DSC 140A - Midterm 01

February 9, 2023

Name:

PID:

Exam Version: A

By signing below, you agree that you will behave honestly and fairly during and after this exam. You should not discuss any part of this exam with anyone who has not yet taken it.

Signature:

Name of student to your **left**:

Name of student to your **right**:

Exam version of student to your **left**:

Exam version of student to your **right**:

(Write "N/A" if a wall/aisle is to your left/right.)

Instructions:

- Write your solutions to the following problems in the spaces provided.
- No calculators are permitted, but a few pages of notes are.
- Write your name or PID at the top of each sheet in the space provided.

(Please do not open your exam until instructed to do so.)

Problem 1.

Consider the data set shown below:

x	y
-1	2
-3	3
2	4
4	7
5	7
6	10

What is the predicted value of y at $x = 3$ if the 3-nearest neighbor rule is used?

Problem 2.

Suppose a linear prediction rule $H(\vec{x}; \vec{w}) = \text{Aug}(\vec{x}) \cdot \vec{w}$ is parameterized by the weight vector $\vec{w} = (3, -2, 5, 2)^T$. Let $\vec{z} = (1, -1, -2)^T$. What is $H(\vec{z})$?

Problem 3.

Suppose a line of the form $H(x) = w_0 + w_1 x$ is fit to a data set of points $\{(x_i, y_i)\}$ in $\mathbb{R}^2$ by minimizing the mean squared error. Let the mean squared error of this predictor with respect to this data set be E_1.

Next, create a new data set by adding a single new point to the original data set with the property that the new point lies *exactly* on the line $H(x) = w_0 + w_1 x$ that was fit above. Let the mean squared error of H on this new data set be E_2.

Which of the following is true?

◯ $E_1 < E_2$

◯ $E_1 = E_2$

◯ $E_1 > E_2$

Problem 4.

Suppose a linear predictor H_1 is fit on a data set $X = \{\vec{x}^{(i)}, y_i\}$ of n points by minimizing the mean squared error, where each $\vec{x}^{(i)} \in \mathbb{R}^d$.

Let $Z = \{\vec{z}^{(i)}, y_i\}$ be the data set obtained from the original by standardizing each feature. That is, if a matrix were created with the ith row being $\vec{z}^{(i)}$, then the mean of each column would be zero, and the variance would be one.

Suppose a linear predictor H_2 is fit on this standardized data by minimizing the mean squared error.

True or False: $H_1(\vec{x}^{(i)}) = H_2(\vec{z}^{(i)})$ for each $i = 1, \ldots, n$.

◯ True

◯ False

Problem 5. (2 points)

a) Let Φ be an $n \times d$ design matrix, let λ be a real number, and let I be a $d \times d$ identity matrix. What type of object is $(\Phi^T \Phi + n\lambda I)^{-1}$?

- ◯ A scalar
- ◯ A vector in $\mathbb{R}^d$
- ◯ A vector in $\mathbb{R}^n$
- ◯ A $d \times d$ matrix
- ◯ A $n \times n$ matrix

b) Let Φ be an $n \times d$ design matrix, and let $\vec{y} \in \mathbb{R}^n$. What type of object is $\Phi^T \vec{y}$?

- ◯ A scalar
- ◯ A vector in $\mathbb{R}^d$
- ◯ A vector in $\mathbb{R}^n$
- ◯ A $d \times d$ matrix
- ◯ A $n \times n$ matrix

c) Let $\vec{w} \in \mathbb{R}^{d+1}$, and for for each $i \in \{1, 2, \ldots, n\}$ let $\vec{x}^{(i)} \in \mathbb{R}^d$ and $y_i \in \mathbb{R}$. What type of object is:

$$\sum_{i=1}^{n} \left(\vec{w} \cdot \text{Aug}(\vec{x}^{(i)}) - y_i \right)^2 ?$$

- ◯ A scalar
- ◯ A vector in $\mathbb{R}^d$
- ◯ A vector in $\mathbb{R}^{d+1}$
- ◯ A vector in $\mathbb{R}^n$
- ◯ A $d \times d$ matrix
- ◯ A $n \times n$ matrix

d) Let $\vec{w} \in \mathbb{R}^{d+1}$, and for for each $i \in \{1, 2, \ldots, n\}$ let $\vec{x}^{(i)} \in \mathbb{R}^d$ and $y_i \in \mathbb{R}$. Consider the empirical risk with respect to the square loss of a linear predictor on a data set of n points:

$$R(\vec{w}) = \frac{1}{n} \sum_{i=1}^{n} (\vec{w} \cdot \text{Aug}(\vec{x}^{(i)}) - y_i)^2$$

What type of object is $\nabla R(\vec{w})$; that is, the gradient of the risk with respect to the parameter vector $\vec{w}$?

