EXAM: SPRING 2021
CIS 472/572
INSTRUCTOR: THIEN HUU NGUYEN

May 13, 2021

The exam is closed book and open notes (1 page, handwritten except with prior permission). You will have 1 hour and 20 minutes to do this exam. Answer the questions in the spaces provided on the question sheets. If you run out of room for an answer, continue on the back of the page.
Undergraduates only: There are 8 problems in this exam. You may skip one of these problems (either Problem 1, 2, 3, 4, 5, 6, 7, or 8). Please write down on the front of your test which problem you are choosing to skip. You will receive full credit on the skipped question.

NAME _____

PROBLEM 1: PERCEPTRONS

Suppose you learn a perceptron on a linearly separable dataset by running the perceptron learning algorithm until convergence. Which of the following could be affected by changing the order of the training examples? For full points, **you must briefly explain each answer** (one sentence is enough).

1. (2 points) The convergence rate (number of iterations before convergence) could change. True or False. **Explain.**

2. (2 points) The final perceptron accuracy on the training data could change. True or False. **Explain.**

3. (2 points) The final perceptron accuracy on a separate, unseen test set could change. True or False. **Explain.**

PROBLEM 2: PERCEPTRON UPDATES

You are training a classifier to distinguish between bubonic plague (+1) and the flu (-1).

Patient	chills	fever	cramps	seizures	gangrene	Disease
	X_1	X_2	X_3	X_4	X_5	Y
1.	1	1	0	1	1	1
2.	1	1	0	0	0	-1
3.	1	1	1	1	0	1
4.	1	0	1	0	0	-1

1. (6 points) Show the weights and bias (w and b) obtained by running perceptron algorithm on this dataset for one iteration. (Here, "one iteration" means going over all of the examples once, in the order shown above. The weights and bias are initialized with zeros.)

PROBLEM 3: OVERFITTING

For each of the following model classes, briefly describe one method for reducing overfitting (1-2 sentences). In some cases, there may be more than one correct answer.

1. (1 point) Decision tree

2. (1 point) Nearest neighbor

3. (2 points) Logistic regression

4. (2 points) Support vector machine with a Gassian kernel

PROBLEM 4: LEARNING POWER

Which of the following classifiers will have zero training error (under 0/1 loss) on the following dataset? Circle YES for classifiers that will have zero error and NO for those that will not. Give a short explanation (one sentence) for each answer.

X1	X2	Category
1	1	0
1	0	1
0	1	1
0	0	0

1. (2 points) Decision trees: YES or NO? Explain.

2. (2 points) 3-nearest neighbor: YES or NO? Explain.

3. (2 points) Logistic regression: YES or NO? Explain.

PROBLEM 5: LINEAR SUPPORT VECTOR MACHINES

Consider the following 1-dimensional training data:

x	-3	0	1.5	2	3	4	5
Class	−	−	+	+	+	+	+

1. (3 points) The decision boundary of a (hard-margin) linear support vector machine (SVM) on this data is a single point on the number line that separates the positive and negative classes. What is the boundary point in this case?

2. (3 points) What are the support vectors for the linear SVM model in this case? How many individual points in this training data would be predicted incorrectly if they are removed from the training data and a linear SVM model is trained on the remaining data to make predictions (on the removed examples)? (this is called the leave one out cross validation error for this SVM on the data set).

PROBLEM 6: KERNELS

Consider a kernelized SVM with bias $b = 0$ and the following instances (support vectors) and weights:

instance (x_i)	label (y_i)	weight (α_i)
(0,0,0)	+1	1.0
(1,1,1)	-1	1.0

What is the predicted label for the instance $x = (-1, -1, -1)$ under each kernel? (Hint: Remember how we make prediction for a new example in SVM based on the support vectors. NOTE: Be careful with positive and negative signs! I recommend showing your work in order to have a chance at partial credit.)

1. (2 points) Linear kernel, $K(x, x') = x \cdot x'$.

2. (2 points) Quadratic kernel, $K(x, x') = (1 + x \cdot x')^2$

3. (2 points) Cubic kernel, $K(x, x') = (1 + x \cdot x')^3$

PROBLEM 7: TRUE/FALSE QUESTIONS

1. (2 points) Classifier A has 90% accuracy on the training set and 75% accuracy on the test set. Classifier B has 78% accuracy on both the training and test sets. Therefore, we can conclude that classifier A is better than classifier B (because it has better mean accuracy). True or False. **Explain.**

2. (2 points) Training a kernelized SVM model is equivalent to training a linear model with an expanded set of features. True or False. **Explain.**

3. (2 points) Consider two logistic regression models, trained on the same dataset with gradient descent until convergence, but initialized with different initial random weights. Given a sufficiently small learning rate and convergence threshold, both models will have the same accuracy on the test set. True or False. **Explain.**

PROBLEM 8: PERCEPTRON AND SVM

1. (3 points) In the learning algorithm of perceptron, we run over the training dataset multiple times and make an appropriate update for each training example we encounter along the way. This is very similar to the way we apply stochastic gradient descent to optimize the loss functions we studied in class. Based on this similarity, please suggest the loss function that perceptron is trying to optimize. Your loss function should be a function of the the i-th training example (i.e., with the input $x_i \in \mathbb{R}^d$ and the output $y_i \in \{-1, +1\}$), and the model parameters $w \in \mathbb{R}^d$ and the bias $b \in \mathbb{R}$: $L(x_i, y_i, w, w_0)$.

   Hint: Remember in class, we mentioned the connection between linear SVM and perceptron and that the loss function of SVM is the hinge loss.

2. (3 points) Once you write down the loss function, please justify it by showing that you can derive the update rules for perceptron based on stochastic gradient descent.