2) a) For a given set of linear Equations —

$$Ax = b$$, where

x = vector
A = matrix and
b = vector.

A and b are the known quantities, whereas x is the unknown for which we find the solution.

If A is a square matrix, then number of equations is equal to number of unknowns and if A is an invertable matrix then solution of x can be found out as $x = A^{-1}b$. Thus in both the above cases calculating x is easier.

But for non square A matrices, calculating inverse is not possible.
So, Pseudo Inverse of a matrix (Moore-Penrose pseudo inverse) helps us to approximately invert the matrix and find a best fit for x that comes closer to solving the equation.

If A is invertable, then pseudo inverse of the matrix (A^+) is equal to A^{-1} and if A is not invertable, then also A^+ is defined.

2) a) i) Underdetermined system of equations :—

here, number of equations are less than no. of unknowns and thus, infinitely many solution will exist for x (where x is the unknown vectos in a given system of linear equation, $Ax = b$). Also A is a short fat matrix and we cannot find its inverse.

There are not enough measurements in b to uniquely determine a single unique solution. We can pick one of these solutions by finding the smallest i.e; we will minimize x.

$$x = A^T (AA^T)^{-1} b \quad \text{where,}$$

$(A^T (AA^T)^{-1})$ is called pseudo-inverse.

ii) Oversdetermined system of equations :—

here, no. of equations are more than the unknowns and we have no solution for x. Here A is a tall matrix and thus we cannot find its inverse.

We try to find an x which will minimize the error between the fit of Ax and b.

$$\|b - Ax\|^2 \Rightarrow x = (A^T A)^{-1} A^T b$$

where $(A^T A)^{-1} A^T$ is called pseudo inverse.

2) b) $x_1 + 3x_2 = 17$
$5x_1 + 7x_2 = 19$
$11x_1 + 13x_2 = 23$

this is an overdetermined system of equations

representing in matrix form —

$$Ax = Y \text{ , where}$$

$$A = \begin{bmatrix} 1 & 3 \\ 5 & 7 \\ 11 & 13 \end{bmatrix}, \quad X = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}, \quad Y = \begin{bmatrix} 17 \\ 19 \\ 23 \end{bmatrix}$$

here A is a non-invertable matrix, so we use pseudo inverse to find out X as follows —

$$X = A^+ Y \quad \text{where,}$$

$$A^+ = \text{pseudo inverse} = (A^T A)^{-1} A^T$$

so $X = (A^T A)^{-1} A^T Y$

$$A = \begin{bmatrix} 1 & 3 \\ 5 & 7 \\ 11 & 13 \end{bmatrix} \qquad A^T = \begin{bmatrix} 1 & 5 & 11 \\ 3 & 7 & 13 \end{bmatrix}$$

$$A^TA = \begin{bmatrix} 1 & 5 & 11 \\ 3 & 7 & 13 \end{bmatrix} \begin{bmatrix} 1 & 3 \\ 5 & 7 \\ 11 & 13 \end{bmatrix}$$

$$= \begin{bmatrix} 1+25+121 & 3+35+143 \\ 3+35+143 & 9+49+169 \end{bmatrix}$$

$$= \begin{bmatrix} 147 & 181 \\ 181 & 227 \end{bmatrix}$$

$$A^TY = \begin{bmatrix} 1 & 5 & 11 \\ 3 & 7 & 13 \end{bmatrix} \begin{bmatrix} 17 \\ 19 \\ 23 \end{bmatrix}$$

$$= \begin{bmatrix} 17+95+253 \\ 51+133+299 \end{bmatrix}$$

$$= \begin{bmatrix} 365 \\ 483 \end{bmatrix}$$

$$(A^TA)^{-1} = \frac{\text{adj}(A^TA)}{|A^TA|}$$

$$\text{adj}(A^TA) = \begin{bmatrix} 227 & -181 \\ -181 & 147 \end{bmatrix}$$

$$|A^T A| = \begin{vmatrix} 147 & 181 \\ 181 & 227 \end{vmatrix}$$

$$= 147 \times 227 - 181 \times 181$$

$$= 608$$

$$(A^T A)^{-1} = \begin{bmatrix} 227/608 & -181/608 \\ -181/608 & 147/608 \end{bmatrix}$$

