1)

$$\hat{y}_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} \cdots \beta_d x_{id}$$

$$J(\beta) = \frac{1}{2} \|y - X\beta\|^2 = \frac{1}{2} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

a) Here regression is finding a relation b/w input variable x & target y, it is done by fitting a best fit 2D st. line or 3D plane Goal is to find the set of β.

why minimizing squared errors? Because it is easy to compute derivatives and it is smooth parabola. Also, this makes to prioritize large errors to bring them closer.

b) $$J(\beta) = \frac{1}{2} (y - X\beta)^T (y - X\beta)$$

$$= \frac{1}{2} (y^T - \beta^T X^T)(y - X\beta)$$

$$\frac{1}{2} (y y^T - y^T X\beta - \beta^T X^T y + X X^T \beta \beta^T)$$

since both are scalar. 1x1

$$J(\beta) = \frac{1}{2} (y y^T - 2 y^T X\beta + X X^T \beta \beta^T)$$

$$\frac{\partial J}{\partial \beta} = -2 y^T X + 2 X X^T \beta = 0$$

$$X y^T = X X^T \beta = X^T y$$

$\Rightarrow \beta = (X^T X)^{-1} X^T y$ —1.1—

c) Finding inverse is very computationally expensive. Also ~~the~~ meeting the conditions for invertibility also becomes difficult.
iterative method → It is easy because the time it would take would be n× one step. much less than the direct one.

(2) a) Core idea of backpropagation :→
i) prediction using forward pass
ii) calculating how error changes with prediction (backpropagation)
iii) error propagation
iv) update weight.

b) $\frac{\partial L}{\partial w_2} = \frac{\partial L}{\partial z_2} \frac{\partial z_2}{\partial w_2}$

$= (a_2 - y)(a_1)$

$\frac{\partial L}{\partial b_2} = (a_2 - y)(1)$

$$\frac{\partial L}{\partial a_1} = \frac{\partial L}{\partial z_2} \frac{\partial z_2}{\partial a_1}$$

$$\frac{\partial L}{\partial z_1} = \frac{\partial L}{\partial a_1} \frac{\partial a_1}{\partial z_1}$$

$$\Rightarrow \frac{\partial L}{\partial w_1} = \frac{\partial L}{\partial z_1} \cdot \frac{\partial z_1}{\partial w_1}$$

$$= (a_2 - y) w_2 a_1 (1 - a_1) x$$

$$\frac{\partial L}{\partial b_1} = \frac{\partial L}{\partial z_1} \frac{\partial z_1}{\partial b_1} = (a_2 - y) w_2 a_1 (1 - a_1)$$

c) Gradient descent

$$w_1 \rightarrow w_1 - \eta \frac{\partial L}{\partial w_1} \qquad \eta \rightarrow \text{learning rate}$$

$$b_1 \rightarrow b_1 - \eta \frac{\partial L}{\partial b_1}$$

$$w_2 \rightarrow w_2 - \eta \frac{\partial L}{\partial w_2}$$

$$b_2 \rightarrow b_2 - \eta \frac{\partial L}{\partial b_2}$$

__/__/__

3) a) ANN vs RNN

ANN → processes input independently. They do not need to maintain the previous inputs.

RNN → processes inputs sequentially They have to save the previous memory also.

b) RNN struggle with long term dependencies due to vanishing of gradient as gradients shrink exponentially during backpropagation through time so it makes early information too small to learn from.

c) Forget gate :→ decides the irrevalent info.
Input Gate :→ decide which new info. is to be taken as input
Output Gate :→ decides the result that's passed on to next layer.

__/__/____

d) Using gated cells.

e) ANN → Image classification
RNN → Speech Recognition
LSTM → language Translation

4) a) The flights that passengers booked with Indigo were affected.
were → flights
Yes, RNN would struggle due to vanishing gradient problem.

b) Suppose the task is to process a paragraph so in that after full stop it is important to wipe out the previous memory for model to work effectively.