CS 440 Exam

Gordon Ng

TOTAL POINTS

32.5 / 48

QUESTION 1

1 Q1 7 / 10

✓ **- 0 pts** (a) correct

- 1 pts (a) partially correct

- 2 pts (a) incorrect

✓ **- 0 pts** (b) correct

- 1 pts (b) partially correct

- 2 pts (b) incorrect

- 0 pts (c) correct

- 1 pts (c) incorrect advantage

- 1 pts (c) incorrect disadvantage

✓ **- 1 pts** Advantage and/or disadvantage explanations require some more information

- 2 pts (c) incorrect

- 0 pts (d) correct

- 1 pts (d) incomplete explanation

✓ **- 2 pts** (d) incorrect

✓ **- 0 pts** (e) correct

- 1 pts (e) partially correct

- 2 pts (e) incorrect

QUESTION 2

2 Q2 8 / 8

✓ **- 0 pts** Correct

- 2 pts a) partial correct

- 4 pts a) incorrect

- 2 pts b) incorrect

- 2 pts c) incorrect

- 1 pts c) partial incorrect

- 1 pts b) partial incorrect

QUESTION 3

3 Q3 7.5 / 12

a(i)

✓ **- 0 pts** Correct

- 1 pts Partially incorrect

- 2 pts Incorrect

a(ii)

✓ **- 0 pts** Correct

- 1 pts Partially incorrect

- 2 pts Incorrect

a(iii)

- 0 pts Correct

- 0.5 pts Plus instead of minus

- 0.5 pts Missing learning rate

- 1 pts Incorrect gradient

✓ **- 0.5 pts** Another type of small mistake

- 2 pts Major mistake

b(i)

- 0 pts Correct

✓ **- 0.5 pts** Incorrect

b(ii)

- 0 pts Correct

✓ **- 0.5 pts** Partially incorrect: m = number of data samples

✓ **- 0.5 pts** Partially incorrect: y = ground truth labels

✓ **- 0.5 pts** Partially incorrect: h(x) = model prediction from data sample

c

- 0 pts Correct

✓ **- 1 pts** Not a correct example.

✓ **- 1 pts** Not a satisfactory explanation.

d

✓ **- 0 pts** Correct

- 1 pts Partially incorrect: one application missing/incorrect.

- 2 pts Incorrect: two applications missing/incorrect.

💬 J parameterized by w

QUESTION 4

4 Q4 7 / 10

- **0 pts** Correct

(a)

✓ - **0 pts** **Correct**

- **1 pts** A is incorrect

- **1 pts** B is incorrect

- **1 pts** C is incorrect

(b)

✓ - **0 pts** **Correct**

- **1 pts** No/incorrect axis labeling

- **1 pts** Incorrect stopping point

- **1 pts** Training/validation curve missing or incorrect

- **0.5 pts** Validation error not increasing after stopping point

- **0.5 pts** unclear which is training and which is validation curve

(c)

- **0 pts** Correct

✓ - **1 pts** **Dropout not named**

✓ - **1 pts** **No/incorrect description of how Dropout achieves such regularization**

- **1 pts** Partial credit for other regularization technique not specifically designed for deep neural networks, e.g. L1/L2 etc.

- **0.5 pts** Incomplete description of how Dropout achieves such regularization

(d)

- **0 pts** Correct

✓ - **1 pts** **Missing row and column labels to indicate which is the *Actual* and which is the *Predicted*.**

- **1 pts** Missing/incorrect class associations

QUESTION 5

5 Q5 3 / 8

- **0 pts** Correct

✓ - **1 pts** **a. not mention one-to-many**

- **1 pts** a. not mention lstm or rnn

✓ - **1 pts** **b. missing conv stride > 1**

✓ - **1 pts** **b. missing pooling methods**

- **2 pts** c. incorrect

- **1 pts** d. only words to describe the chain rule, no math derivation

- **1 pts** d. error in math formulation

✓ - **2 pts** **d. incorrect**

gng8@bu.edu

Name: Gordon Ng

BU ID: U82744816

CS 440 – Artificial Intelligence

Exam

Fall 2021

Instructions:

1- Print your name and BU ID clearly on the top right of this first page

2- You have 1 hour and 15 minutes to solve the exam

3- You may use pen or pencil, please write neatly and clearly.

