MA15042

DATE PAGE

Ans - 1.

Q1) (a) Regression builds relationship between input features and output. It aims to predict the output using linear relationship $y = XB + \epsilon$ in case of linear regression.

- Due to squaring, there is more impact on residuals.
- Cost function becomes smooth function, which ensures a global minimum point.

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

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

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

$\frac{\partial J}{\partial B} = 0 \rightarrow \nabla J(B) = X^T XB - X^T y$

$\left(\begin{array}{l} \nabla (y^T XB) = X^T y \\ \nabla (B^T X^T XB) = 2X^T XB \end{array} \right) \rightarrow \nabla J(B) = 0$

$$\boxed{B = (X^T X)^{-1} X^T y}$$

(c) Direct inversion involves ~~calculation~~ high computational cost calculations like calculating $X^T X$ and then inverse of it, so computationally expensive.

Gradient descent (iterative method) is preferred as it minimises cost function iteratively which is good for real world datasets. Here no matrix inversions are required.

DATE___PAGE___

Q2.) Backpropogation is an algorith which iteratively
(a) updates weights and biases through gradient descent to minimise the cost function.

As neural network is composition of functions using chain rule we first obtain the gradient of loss from output layer the moving backwards layer by layer.

(b) $L = -\left(y \log(a_2) + (1-y)\log(1-a_2)\right)$

$$\frac{\partial L}{\partial w_2} = \frac{\partial L}{\partial a_2} \times \frac{\partial a_2}{\partial z_2} \times \frac{\partial z_2}{\partial w_2} \quad \left(z_2 = w_2 a_1 + b_2\right)$$

$$= (a_2 - y) \times (a_1)$$

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

$$\frac{\partial L}{\partial b_2} = \frac{\partial L}{\partial z_2} \times \frac{\partial z_2}{\partial b_2} = (a_2 - y)(1)$$

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

$$\frac{\partial L}{\partial w_1} = \frac{\partial L}{\partial z_2} \times \frac{\partial z_2}{\partial a_1} \times \frac{\partial a_1}{\partial z_1} \times \frac{\partial z_1}{\partial w_1}$$

$$\frac{\partial L}{\partial z_2} = a_2 - y, \quad \frac{\partial z_2}{\partial a_1} = w_2, \quad \frac{\partial z_1}{\partial w_1} = x$$

$$\frac{\partial a_1}{\partial z_1} = \frac{1}{(1+e^{-z})} \times \frac{e^{-z}}{1+e^{-z}} = a_1(1-a_1)$$

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

DATE____PAGE____

$$\frac{\partial L}{\partial b_1} = \frac{\partial L}{\partial z_2} \times \frac{\partial z_2}{\partial a_1} \times \frac{\partial a_1}{\partial z_1} \times \frac{\partial z_1}{\partial b_1}$$

$$\left(\frac{\partial z_1}{\partial b_1} = 1\right)$$

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

(c) Using Backpropogation, I would start updating parameters of network from output layer iteratively.

$$\boxed{w_n = w_n - \eta \frac{\partial L}{\partial w_n}}$$

$$\boxed{B_n = B_n - \eta \frac{\partial L}{\partial B_n}}$$

$$w_1 = w_1 - \eta (a_2 - y) w_2 \times a_1 (1 - a_1)$$

$$w_2 = w_2 - \eta a_1 (a_2 - y)$$

$$b_1 = b_1 - \eta (a_2 - y) w_2 a_1 (1 - a_1)$$

$$b_2 = b_2 - \eta (a_2 - y)$$

η → learning rate : It is the step size for updation of parameters

If it is too large, we may skip minimum Loss function value and if too small, we will require very long time for convergence.

Proper and optimal choice of η should be taken for good learning.

DATE PAGE

Q3) (a) ANN processes inputs in independent forward as there is no memory of previous inputs, and output is obtained from present input.

RNN has ability to retain memory of previous time step. It has hidden state that transfers information from previous to next steps sequentially. output depends on current input and hidden state.

(b) RNN face problem of vanishing gradient during BTT (Backpropagation through time). Also the gradient terms are multiplied so it gets shrink slowly so RNN looses information from previous steps.

(c) LSTMs are more concise than RNN as it has 3 gates as forget gate : It looses unnecessary information
input gate : decides what to store
output gate : decides result which is passed in next layer
So LSTM are preferred in long term sequences.

(d) They have cell state.

(e) ANN : Image recognition
RNN : Speech recognition
LSTM : Language Translation

Q4) (a) "When mobile dropped on floor, it broke down"

Here, to understand "it", model must remember that even though "floor" is close to "it", but "it" resembles here to the "mobile". And these 2 words have a large gap between them, so model must have long-long memory to remember.

DATE___ PAGE___

In this case RNN would struggle as RNN have short term memory ~~and they are~~ due to vanishing gradient problem during backpropagation.

The gradients shrinks down gradually and model loses memory.

(b) LSTM has long term memory due to its cell state which carries information without significant change in it through many steps.

Gating mechanism properly ensures relevant information to be transferred.
Forget → discard irrelevant information
Input → decides what to be stored
Output → decides what to be passed on

In LSTM, due to cell state, gradients do not shrink, the information is carried over many steps.

If forget gate close to 0,
the model must forget information of the previous step.

Eg: "Yesterday, I went to hotel. Today, I am going to my home".

Here model should forget information of previous day and focus on new day (Today).