USC CSCI 467
Intro to Machine Learning

Midterm Exam

March 7, 2024, 3:30-4:50pm

Spring 2024
Instructor: Robin Jia

Name: ____

USC e-mail: ____@usc.edu

Answer the questions in the spaces provided. **If you run out of space, continue your work on the last two pages, and indicate that your answer is there**. You may use the backs of pages for scratch work only. **Please use pen for ease of grading**. This exam has 6 questions, for a total of 100 points.

Question 1: Fetal Weight Estimation (26 points)

In medicine, various formulas have been developed to estimate the weight of a developing fetus during pregnancy. This is important to estimate because low weight or excessive weight can both cause certain medical complications.

Fetal weight (abbreviated as FW) is difficult to measure directly; however, using ultrasound imaging, various lengths and circumferences of the fetus can be measured, from which FW can be estimated. In particular, the following can be measured using ultrasound (all in centimeters):

- Head circumference (HC).
- Abdominal circumference (AC).
- Femur length (FL): Length of the thigh bone.
- Biparietal Diameter (BPD): Distance between two parietal bones of the skull.

Ryan wants to train a machine learning model to predict the fetal weight using these measurements. He obtains a training dataset $D = \{(x^{(1)}, y^{(1)}), \ldots, (x^{(n)}, y^{(n)})\}$, where each $x^{(i)} \in \mathbb{R}^4$ contains the HC, AC, FL, and BPD measurements for a fetus (in that order), and $y^{(i)}$ is the corresponding fetal weight (in grams).

(a) As a baseline approach, Ryan decides to train a linear model directly on this dataset with no other pre-processing.

i. (1 point) Should he use linear regression, logistic regression, or softmax regression?

ii. (2 points) Ryan will learn a weight w and bias b. State whether each one is a scalar, vector, or matrix. For each vector or matrix, state its dimension(s).

iii. (2 points) At test time, Ryan's model is given the HC, AC, FL, and BPD measurements for a new fetus. Write the formula that describes the prediction that the model will make in terms of these four measurements and the parameters of the model.

(b) After training this baseline linear model, Ryan then learns that in some cases, only AC, FL, and HC can be measured, but not BPD. In these cases, he would need to predict FW using only the first three features.

i. (4 points) Ryan has an idea: He can just set BPD equal to 0 and use the trained model from the previous part to make a prediction. Is this likely to work well or not? Explain your reasoning in 1-2 sentences.

ii. (2 points) Name one method we learned about in class that can be used even if some features, like BPD, are not observed. You may not use the same answer as part a(i).

(c) Ryan reasons that the weight should be proportional to volume, which is in units of cubic centimeters. Therefore, he think that a good formula for FW should involve features where 3 of the original features are multiplied together.

i. (3 points) In 1-2 sentences, explain why the model from the previous part cannot learn the type of formula that Ryan wants.

ii. (4 points) Describe **two** different ways Ryan could train a model to learn the type of formula he is looking for. Explain your answer in detail.

(d) (8 points) One actual formula used to estimate fetal weight is the following:

$$FW = 10^{1.3596 - 0.00386 \times AC \times FL + 0.0064 \times HC + 0.00061 \times BPD \times AC + 0.0424 \times AC + 0.174 \times FL}$$

How do you think the six numbers in this formula (the 1.3596, −0.00386, etc.) were chosen? Describe in detail a procedure that could be used to come up with these numbers. Assume you have access to the same training dataset as in the previous parts.

Question 2: Combining Kernels (16 points)

(a) (8 points) In class, we discussed that many different functions $k(\mathbf{x}, \mathbf{z})$ can be used as kernels. It turns out that not *every* function is suitable as a kernel. To prove that a function k is a valid kernel, a necessary and sufficient condition is to show that there exists a feature mapping ϕ such that $k(\mathbf{x}, \mathbf{z}) = \phi(\mathbf{x})^\top \phi(\mathbf{z})$ for all $\mathbf{x}$ and $\mathbf{z}$.

Let k_1 and k_2 be valid kernels; that is, $k_1(\mathbf{x}, \mathbf{z}) = \phi_1(\mathbf{x})^\top \phi_1(\mathbf{z})$ and $k_2(\mathbf{x}, \mathbf{z}) = \phi_2(\mathbf{x})^\top \phi_2(\mathbf{z})$, for some feature mappings ϕ_1 and ϕ_2, respectively. Show that the function

$$k(\mathbf{x}, \mathbf{z}) = k_1(\mathbf{x}, \mathbf{z}) + k_2(\mathbf{x}, \mathbf{z})$$

is a valid kernel by explicitly constructing a corresponding feature mapping $\phi(\mathbf{z})$.