multiplying, $(A^T A)^{-1}$ with $A^T y$ to get x

$$= \begin{bmatrix} 227/608 & -181/608 \\ -181/608 & 147/608 \end{bmatrix} \begin{bmatrix} 365 \\ 483 \end{bmatrix}$$

$$= \begin{bmatrix} -7.51 \\ 8.12 \end{bmatrix}$$

so,

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} -7.51 \\ 8.12 \end{bmatrix}$$ ans

2)c)i) The estimated model of linear regression can be written as —

$$Y = X\theta + \varepsilon \cdot$$, where

if m = number of rows
n = number of columns then

~~Y is a~~ $m \times 1$ Y is a matrix of size $m \times 1$
Y is a $(m \times 1)$ matrix of target variables,
X is a $(m \times n)$ matrix of features,
θ is a $(n \times 1)$ matrix of parameters and
ε is a $(m \times 1)$ matrix of residuals.

$$\varepsilon = Y - X\theta$$

to find the least square estimator, we write sum of square of residuals as follows —

$$S(\theta) = \sum \varepsilon_i^2 = \varepsilon^T \varepsilon = (Y - X\theta)^T (Y - X\theta)$$

$$= Y^T Y - Y^T X\theta - \theta^T X^T Y + \theta^T X^T X\theta$$

minimizing $S(\theta)$ is setting derivative, to 0, equal

$$-2X^T Y + 2X^T X\theta = 0$$

$$\Rightarrow \theta = (X^T X)^{-1} X^T Y$$; this is the normal equation.

2) c) ii) We prefer iterative methods like gradient Descent rather than using closed form solutions because gradient descent is computationally cheaper. For example, if number of parameters is 10000 then we have to compute 10000^2 values for X^TX in closed form equation, which becomes very expensive to compute parameters.

2)d) ii) **RMSE :-**

RMSE stands for root mean squared error.

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2} \quad \text{where}$$

N = total number of data points
$\hat{y}_i$ = predicted target value
y_i = actual target value

It measures the distance between the predicted and actual values i.e the residuals. It measures the standard deviation of the residuals i.e how spread out the residuals are or how concentrated the data is around the line of best fit. gives high weightage to large errors hence should be used when large errors are undesirable. Lower value means better fit of the model. Ranges from 0 to ∞.

MAE :-

MAE stands for mean absolute error.

$$MSE = \frac{1}{N}\sum_{j=1}^{N}|y_j - \hat{y}_j| \quad \text{where}$$

N = total number of data points
$\hat{y}_j$ = predicted target value
y_j = actual target value

Its value ranges from 0 to ∞

It measures absolute average distance between real data and predicted data. Here, all errors are given equal weightage; MAE fails to give weightage to large errors. Hence must be used when large errors are not present.

MSE : —

MSE stands for mean squared error.

~~MSE =~~ $\sqrt{\quad}$

$$MSE = \frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 \text{ where}$$

N = total number of data points
y_i = actual target value
$\hat{y}_i$ = predicted target value

It ranges between 0 to ∞. MSE puts large weightage on outlier errors and squaring is done it magnifies the error. Hence must be used when the model has less or no outliers. However if we want to ignore outliers then MSE will be a bad choice.

