

STAT 451 Midterm Exam NetID (mine is “jgillett” from “jgillett@wisc.edu”): _____

First name Last name (please write clearly so Gradescope’s OCR can read your name):

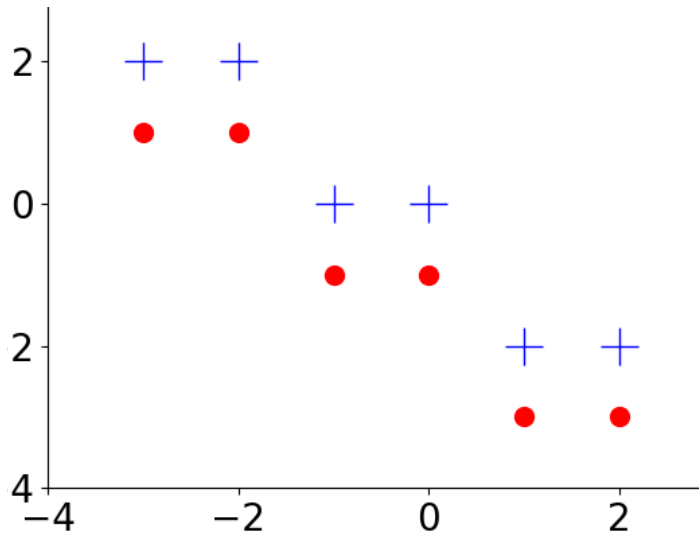
Instructions:

1. Please sit in columns with an empty seat separating columns.
2. Do not open the exam until I say “go.”
3. Put away everything except a pencil or pen, a calculator, and your two one-page (two sides each) notes sheets.
4. For questions with circles in front of the answers, fill in one circle completely. Ok?

☒ Yes
☐ No
5. Show your work. Correct answers without at least a minimal version of the work normally required may receive no credit. For a question with an answer box “,” write only the answer in the box; work should be outside the box.
6. If you continue writing or do not turn in your exam when I say time is up, you risk a penalty. (The alternative, that you get more time than your peers, is unfair.)
7. If a question is ambiguous, resolve the ambiguity in writing. We will consider grading accordingly. e.g.
 - I think “average” refers to the population mean μ (not the sample mean $\bar{X}$).
 - I think “linear regression” refers to OLS, not ridge or lasso.

Problem	Possible	Earned
1	8	
2	4	
3	12	
4	5	
5	9	
6	16	
7	12	
8	14	
9	16	
Total	100	

1. Consider this graph of 2D data points:



Which classifiers could yield accuracy = 1.0 on training data for these data? Select all that apply.

- ☐ linear hard-margin SVM
- ☐ logistic regression
- ☐ decision tree with maximum depth 1
- ☐ decision tree with maximum depth 2
- ☐ decision tree with maximum depth unlimited ☒
- ☐ linear soft-margin SVM
- ☐ SVM with RBF kernel ☒
- ☐ k -NN ☒

2. Consider applying gradient descent with step size $\alpha = 0.1$ to find the $\mathbf{x}$ that minimizes the function $f(\mathbf{x}) = f(x_1, x_2) = (x_1 - 2)^3 + (x_2 + 1)^2$ starting from $\mathbf{x}_0 = (1, 0)$. Find the value $\mathbf{x}_1$ after one iteration.

$\mathbf{x}_1 =$ ANSWER:

$\nabla f(\mathbf{x}) = (3(x_1 - 2)^2, 2(x_2 + 1))$, which is $(3, 2)$ at $\mathbf{x} = (1, 0)$.

Move to $\mathbf{x}_1 = \mathbf{x}_0 - \alpha \nabla f(\mathbf{x}_0) = (1, 0) - 0.1(3, 2) = (0.7, -0.2)$.

3. Consider 3-NN (three nearest neighbors) with the Minkowski distance using $p = 2$.

(a) Find the distance from $\mathbf{z} = (3, 4)$ to each of the other points $\mathbf{x}$:

$\mathbf{x}$	y	Distance from $\mathbf{z}$ to $\mathbf{x}$
(0, 0)	0	ANSWER: 5

(1, 4)	1	ANSWER: 2
--------	---	-----------

(3, -2)	1	ANSWER: 6
---------	---	-----------

(3, 0)	0	ANSWER: 4
--------	---	-----------

(b) How does 3-NN classify $\mathbf{z}$?