Notes:

** There are five questions.*

** Total points: 48 (+2 BONUS POINTS)*

Good luck ☺

gng8@bu.edu

Q1. [10 points] Symbolic AI and Search Strategies

a) [2 points] When does *Symbolic AI* fail?

you need a human to understand what'n is happening

Symbolic AI fails when there could be meaning to it

or there is more to the problem than just looking at symbols, like a ~~sad bird would just~~ be a sad bird, but humans can tell the story behind the ~~scenes~~ with one look.

b) [2 points] When do *Neural Networks* fail?

- NN fail when there is not enough data for the NN

NNs are very data hungry.

c) [2 points] Iterative deepening is sometimes used as an alternative to breadth first search. Give one advantage of iterative deepening over BFS, and give one disadvantage of iterative deepening as compared with BFS.

Advantage: Iterative deepening gets the parameters of the model and backtracks if it is not correct.

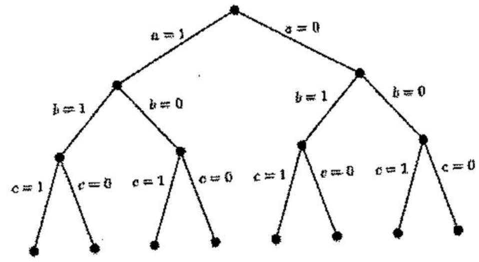
Disadvantage: Sometimes if the solution is in the front, BFS is the faster algorithm.

d) [2 points] Circle True/False and explain why.
Let $h_1(s)$ be an admissible A* heuristic. Let $h_2(s) = 2\,h_1(s)$. Then:

[✓ **True**] **False** The solution found by A* tree search with h_2 is guaranteed to be an optimal solution.

Explanation: Unless you have very high costs with the $2h_1(s)$ heuristics, when you trace the A* algorithm, you should end up with lower costs when you use the formula A* cost = costs of previous arcs + heuristic of current node.

a) [2 points] What is the disadvantage of using breadth-first search to solve the satisfiability problem on a tree like the one demonstrated below?



If the answer is at the last brance of a = 0, b = 0, c = 0, then you would have to traverse all nodes, which is very time consuming.

grg8@bu.edu

[redacted]

Q2. [8 points] Constraint Satisfaction, and Logic

a) [4 points] Akamai Technologies, Inc. is a global content delivery network (CDN), cybersecurity, and cloud service company. Akamai runs a network of thousands of servers and the servers are used to distribute content on Internet. They install a new software or update existing softwares pretty much every week. The update cannot be deployed on every server at the same time, because the server may have to be taken down for the install. Also, the update should not be done one at a time, because it will take a lot of time. There are sets of servers that cannot be taken down together, because they have certain critical functions. Map this problem to a constraint satisfaction problem whose result would inform us of the minimum number of passes needed to install the updates.

- Thousands of servers are the nodes
- Nodes where you can't update at the same time are notconnectedbyarcs.
- there should be seperate update times depending on availability
- Constraint is added where servers that can't be taken down together is dissasociated from the algoritm to not have a fail case.

b) [2 points] Use a truth table to show that $\neg(\neg(p \wedge q) \vee p)$ is a ***contradiction***.

p	q	$\neg p \wedge q$	$\neg(p \wedge q) \vee p$	$\neg(\neg(p \wedge q) \vee p)$
T	T	F	T	F
T	F	T	T	F
F	T	T	T	F
F	F	F	T	F

Because the last ~~row~~ col is all false, it is a contradiction

c) [2 points] Prove that $((p \rightarrow q) \wedge \neg q) \rightarrow \neg p$ is a ***tautology*** *using propositional equivalencies.*

$(\neg p \wedge q \wedge \neg q) \rightarrow \neg p$ implicative

$\neg(\neg p \wedge q \wedge \neg q) \wedge \neg p$ implicative

T	T	F	F
T	F	T	T
F	T	F	T
F	F	T	F

$\neg(p \vee q \wedge \neg q) \wedge \neg p$ demorgans

$p \vee T \vee \neg p$

T tautology law

angl@bu.edu

Q3. [12 points] General Concepts

Answer the following questions in brief one or two sentence answers.

a) [6 points] In gradient descent, mini-batches are typically used to compute the update step for a parameter w.

