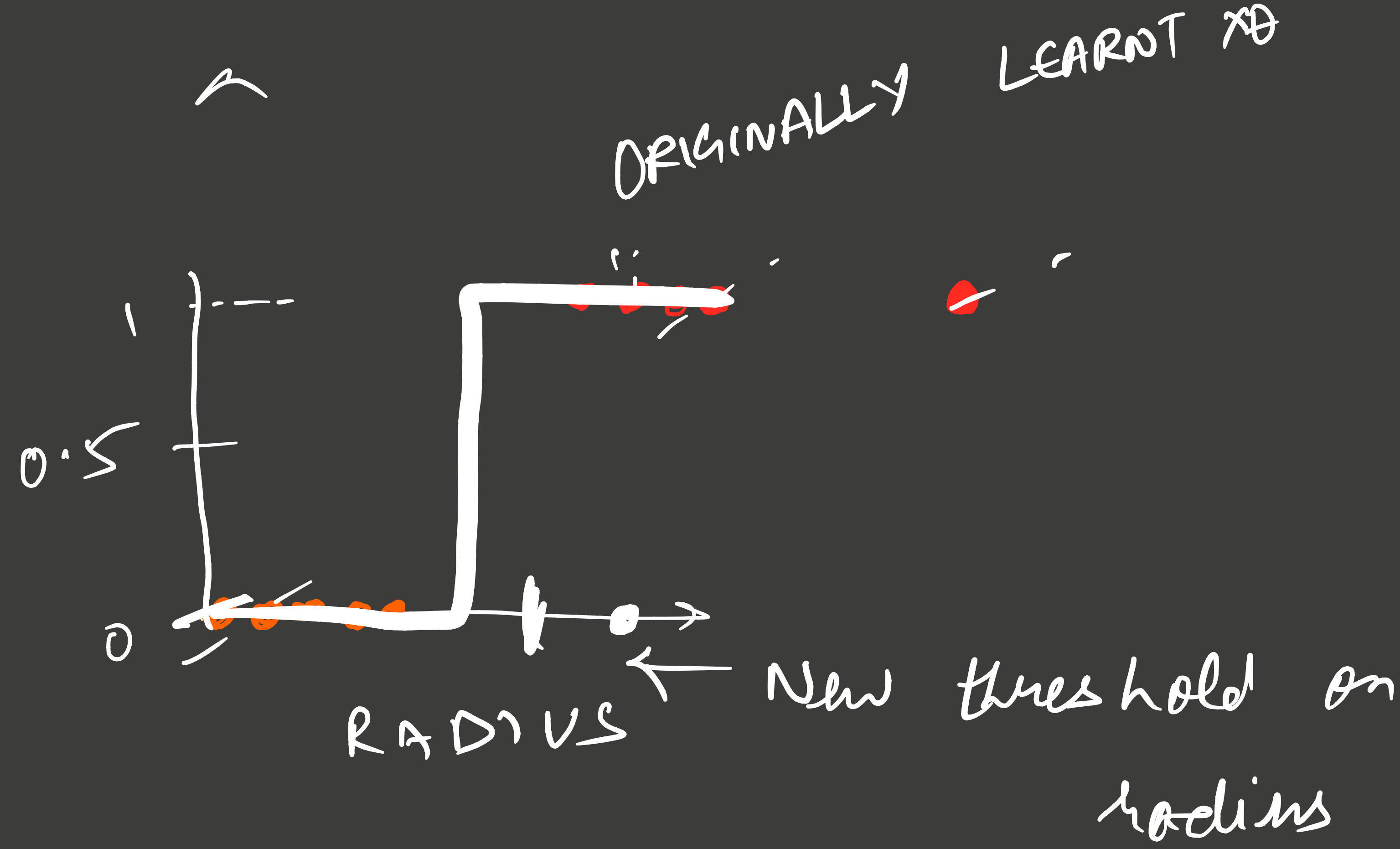
But $P(y=1 | \ldots) \in [0,1]$



Q) CAN WE STILL USE LINEAR REGRESSION?

YES! TRANSFORM $\hat{y} \rightarrow [0, 1]$

How?



Q) CAN WE STILL USE LINEAR REGRESSION?

YES! TRANSFORM $\hat{y} \to [0, 1]$

How?

