Assignment-2

Q.1. Compute the conditional probability using MLE for feature color:
P(C=Black|Good), P(C=Green|Good), P(C=White|Good).

Sol. Suppose, $P(C=Black|Good) = \theta_1$
$P(C=Green|Good) = \theta_2$
$P(C=White|Good) = 1-\theta_1-\theta_2$

Based on given dataset, the likelyhood function to observe the above result is.

$$f(x,\theta_1,\theta_2) = \arg\max_{\theta}(\theta_1^4\ \theta_2^3\ (1-\theta_1-\theta_2))$$

$$= \log(\arg\max_{\theta}(\theta_1^4\ \theta_2^3\ (1-\theta_1-\theta_2))$$

$$= \log\theta_1^4 + \log\theta_2^3 + \log(1-\theta_1-\theta_2)$$

$$= 4\log\theta_1 + 3\log\theta_2 + \log(1-\theta_1-\theta_2)$$

$$\frac{\partial L_1}{\partial\theta_1} = \frac{4}{\theta_1} + 0 + \frac{1}{1-\theta_1-\theta_2}\times(-1)$$

for maximum $\frac{\partial L}{\partial\theta_1} = 0$

$$0 = \frac{4}{\theta_1} + -\frac{1}{1-\theta_1-\theta_2}$$

$$\frac{4}{\theta_1} = \frac{1}{1-\theta_1-\theta_2}$$

$$4-4\theta_1-4\theta_2 = \theta_1 \Rightarrow 4(1-\theta_2) = 5\theta_1 \Rightarrow \theta_{1_{max}} = \frac{4}{5}(1-\theta_2)$$

Now, lets compute $\frac{\partial L}{\partial \theta_2} = 0$

$$\frac{\partial L}{\partial \theta_2} = \log \theta_1^4 + \log \theta_2^3 + \log (1-\theta_1-\theta_2)$$

$$= \frac{\partial}{\partial \theta_2} (4 \log \theta_1 + 3 \log \theta_2 + \log (1-\theta_1-\theta_2))$$

$$= 0 + \frac{3}{\theta_2} + \frac{1}{1-\theta_1-\theta_2} \times (-1)$$

$$= \frac{3}{\theta_2} - \frac{1}{1-\theta_1-\theta_2}$$

for maximum $\frac{\partial L}{\partial \theta_2} = 0$

$$\frac{3}{\theta_2} = \frac{1}{1-\theta_1-\theta_2}$$

$$3 - 3\theta_1 - 3\theta_2 = \theta_2$$

$$3(1-\theta_1) = 4\theta_2$$

$$\Rightarrow \underset{max}{\theta_2} = \frac{3}{4}(1-\theta_1)$$

By placing θ_2 in θ_1 equation,

$$\underset{max}{\theta_1} = \frac{4}{5}\left(1 - \frac{3}{4}(1-\theta_1)\right)$$

$$= \frac{4}{5} - \frac{3}{5}(1-\theta_1)$$

$$\underset{\text{max}}{\theta_1} = \frac{4}{5} - \frac{3}{5} + \frac{3}{5}\theta_1$$

$$\theta_1 - \frac{3}{5}\theta_1 = \frac{1}{5}$$

$$\frac{5\theta_1 - 3\theta_1}{5} = \frac{1}{5}$$

$$2\theta_1 = 1$$

$$\theta_1 = \frac{1}{2}$$

So, $P(C = \text{Black} | \text{Good}) = \frac{1}{2}$.

Now, using θ_1 we will calculate θ_2

$$\theta_2 = \frac{3}{4}(1 - \theta_1)$$

$$= \frac{3}{4}\left(1 - \frac{1}{2}\right)$$

$$= \frac{3}{4} \times \frac{1}{2}$$

$$\theta_2 = \frac{3}{8}$$

Therefore, $P(C = \text{Green} | \text{Good}) = \frac{3}{8}$

Then, calculate $P(C = \text{white} \mid \text{Good})$

$$P(C = \text{white} \mid \text{Good}) = 1 - \theta_1 - \theta_2$$

$$= 1 - \frac{1}{2} - \frac{3}{8}$$

$$= \frac{8 - 4 - 3}{8}$$

$$= \frac{1}{8}$$

Answer → ① $P(C = \text{Black} \mid \text{Good}) = \frac{1}{2}$

② $P(C = \text{Green} \mid \text{Good}) = \frac{3}{8}$

③ $P(C = \text{White} \mid \text{Good}) = \frac{1}{8}$

Q.2. The following data set shows the different parameter depending on which a person may or may not buy a computer. Use Naive Bayes classifier to find out if a person with age <=30, income = medium, student = yes, credit_rating = fair, will or will not buy a computer. Show individual probability calculated as well as the probability for the final class.

