Assignment 1

Anuska keshri

240165

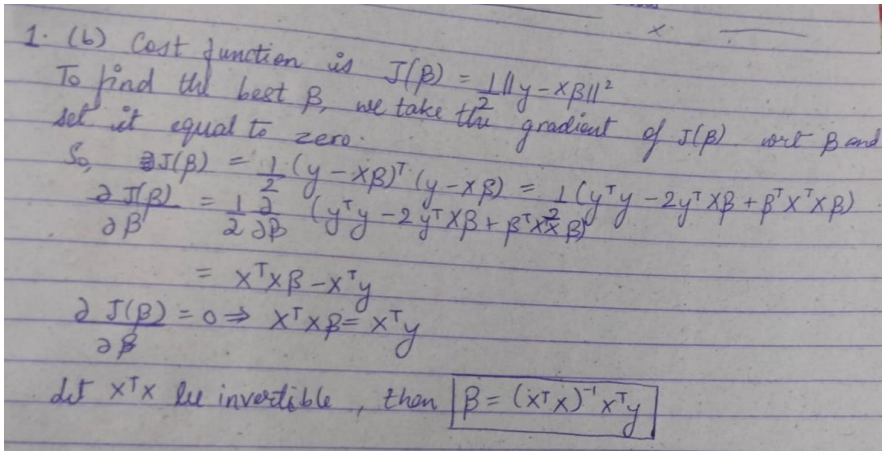
1.

(a) Regression is used to predict a continuous value (like price, marks, temperature) using input features. In linear regression, we assume that the output is a linear combination of the input features and an intercept. The goal is to choose parameters beta such that the predicted values are as close as possible to the true values.

We minimize squared error because:

- It measures how far predictions are from actual values.
- Larger mistakes are penalized more than small ones.
- The cost function becomes smooth and convex, so it has one best solution.
- It makes the math easier and gives a closed-form solution.

(b)



(c) Directly computing XTX can be a problem because:

- Matrix inversion is slow when the number of features is large.
- If features are highly correlated, it may not be invertible.
- It requires a lot of memory for large datasets.

Because of this, iterative methods like gradient descent are preferred:

- They work well for large datasets.
- They avoid matrix inversion.
- They are more stable and efficient.

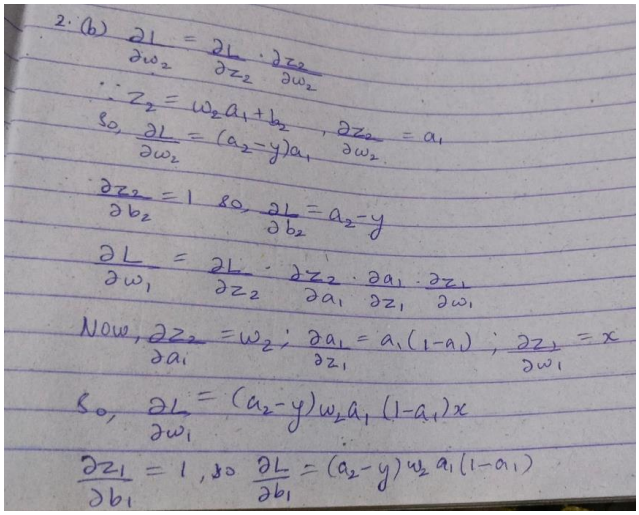
2.

(a) Backpropagation is an algorithm used to train neural networks.
Its main idea is to compute how much each weight and bias contributes to the final error, and then adjust them to reduce the error.

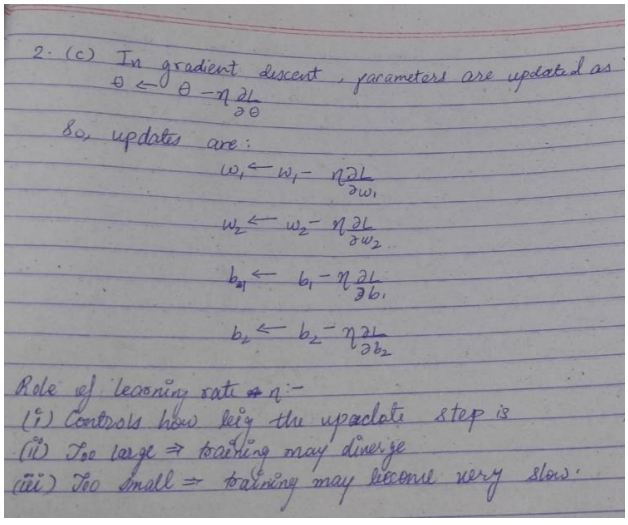
The chain rule from calculus allows us to break a complicated function into smaller parts.
Since a neural network is made of many layers connected together, the chain rule helps us:

- compute gradients layer by layer
- reuse intermediate results
- efficiently update all parameters

(b)



(c)



3.

(a) ANN (Artificial Neural Network)

- Takes fixed-size inputs
- Each input is processed independently
- No memory of past inputs
- Example: predicting house price from features

RNN (Recurrent Neural Network)

- Designed for sequential data
- Processes input one step at a time
- Keeps information from previous steps using a hidden state
- Example: reading a sentence word by word

(b) In long sequences, information from earlier steps becomes very weak due to which gradients become very small (vanishing gradients) during backpropagation and as a result, RNNs forget old information. This makes learning long-range relationships difficult.

(c) An LSTM has three gates that control the flow of information:

1. Forget gate
    - Decides what information to remove from memory
    - If something is not useful anymore, it is forgotten
2. Input gate
    - Decides what new information should be stored in memory
    - Allows important new information to enter the memory cell
3. Output gate
    - Decides what information to send to the next step/output
    - Controls what part of the memory is revealed

Gates in LSTMs control what information to forget, store, and output, allowing the network to preserve important information over long sequences.

(d) LSTMs have a memory cell that allows information to flow unchanged which creates a stable gradient path. As a result, gradients do not vanish easily. This helps LSTMs learn long-term dependencies.

(e) ANN- House price prediction

RNN- Simple text or speech sequences

LSTM- Language translation

4.

(a) Example sentence:

"The book that I bought last year from the store **is** interesting."

- The verb "is" depends on "book", not "store"
- The words are far apart implies long-range dependency
- A standard RNN would likely struggle to remember "book" until the end

Yes, a standard RNN would struggle due to vanishing gradients.

(b) The memory cell stores important information for a long time. Gates decide: what to keep, what to forget and what to output.

Machine Translation Example:

Sentence:

“I went to the market yesterday because I needed milk.”

- The time word “yesterday” must be remembered until the verb is translated
- The forget gate stays close to 1 to keep this information
- This helps produce the correct tense in the translated sentence