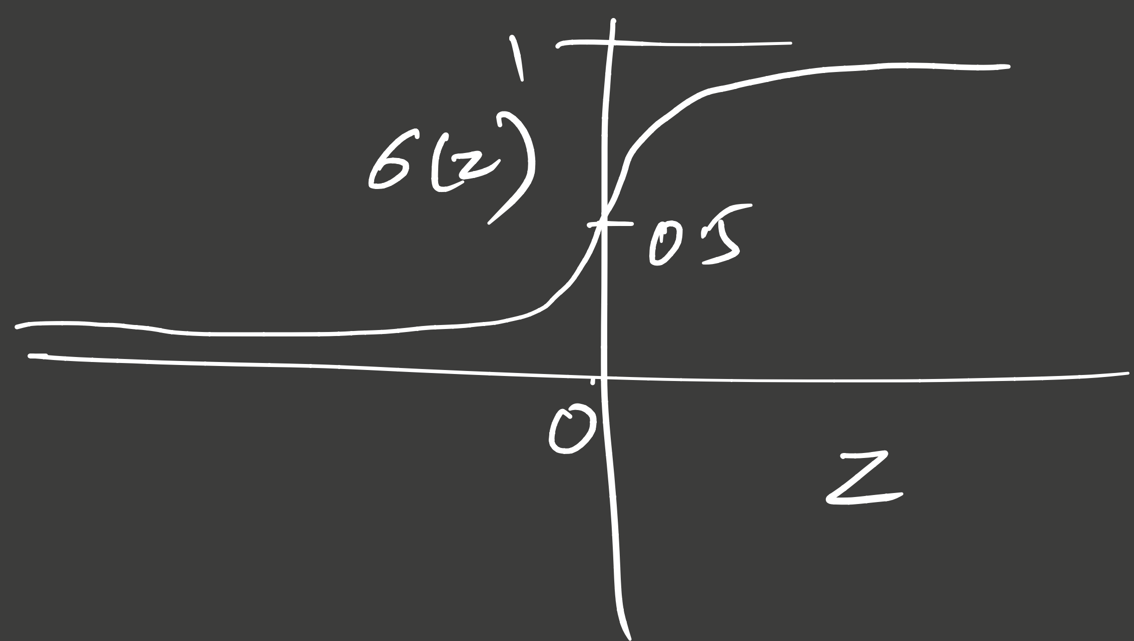
LOGISTIC | SIGMOID FUNCTION

$\hat{y} \in (-\infty, \infty)$

$\phi =$ SIGMOID | LOGISTIC FUNCTION (σ)

$\phi(\hat{y}) \in [0,1]$

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



$z \to \infty$

$\sigma(z) \to 1$

$z \to -\infty$

$\sigma(z) \to 0$

$z = 0$

$\sigma(z) = 0.5$

Q) Could you use some other transformation (ϕ) of $\hat{y}$ s.t.

$$\phi(\hat{y}) \in [0, 1]$$

Yes! But Logistic Regression works

$$P(y=1|X) = \sigma(X\theta) = \frac{1}{1+e^{-X\theta}}$$

Q) write $X\theta$ in a more convenient form (as $P(y=1|x)$, $P(y=0|X)$)

$$P(y=1|X) = \sigma(X\theta) = \frac{1}{1+e^{-X\theta}}$$

Q) Write $X\theta$ in a more convenient form (as $P(y=1|X)$, $P(y=0|X)$)

$$P(y=0|X) = 1 - P(y=1|X) = 1 - \frac{1}{1+e^{-X\theta}} = \frac{e^{-X\theta}}{1+e^{-X\theta}}$$

$$\therefore \frac{P(y=1|X)}{1-P(y=1|X)} = e^{X\theta} \quad \text{or} \quad X\theta = \log\left\{\frac{P(y=1|X)}{1-P(y=1|X)}\right\}$$

Odds (used in betting)

$$\frac{P(win)}{P(loss)}$$

Here

$$\text{odds} = \frac{P(y=1)}{P(y=0)}$$

$$\boxed{\text{Log-odds} = \log\left\{\frac{P(y=1)}{P(y=0)}\right\} = X\theta}$$

Q) What is decision boundary for Logistic Regression?

Q) What is decision boundary for Logistic Regression?