(b) (4 points) Soumya is training a kernelized SVM on a classification task. He tries a polynomial kernel of degree 3 and gets 100% training accuracy but 74% development accuracy. He also tries an RBF kernel and gets 100% training accuracy but 75% development accuracy. Soumya wonders if he could get much better accuracy by using the combined kernel that adds the polynomial kernel and RBF kernel (as in the previous part). Is this likely to make his model much better? Explain your reasoning.

(c) (4 points) Soumya wants to keep using SVM with an RBF kernel. Suggest two things Soumya should try to improve the performance of this model.

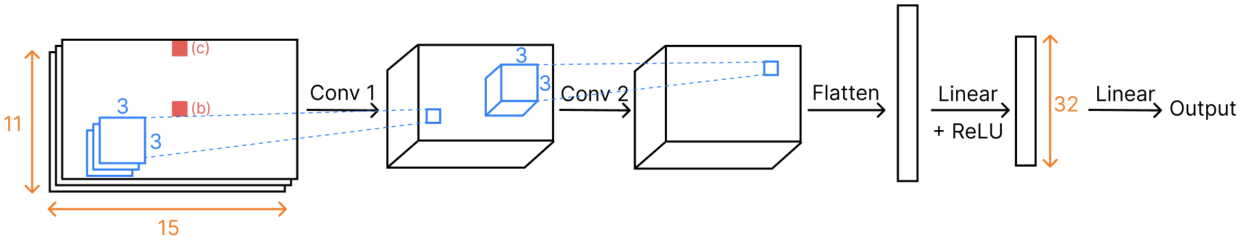
Question 3: Receptive Fields in CNNs (18 points)

Recall from class that in a Convolutional Neural Network, one of the motivations is that "each neuron should only look at a small patch of input." The portion of the input that affects the activation of a neuron is called its receptive field. Let's make this notion concrete.

Consider a CNN defined by the following operations:

- Input: $3 \times 15 \times 11$ pixel image (number of channels × width × height)
- 1st Convolutional Layer: 3×3 kernel, 4 output channels, no padding, stride is 1. ReLU is applied afterwards.
- 2nd Convolutional Layer: 3×3 kernel, 4 output channels, no padding, stride is 1. ReLU is applied afterwards.
- Flatten into a single vector, feed to an MLP with hidden state size of 32 to make a prediction.

Note that there is no max-pooling in this CNN. The diagram below visualizes the entire model, as well as the pixels referred to in parts (b) and (c).



(a) What are the sizes of the activations produced by each convolutional layer? Give your answers in the form $C \times W \times H$ where C is the number of channels, W is width, and H is height.

i. (3 points) First convolutional layer: ______________________________

ii. (3 points) Second convolutional layer: ______________________________

(b) Suppose we modify the value of the pixel in the **exact middle** of the input image.

i. (2 points) For the **first** convolutional layer, how many of the output values would be affected? Remember that there are multiple output channels.

ii. (2 points) For the **second** convolutional layer, how many of the output values would be affected? Remember that there are multiple output channels.

(c) Now suppose we modify the value of the pixel in the **middle of the top edge** of the input image.

i. (2 points) For the **first** convolutional layer, how many of the output values would be affected? Remember that there are multiple output channels.

ii. (2 points) For the **second** convolutional layer, how many of the output values would be affected? Remember that there are multiple output channels.

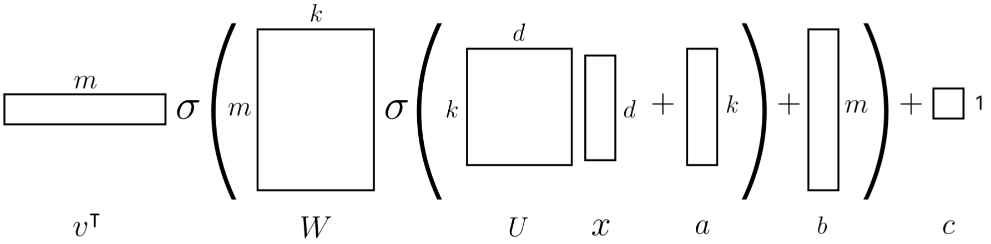
(d) (4 points) Based on your responses, does the first or second convolutional layer have a larger receptive field? Explain your answer.