Sol. E = age<=30, income = medium, student = yes, creditRating = fair
E_1 is age <=30
E_2 is Income = medium
E_3 is studen = yes
E_4 is credit_rating = fair

$$P(yes|E) = \frac{P(E_1|yes)\,P(E_2|yes)\,P(E_3|yes)\,P(E_4|yes)\,P(yes)}{P(E)}$$

$$P(yes) = \frac{9}{14} = 0.643$$

$$P(E_1|yes) = \frac{2}{14} \Big/ \frac{9}{14} = \frac{2}{9} = 0.222$$

$$P(E_2|yes) = \frac{4}{14} \Big/ \frac{9}{14} = \frac{4}{9} = 0.444$$

$$P(E_3|yes) = \frac{6}{14} \Big| \frac{9}{14} = \frac{6}{9} = 0.667$$

$$P(E_4|yes) = \frac{6}{14} \Big| \frac{9}{14} = \frac{6}{9} = 0.667$$

$$P(yes|E) = \frac{(0.222) \times (0.444) \times (0.667) \times (0.667) \times (0.643)}{P(E)}$$

$$= \frac{0.028}{P(E)}$$

$$P(no \mid E) = \frac{P(E_1|no)\, P(E_2|no)\, P(E_3|no)\, P(E_4|no)\, P(no)}{P(E)}$$

$$P(no) = \frac{5}{14} = 0.357$$

$$P(E_1|no) = \frac{3}{14} \Big| \frac{5}{14} = \frac{3}{5} = 0.600$$

$$P(E_2|no) = \frac{2}{14} \Big| \frac{5}{14} = \frac{2}{5} = 0.400$$

$$P(E_3|no) = \frac{1}{14} \Big| \frac{5}{14} = \frac{1}{5} = 0.200$$

$$P(E_4|no) = \frac{2}{14} \Big| \frac{5}{14} = \frac{2}{5} = 0.400$$

$$P(no|E) = \frac{(0.600) \times (0.400) \times (0.200) \times (0.400) \times (0.357)}{P(E)}$$

$$= \frac{0.007}{P(E)}$$

Therefore, the Naive Bayes Classifier predicts buy_computer = yes for the given set of situtions.

Q.3. The prediction of an Naive Bayes Classifier can be described as

$$h_{nb}(x) = \underset{c \in Y}{\operatorname{argmax}} P(c) \prod_{i=1}^{d} P(x_i | c).$$

When the data dimension is high ($d \to \infty$), the production of the probability of any $c \in Y$ will quickly converage to zero. Propose your solution to prevent this from happening.

Sol.

To prevent this production rule to converge to zero because of dimension equal to ∞. We will take log and convert multiplication to summation.

$$h_{nb}(x) = \underset{c \in Y}{\operatorname{argmax}} P(c) \prod_{i=1}^{d} P(x_i | c)$$

$$= \log \left(P(c) \prod_{i=1}^{d} P(x_i | c) \right)$$

$$= \log P(c) + \sum_{i=1}^{d} \log P(x_i | c)$$

Other than this, we can use laplace smoothing technique that handles the problem of zero probability. Using Laplace smoothing, we can represent

$$p(x_i|C) = \frac{\text{number of reviews with } x_i \text{ and } y = C + \alpha}{N + \alpha * k}$$

Here,

alpha represents the smoothing parameter,
k represents the number of dimension in the data.
N represents the number of review with $y = C$

If we choose a value of alpha != 0, the probability will no longer be zero even if a parameter is not present in the training dataset.

Q.4. What is sigmoid function and its use in logistic regression? Explain it with graph. Explaing the cost & gradient for logistic regression.