~~3)c) tanh x is defined as~~

3)c) tanh x is defined as →

$$\tanh x = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}} = g(x)$$

let $h_\theta(x) = g(\theta^T x) = \frac{e^{\theta^T x} - e^{-\theta^T x}}{e^{\theta^T x} + e^{-\theta^T x}}$

where θ = parameters of the model
x = independant features of the model

calculating the derivative of tanh(z) →

$$\frac{d}{dz}(g(z)) = \frac{d}{dz}\left(\frac{e^{z} - e^{-z}}{e^{z} + e^{-z}}\right)$$

$$\Rightarrow \frac{(e^{z} + e^{-z})(e^{z} + e^{-z}) - (e^{z} - e^{-z})(e^{z} - e^{-z})}{(e^{z} + e^{-z})^2}$$

$$\Rightarrow 1 - \frac{(e^{z} - e^{-z})^2}{(e^{z} + e^{-z})^2}$$

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

$$\Rightarrow 1 - \tanh(z)^2$$

$$\Rightarrow 1 - (g(z))^2 \quad \ldots (1)$$

in logistic regression, we predict the probability of any of the class, the probability of one class can be calculate with the 1st one.

$$P(y=1 \mid x;\theta) = h_\theta(x) = \frac{e^{\theta^T x} - e^{-\theta^T x}}{e^{\theta^T x} + e^{-\theta^T x}}$$

then $P(y=0 \mid x;\theta) = 1 - P(y=1 \mid x;\theta)$

$$= 1 - \left(\frac{e^{\theta^T x} - e^{-\theta^T x}}{e^{\theta^T x} + e^{-\theta^T x}} \right)$$

combining the two equations —

$$P(y \mid x;\theta) = (h_\theta(x))^y \, (1 - h_\theta(x))^{1-y}$$

probability of y given x parameterised by θ

$L(\theta)$ = likelihood of θ ~~$= L(\theta;$~~

maximizing $L(\theta)$ so that data points best fit the model

$$\max(L(\theta)) = \max \left(P(\vec{y} \mid x;\theta) \right)$$

[m = no. of instances] $$= \prod_{i=1}^{m} P(y^i \mid x^i;\theta)$$

$$= \prod_{i=1}^{m} h_\theta(x^i)^{y^i} - (1 - h_\theta(x^i))^{1-y^i}$$

maximising $L(\theta)$ = maximising $l(\theta)$ where $l(\theta)$ = log likelihood function thus

$$\max l(\theta) = \max \log L(\theta) =$$

$$\sum_{i=1}^{m} \log (h_\theta(x^i))^{y^i} + \sum_{i=1}^{m} \log (1 - h_\theta(x^i))^{1-y^i}$$

here, $\theta = \theta + \alpha \nabla_\theta l(\theta)$, we are maximising $l(\theta)$, i.e gradient ascent.

$$\nabla_{\theta_j} \cdot l(\theta_j)$$

$$= \frac{\partial}{\partial \theta_j} \left(y^i \log (h_\theta(x^i)) \right) + (1-y^i) \log (1 - h_\theta(x^i))$$

$$= \frac{y^i}{h_\theta(x^i)} \frac{\partial}{\partial \theta_j} h_\theta(x^i) - \frac{1-y^i}{(1-h_\theta(x^i))} \frac{\partial (h_\theta(x^i))}{\partial \theta_j}$$

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

$$= \left(\frac{y^i}{h_\theta(x^i)} - \frac{1-y^i}{1-h_\theta(x^i)} \right) \left[1 - h_\theta(x^i)^2 \right] \times x_j^i$$

$$= \frac{y^i - y^i h_\theta(x^i) - h_\theta(x^i) + h_\theta(x^i) y^i}{h_\theta(x^i) \times (1 - h_\theta(x^i))} \times \left[1 - h_\theta(x^i)^2 \right] \times x_j^i$$

$$= \frac{y^i - h_\theta(x^i)}{h_\theta(x^i) * (1 - h_\theta(x^i))} * (1 + h_\theta(x^i))(1 - h_\theta(x^i)) * x_j^i$$

$$= \frac{y^i - h_\theta(x^i)}{h_\theta(x^i)} * (1 + h_\theta(x^i)) * x_j^i$$

* thus the update rule for logistic regression =

$$\theta_j = \theta_j + \alpha\left(\frac{y^i - h_\theta(x^i)}{h_\theta(x^i)} * x_j^i * (1 + h_\theta(x^i))\right)$$

→ update rule for logistic regression using tanh function as the decision boundary