ANSWER: 0

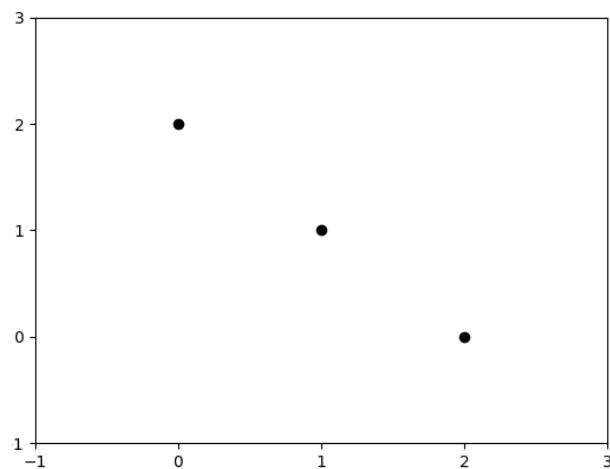
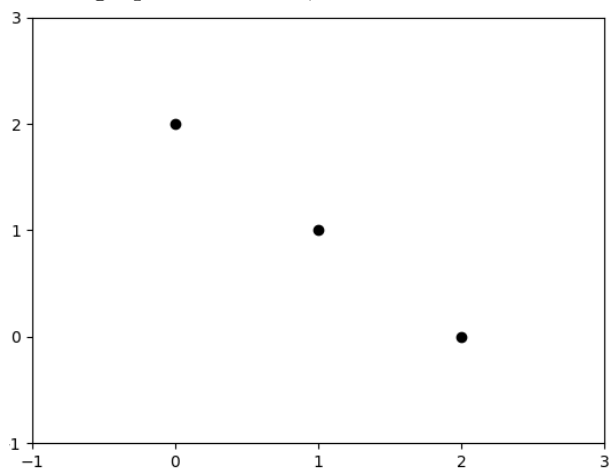
(c) How does weighted 3-NN classify $\mathbf{z}$?

ANSWER: 1

4. Consider these training examples (with 1D $\mathbf{x} = x$):

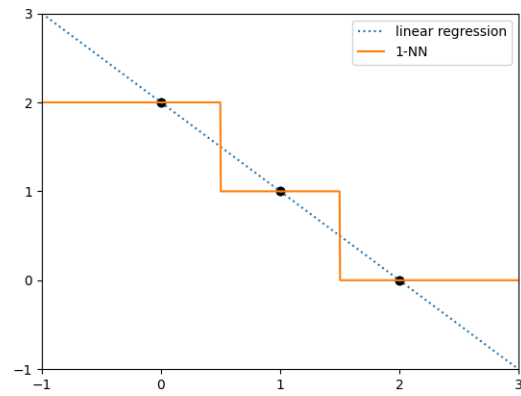
$\mathbf{x}$	y
0	2
1	1
2	0

(a) Draw the linear regression model fitted to these data over the whole x domain $(-1, 3)$ of the graph on the left, below.



(b) Draw the 1-NN regression model fitted to these data over the whole x domain $(-1, 3)$ of the graph on the right, above.

ANSWER:



5. Ordinary least squares, Lasso, and ridge regression models are trained, each with several (undisclosed) values of α . For each sequence of rows of coefficient vectors $\mathbf{w}$, indicate the corresponding model and alpha sequence. (Hint: There may be more than one correct answer; selecting at least one correct answer will earn full credit.)

(a)

α	$\mathbf{w}$						
	w_1	w_2	w_3	w_4	w_5	w_6	w_7
α_1	-2.04	0.67	0.92	0.98	0.45	-0.38	-0.22
α_2	-2.03	0.67	0.89	0.98	0.46	-0.38	-0.22
α_3	-1.94	0.73	0.72	0.91	0.54	-0.36	-0.23
α_4	-1.58	0.85	0.26	0.6	0.61	-0.22	-0.19

- ☐ ordinary least squares with α increasing
☐ ordinary least squares with α decreasing
☐ Lasso with α increasing
☐ Lasso with α decreasing
☐ ridge with α increasing ☒
☐ ridge with α decreasing

(b)

α	$\mathbf{w}$						
	w_1	w_2	w_3	w_4	w_5	w_6	w_7
α_1	-2.04	0.67	0.92	0.98	0.44	-0.38	-0.22
α_2	-2.04	0.67	0.92	0.98	0.44	-0.38	-0.22
α_3	-2.04	0.67	0.92	0.98	0.44	-0.38	-0.22
α_4	-2.04	0.67	0.92	0.98	0.44	-0.38	-0.22

