NAME: Vikram Nanda Victor

STUDENT ID: 6081050

Directions:

- ❑ This is an open-book, open-note and open-internet exam.
- ❑ You must answer the questions in your own words.
- ❑ Your answers should be brief, concise and readable.
- ❑ You have 120 minutes to complete the exam. Good luck!

1. (3 pts) Name **THREE** major factors that contribute to the recent success of deep learning.
   - Data → Due to increase in the availibility of data we have alot of it
   - Computation Power → Advancement of technology help shortern process time of modeling.
   - Algorithm → Better Algorithm have been found due to more people are in the field

2. (3 pts) True or False:
   a. F Deep learning works well even when we have a small training set.
   b. T Training a deep network requires a lot of computational power.
   c. T A deep network is basically a neural network with many layers.

3. (3 pts) Consider a neural network that outputs $\hat{y} = \{0,1\}$ and was trained on a dataset $\{(x_1,y_1),(x_2,y_2),\ldots,(x_n,y_n)\}$. Can you suggest **a loss function** and **a cost function** that we can use for this network? You must write out the functions mathematically using notations provided in the context.

Loss → $l(y',y) = -(y\log(y') + ((1-y)\log(1-y'))$

Cost → $J(w,b) = \frac{1}{m} \cdot \sum_{i=1}^{n} [l(y'_i, y_i)]$

4. (3 pts) What is vectorization and why is it important in deep learning?

Vectorization: is a way to optimize code to remove for loop and help execution time

it important to DL cause our data set is very big

with vectorization we can reduce the computation time greatly.

5. (3 pts) Write down **THREE** activation functions that are commonly used in deep networks, along with their derivatives.

- Sigmoid: $g(z) = \frac{1}{1+e^{-z}}$, $g'(z) = g(z) \cdot (1 - g(z))$
- Tanh: $g(z) = \frac{e^{z} - e^{-z}}{e^{z} + e^{-z}}$, $g'(x) = 1 - g(z)^2$
- RELU: $g(z) = \max(0, z)$, $g'(x) = \begin{cases} 0 \text{ if } z < 0 \\ 1 \text{ if } z \geq 0 \end{cases}$

6. (3 pts) Consider the following function.

$$f(x) = \begin{cases} 0 \text{ if } x < 0 \\ 1 \text{ otherwise} \end{cases}$$

Can we use this function as an activation function? If not, explain why not

Yes Cause RELU

Can't cause we don't have a back ward prop. or derivatives of it.

7. (3 pts) Why do we need non-linear activation functions in a neural network?

- What ever the activation func. is the oupt will also be linear if we use linear activation func.
- If the output is a real number it's fine but if not it will cause problem; If output is even negative RELU is better.

8. (3 pts) What is the learning rate? Where do we use it?

Learning rate is the rate which determind how big a gap is to loss function min.

we use learnig rate as a hyperparameters ~~case~~ it is used in optimizing loss function.

9. (3 pts) In logistic regression, we can initialize all the weights with zeros but, in neural network, we cannot do so. Why?

- We will be computing the same function in all units.
- The update in the gradient decent will be the same

10. (3 pts) Explain why it is not recommended to initialize weights of a neural network with large values (large positive or small negative).

Case for tanh and sigmoid if the weight is to big it will slow down the computation time it will be saturated.

11. (3 pts) Explain the difference between train set, dev set, and test set.

Train set : Data train model

Dev set : Data to check and train for adjustment for overfitting and underfitting.

Test set : Set of Data we test how good is our final model after training and config with dev set.

12. (6 pts) Consider the following situations. Indicate whether it is overfitting, underfitting or something else. Also, suggest what to do next to improve performance.

a. Bayes error: 2%, Training error: 10%, Test error: 12%

Choose one or write your answer: (overfitting, underfitting, _____)
Your suggestion:

High Bi

b. Bayes error: 5%, Training error: 1%, Test error 15%

Choose one or write your answer: (overfitting, underfitting, _____)
Your suggestion:

c. Bayes error: 7%, Training error: 6%, Dev error: 7%, Test error: 20%

Choose one or write your answer: (overfitting, underfitting, _____)
Your suggestion:

High Bias Overfitting

13. (3 pts) Explain what is vanishing/exploding gradient problem? How can we avoid it?

- Vanishing Gradient : derivatives are too small
  happen when $W^{[L]} < I$ (Identity matrix)
- Exploding Gradient : drivative are too big
  happen when $W^{[L]} > I$ (Identity matrix)
- Best current way to slove it to random Initial weights.
  or RELU + Weight with variance.

14. (3 pts) What is dropout? How does it help regularizing a network?

Disable / kicking out a node by checking from "keep-prob"
this help overfitting greatly as it reduce the model
computation power.

15. (4 pts) How many weight updates the following algorithms perform on ONE pass over the same training data? Assume 100 training examples.

100/1 a. 100 Stochastic Gradient Descent
100/100 b. 1 Batch Gradient Descent
100/10 c. 10 Mini-Batch Gradient Descent (batch size = 10)
100/50 d. 2 Mini-Batch Gradient Descent (batch size = 50)

16. (3 pts) Explain how learning rate decay could help in the training process.

- It reduce chance of reaching optimum point hence it help gradient descent func.

17. (3 pts) What is Batch Normalization? What are the benefits of batch normalization?

Batch Norm. : Form of normalization where you get mean and variance of mini-batch and subtract it of

Pro
- Reduce input shifting
- Regulize the data a bit
- help compute time