Sol

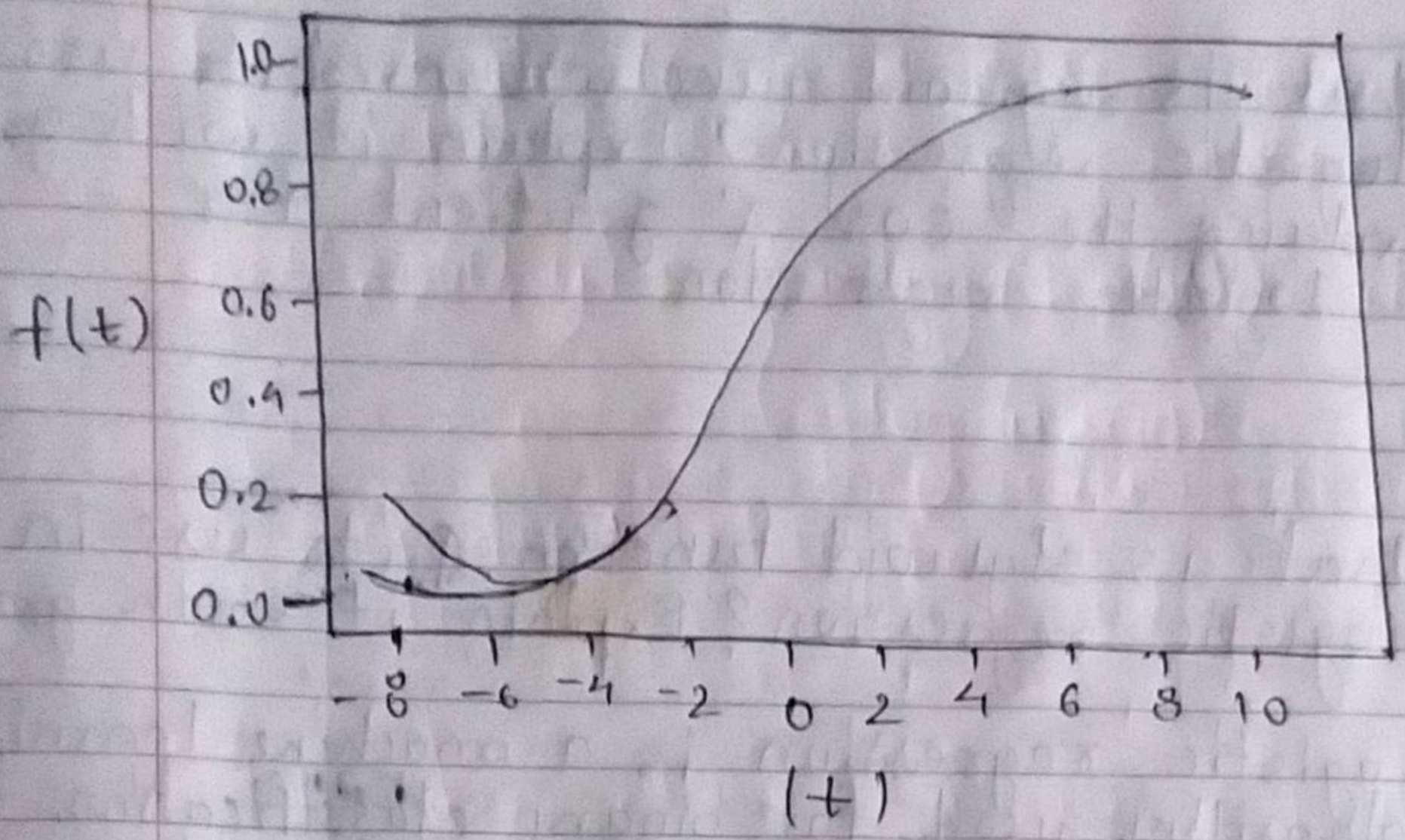
(a) What is sigmoid function & its use in logistic regression? Explain it with graph.

Ans:- Logistic regression is a machine learning algorithm used for binary classification. It predicts the probability of a binary outcomes using a logit function.

Sigmoid function -

It is a mathematical function having a characteristics that can take real value and map it to between 0 to 1 on graph shaped like 'S'. It is a special case of logistic function.

$$y = \frac{1}{1+e^{-z}}$$



Sigmoid function

So, if the value of z goes to positive infinity then the predicted value of y will become 1 and if it goes to negative infinity then the predicted value of y will become 0. Therefore, if the outcome of the sigmoid function is more than 0.5 then we classify that label as class 1 or positive class and if it is less than 0.5 then we can classify it to negative class or label as class 0.

Why do we use the Sigmoid function?