- ◯ A scalar
- ◯ A vector in $\mathbb{R}^d$
- ◯ A vector in $\mathbb{R}^{d+1}$
- ◯ A vector in $\mathbb{R}^n$
- ◯ A $d \times d$ matrix
- ◯ A $n \times n$ matrix

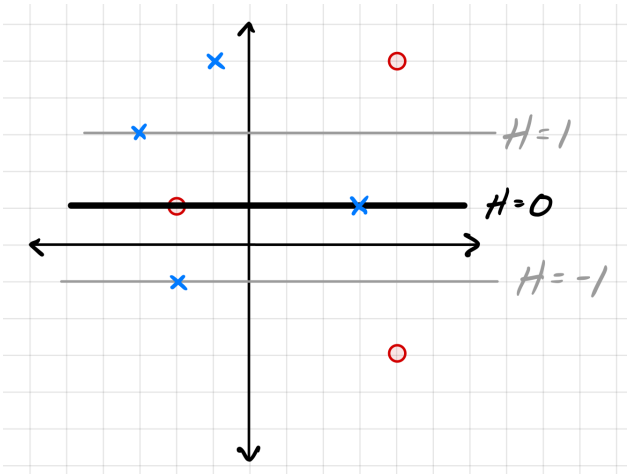
Problem 6.

Let $\vec{x}^{(1)} = (-1, -1)^T$, $\vec{x}^{(2)} = (1, 1)^T$, and $\vec{x}^{(3)} = (-1, 1)^T$. Suppose H is a linear prediction function, and that $H(\vec{x}^{(1)}) = 2$ while $H(\vec{x}^{(2)}) = -2$ and $H(\vec{x}^{(3)}) = 0$.

Let $\vec{x}^{(4)} = (1, -1)^T$. What is $H(\vec{x}^{(4)})$?

Problem 7.

Consider the data set shown below. The "×" points have label +1, while the "∘" points have label -1. Shown are the places where a linear prediction function H is equal to zero, 1, and -1.



For each of the below subproblems, calculate the **total** loss with respect to the given loss function. That is, you should calculate $\sum_{i=1}^{n} L(\vec{x}^{(i)}, y_i, \vec{w})$ using the appropriate loss function in place of L. Note that we have most often calculated the *mean* loss, but here we calculate the *total* so that we encounter fewer fractions.

a) What is the **total** square loss of H on this data set?

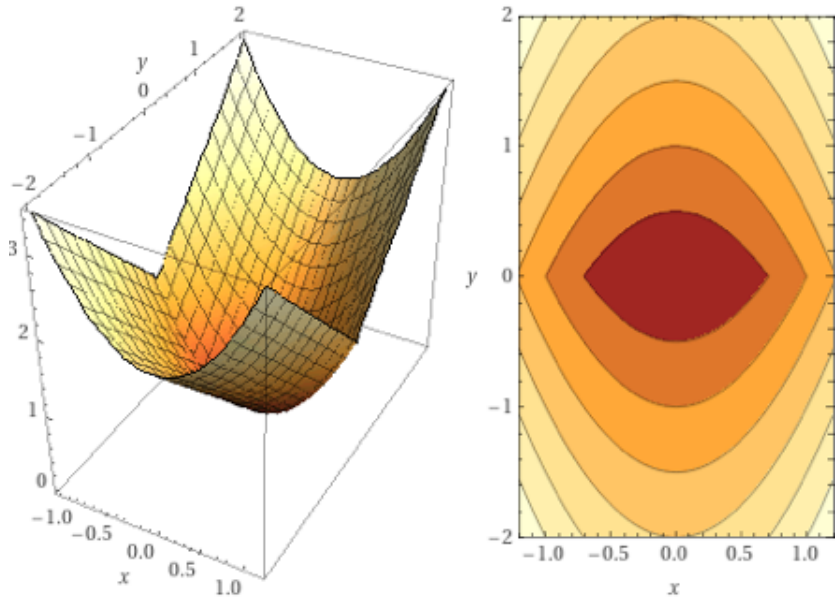
b) What is the **total** perceptron loss of H on this data set?

c) What is the **total** hinge loss of H on this data set?

PID or Name: _____

Problem 8.

Consider the function $f(x, y) = x^2 + |y|$. Plots of this function's surface and contours are shown below.



Which of the following are subgradients of f at the point $(0, 0)$? Check all that apply.