Decision boundary : $P(y=1|x) = P(y=0|x)$

or $\dfrac{1}{1+e^{-X\theta}} = \dfrac{e^{-X\theta}}{1+e^{-X\theta}}$

or $e^{X\theta} = 1$

or $X\theta = 0$

Notebook

LEARNING PARAMETERS

Could we use cost function as:

$$J(\theta) = \Sigma(y_i - \hat{y}_i)^2$$

$$\hat{y}_i = \sigma(X\theta)$$

Answer: NO (Non convex)

{ See Jupyter Notebook }

LIKELIHOOD $= p(D|\theta)$

$$p(y|X,\theta) = \prod_{i=1}^{n} p(y_i|x_i,\theta)$$

$y \rightarrow$ 0 or 1

LIKELIHOOD = $p(D|\theta)$

$$p(y|X,\theta) = \prod_{i=1}^{n} p(y_i|x_i,\theta)$$

$y \to$ 0 or 1

$$= \prod_{i=1}^{n} \left\{\frac{1}{1+e^{-x_i\theta}}\right\}^{y_i} \left\{1 - \frac{1}{1+e^{-x_i\theta}}\right\}^{1-y_i}$$

[Above: similar to $P(D|\theta)$ for Linear Reg.
Difference Bernoulli instead of Gaussian]

$-\log p(y|X,\theta)$ = Negative log. likelihood
= Cost function we're minimizing
= $J(\theta)$

$$J(\theta) = -\log\left\{\prod_{i=1}^{n}\left\{\frac{1}{1+e^{-x_i\theta}}\right\}^{y_i}\left\{1-\frac{1}{1+e^{-x_i\theta}}\right\}^{1-y_i}\right\}$$

$$J(\theta) = -\left\{\sum_{i=1}^{n} y_i \log(\sigma_\theta(x_i)) + (1-y_i)\log(1-\sigma_\theta(x_i))\right\}$$

$$\frac{\partial J(\theta)}{\partial \theta_j} = -\frac{\partial}{\partial \theta_j}\left\{\sum_{i=1}^{n} y_i \log(\sigma_\theta(x_i)) + (1-y_i)\log(1-\sigma_\theta(x_i))\right\}$$

$$= -\sum_{i=1}^{n}\left[y_i \frac{\partial}{\partial \theta_j}\log(\sigma_\theta(x_i)) + (1-y_i)\frac{\partial}{\partial \theta_j}\log(1-\sigma_\theta(x_i))\right]$$

$$\frac{\partial J(\theta)}{\partial \theta_j} = -\sum_{i=1}^{n} \left[y_i \frac{\partial}{\partial \theta_j} \log(\sigma_\theta(x_i)) + (1-y_i) \frac{\partial}{\partial \theta_j} \log(1-\sigma_\theta(x_i)) \right]$$

$$= -\sum_{i=1}^{n} \left[\frac{y_i}{\sigma_\theta(x_i)} \frac{\partial}{\partial \theta_j} \sigma_\theta(x_i) + \frac{1-y_i}{1-\sigma_\theta(x_i)} \frac{\partial}{\partial \theta_j} (1-\sigma_\theta(x_i)) \right]$$

..... 1

ASIDE

$$\frac{\partial}{\partial z} \sigma(z) = \frac{\partial}{\partial z} \frac{1}{1+e^{-z}} = -(1+e^{-z})^{-2} \frac{\partial}{\partial z} (1+e^{-z})$$

$$= \frac{e^{-z}}{(1+e^{-z})^2} = \left(\frac{1}{1+e^{-z}}\right)\left(\frac{e^{-z}}{1+e^{-z}}\right) = \sigma(z) \left\{ \frac{1+e^{-z}}{1+e^{-z}} - \frac{1}{1+e^{-z}} \right\}$$

$$= \sigma(z)(1-\sigma(z))$$

Resuming From 1

$$\frac{\partial J(\theta)}{\partial \theta_j} = -\sum_{i=1}^{n}\left[\frac{y^i}{\sigma_\theta(x^i)}\frac{\partial}{\partial \theta_j}\sigma_\theta(x^i) + \frac{(1-y_i)}{1-\sigma_\theta(x^i)}\frac{\partial}{\partial \theta_j}(1-\sigma_\theta(x^i))\right]$$

$$-\sum_{i=1}^{n}\left[\frac{y_i\ \cancel{\sigma_\theta(x_i)}(1-\sigma_\theta(x_i))}{\cancel{\sigma_\theta(x_i)}}\frac{\partial}{\partial \theta_j}(x_i\theta) - \frac{(1-y_i)\ \sigma_\theta(x_i)\ \cancel{(1-\sigma_\theta(x_i))}}{\cancel{(1-\sigma_\theta(x_i))}}\frac{\partial}{\partial \theta_j}|1-\sigma_\theta(x_i)|\right]$$