Sigmoid function acts as an activation function and used to add non-linearity to a machine learning model. Its use in logistic

regression is to convert outcomes into categorical value. There are many examples where we can use logistic regression for example, it can be used for fraud detection, spam detection, cancer detection etc!

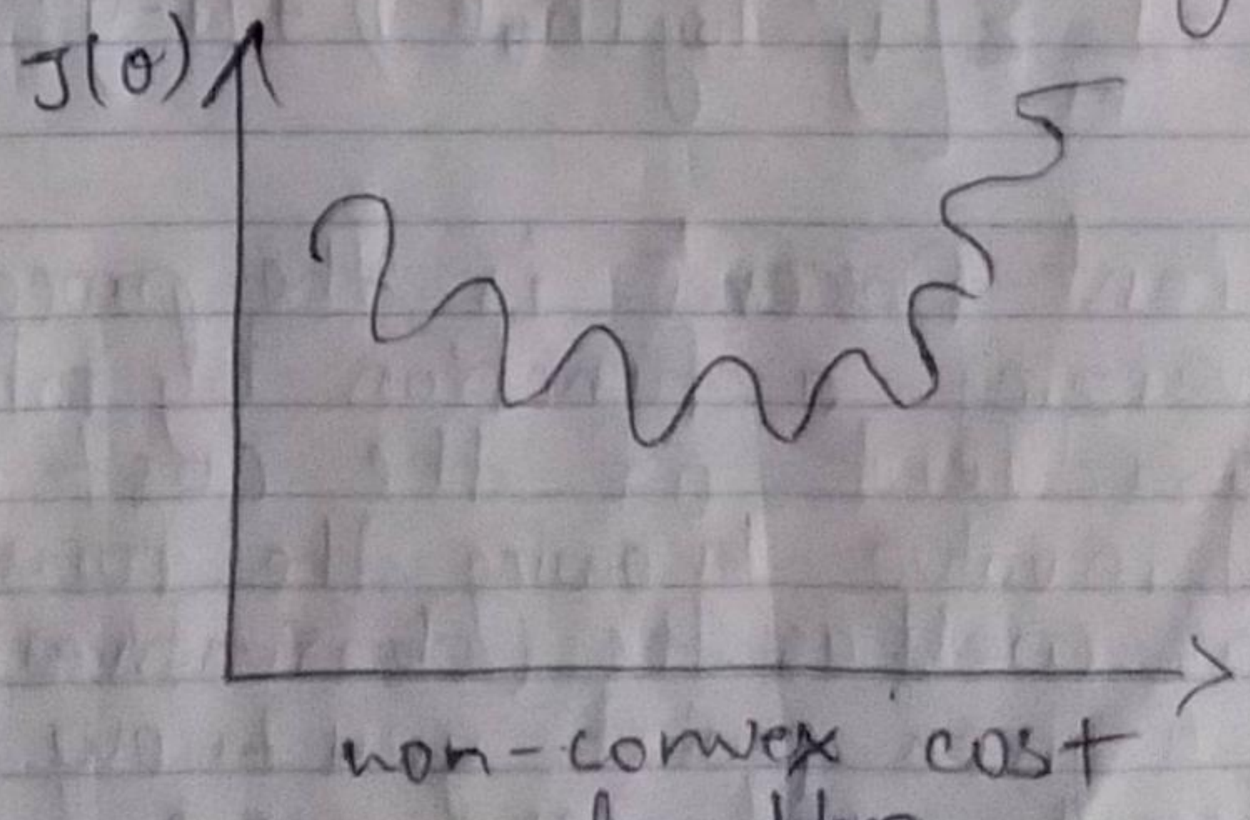
(b) Explain the cost and gradient for logistic regression.

Ans:-

(i) The cost function $J(\theta)$ in the linear regression is:

