Name:_____ USC ID:_____

DSCI 565 MIDTERM

26 October 2023
6:00PM - 7:50PM

This exam contains 14 pages (including this page and extra blank pages at the end).

The are 7 top-level questions with a total of 100 points.

This exam is 110 minutes **including** submission to BlackBoard.

Submit the exam to BlackBoard by 7:50PM.

For this exam:

- One page of notes is allowed.
- Calculators are allowed.
- Not smartphones, laptops, or any device with internet connection during the exam.

When you are ready to submit the exam:

- Put away your pens and pencils.
- Then, take out your mobile device to submit the exam.

1. (20 points) Multiple Choice Questions
   1.1. (3 points) Which of the following statement(s) is/are NOT true for Gradient Descent (GD), Stochastic Gradient Descent (SGD), and Mini-Batch Gradient Descent?

Statement 1: In GD and SGD, smaller learning rates always result in faster convergence.

Statement 2: In SGD, updates are computed based on individual data points, which can introduce noise into the optimization process.

Statement 3: In Mini-Batch Gradient Descent, updates are calculated based on a subset of data points, reducing noise compared to SGD.

A) Only 1

B) Only 2

C) Only 3

D) 1 and 2

E) 2 and 3

F) 1 and 3

G)1, 2, and 3

1.2. (4 points) Distribution shift refers to situations where the distribution of data in the real-world deployment environment differs significantly from the distribution of the training data. Which of the following are causes of distribution shift in machine learning? (Multiple Correct)

A) The actual deployment environment may be different from the assumed population distribution, leading to a mismatch in data characteristics.

B) A machine learning model with a high VC dimension, which can handle complex data, increasing the likelihood of distribution shift.

C) The environment adapts to the model's predictions, such as in spam detection, altering the data distribution.

D) Changes in the patient population demographics over time, resulting in a shift in the data distribution.

1.3. (4 points) In the context of fairness, accountability, and transparency in machine learning, which of the following steps are recommended before deploying a machine learning model in real-world applications? (Multiple Correct)

A) Ensuring the model is trained on the largest possible dataset to maximize accuracy, without considering potential biases or ethical implications.

B) Conducting an impact analysis to assess how the model's decisions may affect various groups or individuals.

C) Verifying that the model's decisions align with ethical and fairness standards across different subpopulations.

D) Establishing a structured framework for continuous monitoring, management, and ethical oversight of the model's behavior.

1.4. (3 points) Which of the following activation functions retains only positive elements and discards negative elements, making it well-behaved with no vanishing gradients?

A. Sigmoid

B. Softmax

C. Tanh

D. ReLU

E. Leaky ReLU

1.5. (2 points) Sigmoid was the most commonly used activation function in neural network, until an issue was identified. The issue is that when the gradients are too large in positive or negative direction, the resulting gradients coming out of the activation function get squashed. This is called saturation of the neuron. That is why ReLU function was proposed, which kept the gradients same as before in the positive direction. A ReLU unit in neural network never gets saturated.

A) TRUE

B) FALSE

1.6. (4 points) In deep learning, what are exploding gradients and vanishing gradients, and how can these issues be addressed? (Select all that apply)

A) Exploding gradients occur when gradients during training become extremely large, leading to numerical instability and divergence.

B) Vanishing gradients describe the phenomenon when gradients vanish as they become too small to be useful, causing slow learning and convergence problems.

C) One solution to mitigate exploding gradients is to use gradient clipping, which limits the gradients to a specific threshold.

D) To address vanishing gradients, using activation functions like ReLU is recommended.

E) To address vanishing gradients, initialization techniques like Xavier (Glorot) for weight initialization can be employed.

2. (19 points) Convolution Neural Network

2.1. (3 points) In the context of image classification, describe the translational invariance principle.

2.2. (3 points) In the context of image classification, describe the local principle.

2.3. (4 points) How does convolution neural network realize these principles. Explain.

Consider a CNN with an input image of size 64x64x3 (height x width x number of channels). The network consists of a convolutional layer with 32 filters, each of size 3x3, a stride of 1, and zero padding of size 1.

2.4. (3 points) After passing through this convolutional layer, what will be the size of the output feature map?

2.5. (3 points) How many parameters are needed to represent this convolutional layer?

2.6. (3 points) How many floating-point operations are needed to compute a forward pass of this layer?

3. (12 points) Normalization
   3.1. (4 points) What is the purpose of batch normalization?

Given this output tensor from a minibatch of size 2:

$$\begin{bmatrix} 0 & 1 & 2 \\ 1 & 0 & 1 \end{bmatrix}$$

   3.2. (4 points) Compute batch normalization of this tensor. Assume scale parameter $\gamma = 2$ and shift parameter $\beta = 0$.

3.3. (4 points) Compute the layer normalization of this tensor.

4. (12 points) ResNet
   4.1. (4 points) Draw and describe a residual block.

   4.2. (4 points) How many parameters are needed to represent a 1x1 convolution?

   4.3. (4 points) Why a 1x1 convolution might be needed for a residual block.

5. Attention

Here is an attention function based on distance:

$$\alpha(q, k_i) = -\frac{1}{2}\|q - k_i\|^2 = q^T k_i - \frac{1}{2}\|k_i\|^2 - \frac{1}{2}\|q\|^2$$

5.1. (4 points) What would be needed to transform it into a dot product attention?

5.2. (4 points) The dot production attention is used by Transformers (Vaswani et al., 2017) is scaled by $1/\sqrt{d}$. What does d refer to? What is the reasoning behind the scaling?

5.3. (4 points) The decoder of the Transformer contains two attention blocks. What are the differences between these two attention blocks?

6. (9 points) N-Gram

Consider the following text corpus:

"Humpty Dumpty sat on a wall,
Humpty Dumpty had a great fall.
All the king's horses and all the king's men,
Couldn't put Humpty together again."

Using a 2-gram model, answer the following:

Note that $P(a|b)$ is the probability that word a follow word b

6.1. (3 points) What is $P(Dumpty \mid Humpty)$?

6.2. (3 points) What is $P(king's \mid men)$?

6.3. (3 points) What is $P(fall \mid put)$?

7. (16 points) Optimization

Consider the function $f(x_1, x_2) = 0.1x_1 + 0.5x_2^2$ for $x_1 \geq 0$ and its gradient $\nabla f(x_1, x_2) = (0.1, x_2)$.

Suppose we are using gradient descent and starting out at point (10, 1).

7.1. (3 points) At what range of learning rates η does gradient descent procedure diverge?

7.2. (4 points) Suppose we set the learning rate to be $\eta = 0.1$. Carry out three iterations of gradient descent. What is the x_1, x_2 point at the end of the three iterations.

7.3. (3 points) In the $x_1$direction, suppose we want the $x_1$to be within 0.1 of the optimum, how many iterations would be needed?

7.4. (3 points) In the $x2_1$direction, suppose we want the $x_2$to be within 0.1 of the optimum, how many iterations would be needed?

7.5. (3 points) Suppose we keep the learning rate at $\eta = 0.1$, how would you improve the gradient descent procedure to improve convergence?

BLANK PAGE

BLANK PAGE