# Assignment 1

## Solution 1

(a) Linear Regression assumes a straight line relationship between the variables. The goal is to find the best fit line that minimizes the difference between predicted value and the actual observed data points.

Once the line is fitted, we can predict the value of dependent variable(Y) for some independent variable(X).

The squared error needs to be minimized because its the difference between the actual observed value and the predicted value by the line. Also since it is squared, large errors are penalized disproportionately more than small errors. Like an error of 10 units contributes 100 to Sum of squered errors while an error of 1 unit contributes 1.

(b) The cost function → $J(\beta) = \frac{1}{2} |Y - X\beta|^2$

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

gradient wrt $\beta$ →

$$\nabla_\beta J(\beta) = -X^T (Y - X\beta) = 0$$

$$X\beta X^T = X^T Y$$

$$\text{so } \beta = (XX^T)^{-1} X^T Y$$

(c) Direct inversion of $XX^T$ is problematic becoz $(X^T X)^{-1}$ has time complexity of $O(p^3)$, and the space complexity is $O(p^2)$ which are very high.

Also when features are linearly dependent, $X^TX$ is non-invertible.

Iterative methods being →

Avoids matrix inversion, scale well for large datasets and can be stopped early for approximate solutions

2 (a)

In Backpropagation, each neuron's output depends on earlier parameters. The chain rule allows gradients to be reused layer by layer, avoiding redundant calculations.

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

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

so $\mathcal{L} = -[y \log a_2 + (1-y)\log(1-a_2)]$

$\frac{\partial \mathcal{L}}{\partial z_2} = a_2 - y$ so $\frac{\partial \mathcal{L}}{\partial w_2} = (a_2 - y)a_1$

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

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

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

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

(c) $\theta \leftarrow \theta - \eta \frac{\partial \mathcal{L}}{\partial \theta}$ The learning rate ($\eta$) is the step size if it too low, the convergence is slow, if too large, then unstable.

PAGE No.
DATE

3

(a) An ANN processes a fixed size input as a single, static unit and its output depends only on that current input. An RNN processes a sequence piece by piece iteratively using a hidden state to pass info.

(b) RNNs struggle with RNNs with long term dependencies due to vanishing gradient problem that occurs during training via Backpropagation through time.

(c) Role of LSTM is to selectively regulate the flow of info into and out of the cell state which acts as network's long term memory.
- Preserves info by establishing a conveyor belt mechanism.

(d) They solve it using additive cell state mechanism which acts as a memory conveyor belt allowing error signals to flow across many steps with minimal decay.

(e) - ANN - Handwritting Recognition
RNN - Simple Sentiment Analysis
LSTM - Machine Translation