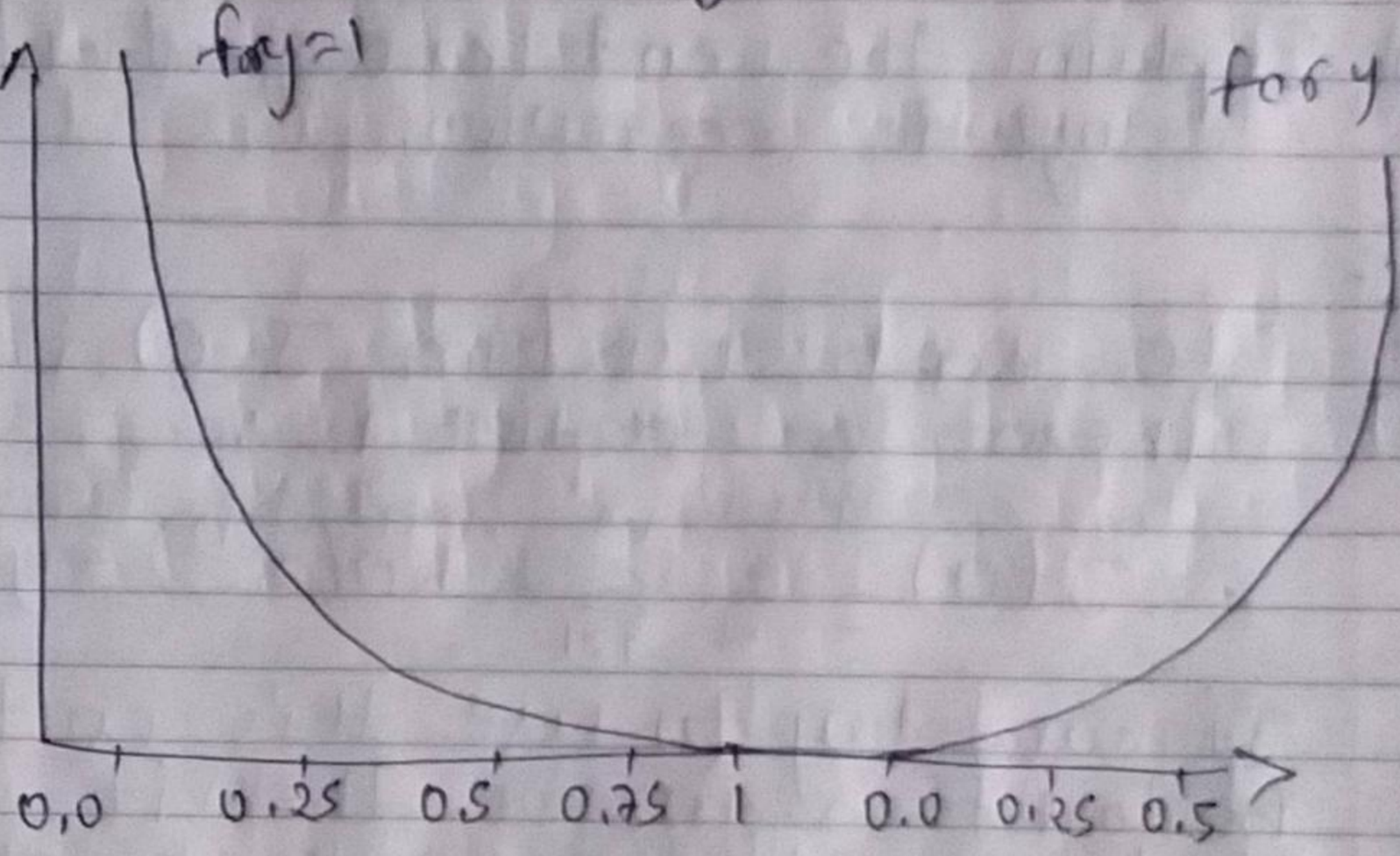
$$J(\theta) = \frac{1}{2} \sum_{i=1}^{m} (h\theta(x^{(i)}) - y^{(i)})^2$$

If we try to use this cost function for logistic regression then it would be end up being non-convex function with many local minima. Hence, it would be difficult to minimize cost value and find global minima.



For logistic regression cost function is defined as:

$$\text{cost}(h_\theta(x), y) = \begin{cases} -\log(h_\theta(x)) & \text{if } y = 1 \\ -\log(1 - h_\theta(x)) & \text{if } y = 0 \end{cases}$$



The above two functions can be compressed into single function i.e.

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

(ii) Gradient Dorent is the process of minimizing a function by following the gradient of the cost function. This involve knowing the form of cost as well as the derivatives so that from a given point you know the gradient and can move in that

direction, e.g. downhill towards the minimum value.

For Graddient Descent we need to minimize cost value i.e. $\min J(\theta)$.

So, we run gradient descent for each parameter:

$$\theta_j = \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta)$$

Want $\min_\theta J(\theta)$

Repeat, $\theta_j := \theta_j - \alpha \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)}) x_j^{(i)}$

Simultaneously update all θ_j