(i) What is a mini-batch?

A subset ~~of~~ the data to compute the gradient descent params.

(ii) List one advantage of using a mini-batch.

- large batches provide a more accurate gradient descent
- small batch provides regularization as noise is added at the beginning

(iii) Write a generic formulation for the update step of gradient descent for a parameter w and a cost function J.

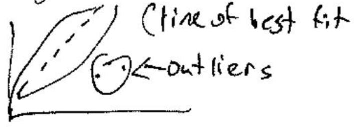
gradient descent $= \theta - \alpha \frac{\partial}{\partial w} J(\theta)$

b) [2 points] Consider the following loss function:

$$\frac{1}{2m}\sum_{i=1}^{m}\left(h_\theta(x^{(i)}) - y^{(i)}\right)^2$$

i) For what task can the above loss function be used?

It can be ~~mea~~ error squared for linearization tasks or regularization tasks:



ii) Describe all the terms included in the formulation.

h_θ = hueristics cost

$x^{(i)}$ = upper bound

$y^{(i)}$ = lower bound

m = iterations

$(h_\theta(x^{(i)}) - y^{(i)})^2$ = squared erron

c) [2 points] Name an example of a data augmentation approach and explain how it helps improve model generalization.

Early stopping stops the ~~train~~ validation fold from being too generalized with the original training data, if it has too much noise because it is so similar with the original data, it fails to be a good predictor for unseen data.

d) [2 points] List two applications that would directly benefit from an AI system that is able to perform action recognition from video.

① - People watching videos with disabilities, reading their actions/emotions and helping them use the mouse of a computer

② - Early childhood, helps ~~children learn for~~ us learn why some children behave in some way due to their actions.

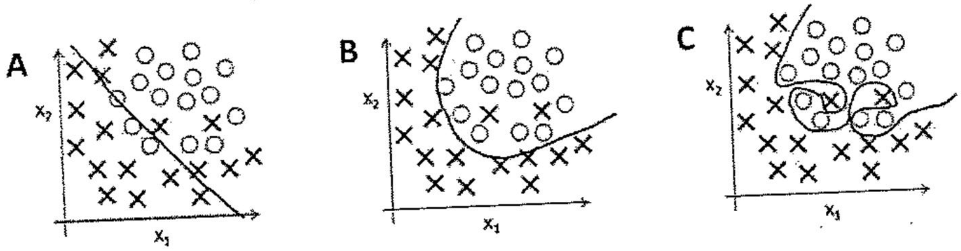
gng8@byu.edu

Q4. [10 points] Regularization and Evaluation

Suppose you want to fit a Logistic Regression model to predict whether an email is spam ($y = 1$) or not spam ($y = 0$) based on the frequency of the words "buy" (feature x_1) and "click" (feature x_2). You have fit three models by minimizing the regularized Logistic Regression cost function

$$J(\theta) = \frac{1}{m}\sum_{i=1}^{m}\left[-y^{(i)}\log\left(h_\theta(x^{(i)})\right) - (1-y^{(i)})\log\left(1-h_\theta(x^{(i)})\right)\right] + \frac{\lambda}{2m}\sum_{j=2}^{n}\theta_j^2$$

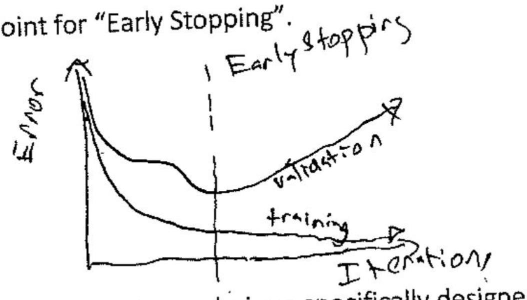
for $\lambda = 10^{-2}, 10^{0}, 10^{2}$. The following are sketches of the resulting decision boundaries.



a) [3 points] Which value of λ goes with each of the plots?