- ☐ $(0, 0)^T$
- ☐ $(0, 1)^T$
- ☐ $(0, -1)^T$
- ☐ $(1, 0)^T$
- ☐ $(-1, 0)^T$

Problem 9.

Suppose gradient descent is to be used to train a perceptron classifier $H(\vec{x}; \vec{w})$ on a data set of n points, $\{\vec{x}^{(i)}, y_i\}$. Recall that each iteration of gradient descent takes a step in the opposite direction of the "gradient".

Which gradient is being referred to here?

- ◯ The gradient of the empirical risk with respect to $\vec{w}$
- ◯ The gradient of the empirical risk with respect to $\vec{x}^{(i)}$
- ◯ The gradient of the perceptron loss with respect to $\vec{w}$
- ◯ The gradient of the perceptron loss with respect to $\vec{x}^{(i)}$
- ◯ The gradient of H with respect to $\vec{w}$
- ◯ The gradient of H with respect to $\vec{x}$

Problem 10.

Let $\{\vec{x}^{(i)}\}$ be a set of n vectors in $\mathbb{R}^d$. Consider the function $f(\vec{w}) = \sum_{i=1}^{n} \vec{w} \cdot \vec{x}^{(i)}$, where $\vec{w} \in \mathbb{R}^d$.

True or False: f is convex as a function of $\vec{w}$.

- ◯ True
- ◯ False

Problem 11.

Let $X = \{\vec{x}^{(i)}, y_i\}$ be a data set of n points where each $\vec{x}^{(i)} \in \mathbb{R}^d$.

Let $Z = \{\vec{z}^{(i)}, y_i\}$ be the data set obtained from the original by standardizing each feature. That is, if a matrix were created with the ith row being $\vec{z}^{(i)}$, then the mean of each column would be zero, and the variance would be one.

Suppose that X and Z are both linearly-separable. Suppose Hard-SVMs H_1 and H_2 are trained on X and Z, respectively.

True or False: $H_1(\vec{x}^{(i)}) = H_2(\vec{z}^{(i)})$ for each $i = 1, \ldots, n$.

◯ True

◯ False

Problem 12.

Suppose a data set $\{\vec{x}^{(i)}, y_i\}$ is linearly-separable.

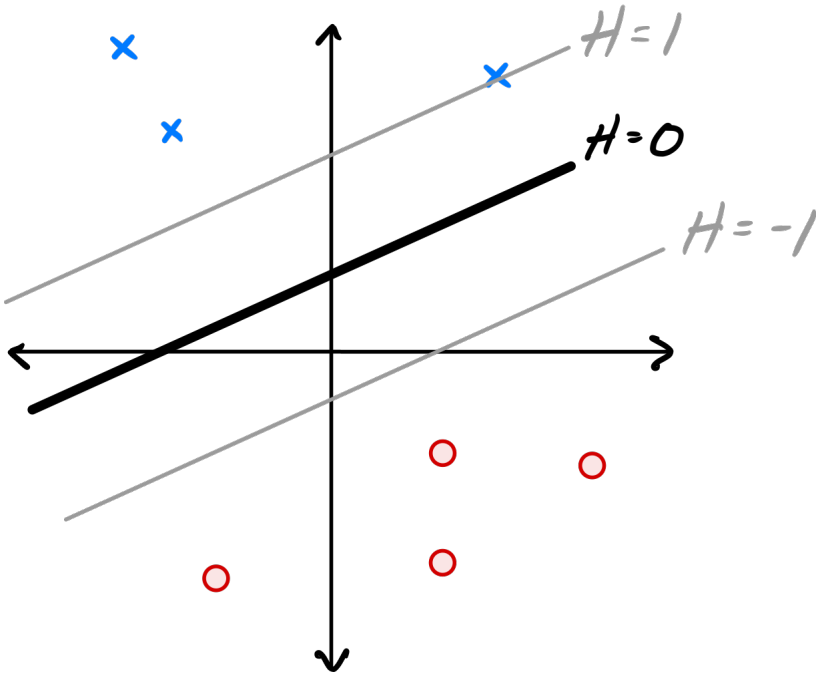
True or false: a least squares classifier trained on this data set is guaranteed to achieve a training error of zero.

◯ True

◯ False

Problem 13.

The image below shows a linear prediction function H along with a data set; the "×" points have label +1 while the "○" points have label -1. Also shown are the places where the output of H is 0, 1, and -1.



True or False: H could have been learned by training a Hard-SVM on this data set.

◯ True

◯ False

PID or Name: _____

Problem 14.