Question 4: Symmetries in Neural Networks (10 points)

Consider a 3-layer fully-connected neural network defined by:

$$g(x) = v^\top \sigma(W\sigma(Ux + a) + b) + c$$

where $U \in \mathbb{R}^{k\times d}$, $a \in \mathbb{R}^k$, $W \in \mathbb{R}^{m\times k}$, $b \in \mathbb{R}^m$, $v \in \mathbb{R}^m$, $c \in \mathbb{R}$ are the parameters, σ is the sigmoid activation function, and $x \in \mathbb{R}^d$ is the input. The diagram below visualizes the dimensions of everything in this formula.



(a) (4 points) Suppose we replace U with U' where two rows i and j of U are swapped, and also replace a with a' where the i-th and j-th entries of a are swapped. All other values stay the same. How do the activations of the original first layer, $\sigma(Ux + a)$, differ from the activations of the new first layer, $\sigma(U'x + a')$?

(b) (6 points) Find values $W' \in \mathbb{R}^{m \times k}$ and $b' \in \mathbb{R}^m$ such that

$$g'(x) = v^\top \sigma(W' \sigma(U'x + a') + b') + c$$

is equal to $g(x)$ for all inputs x.

Question 5: Short Response (10 points)

Answer the following questions and **explain your reasoning fully**. You may also draw explanatory diagrams when appropriate.

(a) (5 points) Vishesh is using Naive Bayes for text classification. He gathers enough training data so that every word in the vocabulary occurs multiple times in his dataset. Because of this, Vishesh decides that he does not need to use Laplace Smoothing, since every word has a non-zero probability of occurring in his dataset. Is Vishesh right? Explain your answer.

(b) (5 points) Consider a linear model with parameter vector $w \in \mathbb{R}^d$ for binary classification. For a given training dataset, we can compute the **zero-one loss** as follows:

$$L(w) = \sum_{i=1}^{n} \mathbb{I}[y^{(i)} w^\top x^{(i)} \leq 0].$$

Recall that $\mathbb{I}[\cdot]$ is the indicator function that is 1 if the input is true, and 0 if it is false. Is it possible to use gradient descent to learn w by minimizing this loss function?

Question 6: Multiple Choice (20 points)

In the following questions, circle the correct answer(s). There is no need to explain your answer.

(a) (2 points) **True** or **False**: The objective function used in L_2-regularized logistic regression is convex.

(b) (2 points) **True** or **False**: If a loss function is convex, then there is a unique value of the model parameters that minimizes this loss function.

(c) (2 points) **True** or **False**: In SVMs, the support vectors are the $x^{(i)}$'s for which the corresponding α_i is equal to 1 or -1.

(d) (2 points) **True** or **False**: Non-parametric methods have no learnable parameters.

(e) (3 points) Which of the following classification methods make a prediction on an input x by computing $\max_y P(y \mid x)$? Choose all that apply.

A. Logistic regression
B. Naive Bayes
C. k-Nearest Neighbors
D. Support Vector Machines

(f) (3 points) Which of the following could be a reasonable choice for an activation function in a neural network? Choose all that apply.

A. $g(z) = \max(z, 0)$
B. $g(z) = \sigma(z)$
C. $g(z) = e^z$
D. $g(z) = -\max(z, 0)$

(g) (3 points) Which of the following are recommended ways to reduce overfitting in a neural network model? Choose all that apply.

A. Increase the number of neurons in the hidden layers.
B. Monitor the training loss and stop running gradient descent if the training loss goes up.
C. Randomly set half the activations to zero during training, and multiply the other activations by 2.
D. Add a term to the loss that encourages the L_2 norm of the weight matrices to be large.

(h) (3 points) Which of the following describe reasonable ways to learn word vectors? Choose all that apply.

A. Train the word vectors so that if two words occur next to each other, their word vectors are made more similar.
B. Use word vectors as part of a language model, and train the language model on a training dataset of text.
C. Train the word vectors so that they can predict which other words each word co-occurs with.
D. Use a training dataset containing word analogies (e.g., "apple is to tree as grape is to vine").

[This page provides extra space for answers]

[This page provides extra space for answers]