ILIKA BERA
240462

ASSIGNMENT-1

1a) Regression models or aims to find a relationship between a set of independant & dependent variables.

We minimise squared error because:-

- It penalises larger errors more than smaller errors because ~~when we~~ after squaring, large errors becomes larger.
- Squaring removes negative signs.
- It brings the predicted values closer to the real values.

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

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

$$\nabla_\beta(J(\beta)) = -X^T(y - X\beta)$$

$$-X^T(y - X\beta) = 0$$

$$\Rightarrow X^TX\beta = X^Ty$$

$$\therefore \boxed{\beta = (X^TX)^{-1}X^Ty}$$

(c)
- Inversion leads to $O(d^3)$ term → large calculation and computational time
- X^TX might be singular matrix ⇒ $(X^TX)^{-1}$ won't exist
- Would take up ~~large~~ a lot of memory for large datasets.

Iterative methods are used —
- well adjusted to data-sets.
- works even if (X^TX) is singular / non invertible.
- No need for inverse calculation.

2. a) Backpropagation is efficient —
- Application of chain rule.
- Reusing derivatives obtained in intermediate steps.
- Error signals are taken from output to input layers.

b) $z_1 = w_1 x + b_1$, $a_1 = \sigma(z_1)$

$z_2 = w_2 a_1 + b_2$ $a_2 = \sigma(z_2)$

$$\frac{\partial L}{\partial z_2} = a_2 - y.$$

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

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

$$\frac{\partial L}{\partial a_1} = (a_2 - y) \cdot w_2$$

$\sigma'(z_1) = a_1(1 - a_1)$.

$$\frac{\partial L}{\partial w_1} = (a_2 - y) w_2 a_1 (1 - a_1) x.$$

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

c) $w \leftarrow w - \eta \frac{\partial L}{\partial w}$

$b \leftarrow b - \eta \frac{\partial L}{\partial b}$.

η controls — step size of updates.

Too large — divergence

Too small — slow convergence.

3. a) ANN w/s RNN

ANN: Processes inputs independently

RNN: Processes sequences & maintains a hidden state.

b) RNNs struggle —
- Gradients explode/vanish over large sequence
- Earlier data is forgotten.

c) Roles in LSTM -

- Input gate - Controls writing to memory
- Forget gate - Controls erasing memory
- Output gate - Controls reading from memory.

d) How LSTMs address vanishing gradients -

- Memory cell allows near constant gradient flow
- Gate regulate information & don't overwrite it.

e) Example tasks :-

ANN : house price prediction
RNN : speech recognition
LTSM : machine translation.

4. a) "The book that the professor who the students admired wrote (was) published"

'Was' prediction depends on book not on students.
RNN would struggle due to vanishing gradients.

b) Gates decide :-
- which data to keep, forget & output

Eg :, When translating a long sentence, once a clause ends, the forget gate should be near 0 to erase irrelevant context before processing the next clause.