A: 10^2 **B:** 10^0 **C:** 10^{-2}

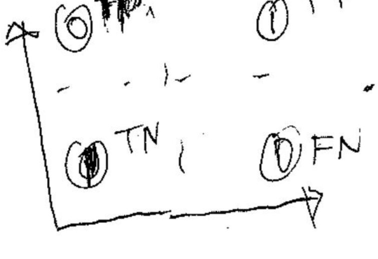
b) [3 points] Draw an example of training and validation curves that illustrate overfitting behavior, and mark the ideal point for "Early Stopping".



c) [2 points] Name the regularization technique specifically designed for Deep Neural networks and briefly describe how it achieves such regularization.

Backpropogation achieves regularization by remaing noise as you go on with the data because it updates gradient. The sigmoid function in LSTM also helps as it sets irrelevant information to 0.

d) [2 points] Sketch how the confusion matrix for this problem would look like, and label True Positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN) entries.



gng8@bu.edu

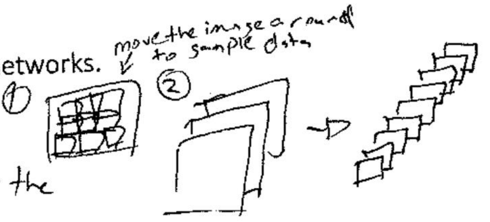
Q5. [8 points + 2 bonus] Neural Networks and Deep Learning

Answer the following questions in brief one or two sentence answers.

a) [2 points] Name a deep learning architecture that can be used for image captioning. Assume the input to your system is an image, and the desired output is a caption describing the input image.

□ → CNN → RNN (language model) → captions

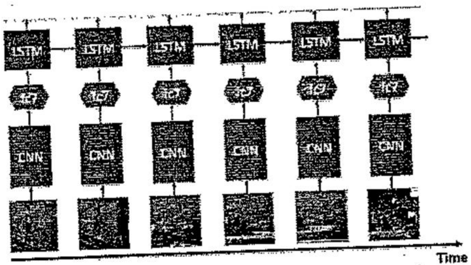
b) [2 points] List two ways to downsize feature maps in convolutional neural networks.



① - Use linearization techniques

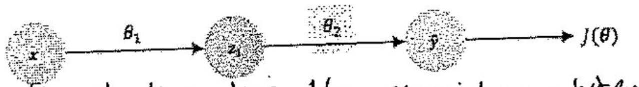
② - Use split image into smaller subsections to train it for the next step, so you get more detailed data/features

c) [2 points] Propose a strategy to determine which video frame(s) is/are ***most important*** for predicting the action happening in a video fed into a CNN-RNN architecture as depicted below:



You backtrack trace from the last LSTM on the top right and do a lot of derivatives to reach the front bottom to set their gradients. ~~You can~~ Then You remove some images ~~one~~ at a time to see which ~~image~~ loss of image provides the highest error, then you'll know which image was ~~[illegible]~~ critical information.

d) [2 points] In the following simplistic model, compute the following: How does a small change in θ_1 affect the final loss $J(\theta)$?



θ_1 is the beginning of the iterative loss, the sigmoid or whatever is used in z_1 will be altered and θ_2 would be different, could be the matter of pos/negative switch or 0/1 switch which would change $J(\theta)$ drastically.

e) [2 points BONUS] Describe why we need to use a discount factor for computing the future reward in Reinforcement Learning, and how this discount factor is applied to future rewards.

A discount factor helps remove noise so it works better for unseen data.