Let

$$\vec{x}^{(1)} = (1, 2, 0)^T$$
$$\vec{x}^{(2)} = (-1, -1, -1)^T$$
$$\vec{x}^{(3)} = (2, 2, 0)^T$$
$$\vec{x}^{(4)} = (0, 2, 0)^T.$$

Suppose a prediction function $H(\vec{x})$ is learned using kernel ridge regression on the above data set using the kernel $\kappa(\vec{x}, \vec{x}') = (1 + \vec{x} \cdot \vec{x}')^2$ and regularization parameter $\lambda = 3$. Suppose that $\vec{\alpha} = (1, 0, -1, 2)^T$ is the solution of the dual problem.

Let $\vec{x} = (0, 1, 0)^T$ be a new point. What is $H(\vec{x})$?

Problem 15.

Let $\{\vec{x}^{(i)}, y_i\}$ be a data set of n points, with each $\vec{x}^{(i)} \in \mathbb{R}^d$. Recall that the solution to the kernel ridge regression problem is $\vec{\alpha} = (K + n\lambda I)^{-1}\vec{y}$, where K is the kernel matrix, I is the identity matrix, $\lambda > 0$ is a regularization parameter, and $\vec{y} = (y_1, \ldots, y_n)^T$.

Suppose kernel ridge regression is performed with a kernel κ that is a kernel for a feature map $\vec{\phi} : \mathbb{R}^d \to \mathbb{R}^k$.

What is the size of the kernel matrix, K?

- ◯ $d \times d$
- ◯ $k \times k$
- ◯ $n \times n$
- ◯ $n \times d$

Problem 16.

Let $f(\vec{w}) = \vec{a} \cdot \vec{w} + \lambda \|\vec{w}\|^2$, where $\vec{w} \in \mathbb{R}^d$, $\vec{a} \in \mathbb{R}^d$, and $\lambda > 0$.

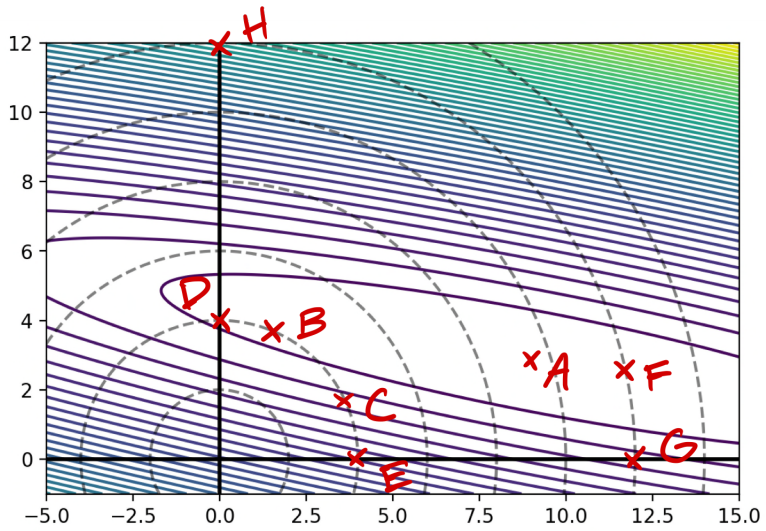
What is the minimizer of f? State your answer in terms of $\vec{a}$ and λ.

$\vec{w}^* =$

Problem 17. (2 points)

Let $R(\vec{w})$ be the unregularized empirical risk with respect to the square loss (that is, the mean squared error) on a data set.

The image below shows the contours of $R(\vec{w})$. The dashed lines show places where $\|\vec{w}\|_2$ is 2, 4, 6, etc.



a) Assuming that one of the points below is the minimizer of the *un*regularized risk, $R(\vec{w})$, which could it possibly be?

- ◯ A
- ◯ B
- ◯ C
- ◯ D
- ◯ E
- ◯ F
- ◯ G
- ◯ H

b) Let the **regularized** risk $\tilde{R}(\vec{w}) = R(\vec{w}) + \lambda\|\vec{w}\|_2^2$, where $\lambda > 0$.
Assuming that one of the points below is the minimizer of the **regularized** risk, $\tilde{R}(\vec{w})$, which could it possibly be?

- ◯ A
- ◯ B
- ◯ C
- ◯ D
- ◯ E
- ◯ F
- ◯ G
- ◯ H

c) Before turning in your exam, please check that your name is on every page.

(You may detach and use this page for scratch work. You do not need to turn it in.)

(You may detach and use this page for scratch work. You do not need to turn it in.)