$$= -\sum_{i=1}^{n}\left[y_i(1-\sigma_\theta(x_i))x_i^j - (1-y_i)\sigma_\theta(x_i)\,x_i^j\right]$$

$$= -\sum_{i=1}^{n}\left[\left(y_i - \cancel{y_i\sigma_\theta(x_i)} - \sigma_\theta(x_i) + \cancel{y_i\sigma_\theta(x_i)}\right)x_i^j\right]$$

$$= \sum_{i=1}^{n}\left[\sigma_\theta(x^i) - y^i\right]x_i^j$$

$$\boxed{\frac{\partial J(\theta)}{\partial \theta_j} = \sum_{i=1}^{N} \left[\sigma_\theta(x_i) - y_i \right] x_i^j}$$

Now, just use Gradient Descent!!

REGULARIZED LOGISTIC REGRESSION

Unreg.

$$J_1(\theta) = -\left\{ \sum_{i=1}^{m} y_i \log(\sigma_\theta(x_i)) + (1-y_i)\log(1-\sigma_\theta(x_i)) \right\}$$

L_2 REG.

$$J(\theta) = J_1(\theta) + \lambda \, \theta^T \theta$$

L1 REG.

$$J(\theta) = J_1(\theta) + \lambda |\theta|$$

MULTI- CLASS PREDICTION

1. Use ONE-VS-ALL ON BINARY LOGISTIC REGRESSION

2. Use ONE-VS-ONE ON BINARY LOGISTIC REGRESSION

3. Extend BINARY LOGISTIC REGRESSION TO MULTI-CLASS LOGISTIC REGRESSION

SOFTMAX

$$Z \in R^d$$

$$f(z_i) = \frac{e^{z_i}}{\sum_{i=1}^{d} e^{z_i}}$$

$$\therefore \sum f(z_i) = 1$$

$f(z_i)$ refers to **probability** of class i

SOFTMAX FOR MULTI-CLASS LOG. REGRESSION

$k = 1, \ldots K$ classes

$$\Theta = \begin{bmatrix} | & | & & | \\ \theta_1 & \theta_2 & \cdots & \theta_K \\ | & | & & | \end{bmatrix}$$

$$\boxed{P(y = k \mid x, \theta) = \frac{e^{x\theta_k}}{\sum_{k=1}^{K} e^{x\theta_k}}}$$

SOFTMAX FOR MULTI-CLASS LOG. REGRESSION

For K=2 classes

$$P(y=k\mid x,\theta) = \frac{e^{x\theta_k}}{\sum_{k=1}^{K} e^{x\theta_k}}$$

$$P(y=0\mid x,\theta) = \frac{e^{x\theta_0}}{e^{x\theta_0}+e^{x\theta_1}}$$

$$P(y=1\mid x,\theta) = \frac{e^{x\theta_1}}{e^{x\theta_0}+e^{x\theta_1}} = \frac{e^{x\theta_1}}{e^{x\theta_1}\left\{1+e^{x(\theta_0-\theta_1)}\right\}}$$

$$= \frac{1}{1+e^{-x\theta'}}$$

= SIGMOID!

MULTI-CLASS LOGISTIC REGRESSION COST

FOR 2 CLASS WE HAD.

$$J(\theta) = -\left\{ \sum_{i=1}^{n} y_i \log(\sigma_\theta(x_i)) + (1 - y_i) \log(1 - \sigma_\theta(x_i)) \right\}$$

EXTEND TO K CLASS

$$J(\theta) = -\left\{ \sum_{i=1}^{n} \sum_{k=1}^{K} I\{y_i = k\} \log \frac{e^{x_i \theta_k}}{\sum_{k=1}^{K} e^{x_i \theta_k}} \right\}$$

$i \rightarrow$ sample #

$k \rightarrow$ class

I : Identity function

$I(\text{TRUE}) = 1$; $I(\text{FALSE}) = 0$

Now;

$$\frac{\partial J(\theta)}{\partial \theta_k} = -\sum_{i=1}^{n} \left[x_i \left\{ I(y_i = k) - p(y_i = k \mid x_i; \theta) \right\} \right]$$