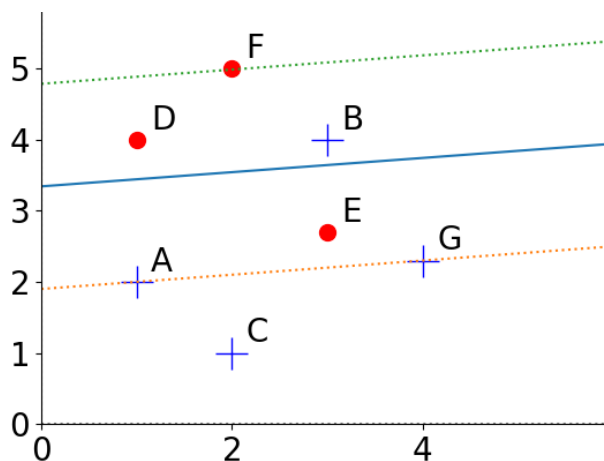
- ☐ ordinary least squares with α increasing ☒
☐ ordinary least squares with α decreasing ☒
☐ Lasso with α increasing
☐ Lasso with α decreasing
☐ ridge with α increasing
☐ ridge with α decreasing

(c)

α	$\mathbf{w}$						
	w_1	w_2	w_3	w_4	w_5	w_6	w_7
α_1	-2.14	1.33	1.43	0.07	-0.11	-0.07	-0.05
α_2	-2.09	1.37	1.34	0.01	-0.09	-0.05	-0.05
α_3	-1.72	1.22	0.82	0.0	0.08	0.0	-0.04
α_4	-0.63	0.0	-0.0	0.44	0.0	0.06	0.06

- ☐ ordinary least squares with α increasing
☐ ordinary least squares with α decreasing
☐ Lasso with α increasing ☒
☐ Lasso with α decreasing
☐ ridge with α increasing
☐ ridge with α decreasing

6. Consider this graph of a soft-margin linear SVM:



(a) Select each example for which the hinge loss is zero.

A ☒, B ☐, C ☒, D ☐, E ☐, F ☒, G ☒

(b) Select each example that the SVM classifies incorrectly.

A ☐, B ☒, C ☐, D ☐, E ☒, F ☐, G ☐

(c) Select each example for which the constraint is satisfied.

A ☒, B ☐, C ☒, D ☐, E ☐, F ☒, G ☒

(d) Select each example that has the largest hinge loss among these examples.

A ☐, B ☐, C ☐, D ☐, E ☒, F ☐, G ☐

(e) What is the accuracy of the SVM on these data?

0 ☐, $\frac{1}{7}$ ☐, $\frac{2}{7}$ ☐, $\frac{3}{7}$ ☐, $\frac{4}{7}$ ☐, $\frac{5}{7}$ ☒, $\frac{6}{7}$ ☐, 1 ☐

(f) There is a value of the hyperparameter C that improves the accuracy of the SVM on these data.

True ☐ ANSWER: ☒ There is a line that leaves only one error, improving the accuracy to $\frac{6}{7}$. , False ☐

(g) There is a value of the hyperparameter C that gives the SVM accuracy 1 on these data.

True ☐, False ☐ ANSWER: ☒ No line can separate E from A, B, C, G.

(h) Using $\mathbf{w} = (1, 1)$ and $b = -2$ would give which accuracy on these data?

☐ Either 0 or 1

☐ Either $\frac{1}{7}$ or $\frac{6}{7}$

☐ Either $\frac{2}{7}$ or $\frac{5}{7}$

- Either $\frac{3}{7}$ or $\frac{4}{7}$ ANSWER: ● The decision boundary would be $1 \cdot x_1 + 1 \cdot x_2 + (-2) = 0$, which goes through $(0, 2)$ and $(2, 0)$. All 7 points are above this line.

I regraded 6(h) in Gradescope, giving everybody its 2 points. I copied the new scores to Canvas.

For this question in the printed exam, I mistakenly trimmed the plot so much that minus signs were missing from the y -axis, which showed $(3, 2, 1, 0, 1, 2, 3)$ instead of $(3, 1, 2, 0, -1, -2, -3)$; and I misread the axes when writing 6(h) and its key answer. I'm sorry for the trouble.

I've corrected the figure here so that I can use this exam as a source of sample questions for future students.

- There is not enough information to answer

7. Here are some questions about decision trees.

- (a) Consider a classification decision tree node containing the set of examples $S = \{(\mathbf{x}, y)\}$ where $\mathbf{x} = (x_1, x_2, x_3)$:

$$S$$

x_1	x_2	x_3	y
4	8	15	1
0	7	12	0
1	11	14	0
5	10	13	0
2	9	17	1
3	6	16	1

- i. The entropy of this node in bits is .

ANSWER:

The node's y values are 1, 0, 0, 0, 1, 1, so $f_{ID3}(S) = \frac{1}{|S|} \sum_{(\mathbf{x}, y) \in S} y = \frac{1}{6}(1 + 0 + 0 + 0 + 1 + 1) = \frac{1}{2}$.

$H(S) = 1$

- ii. The (feature, threshold) pair (j, t) that yields the best split for this node is feature

$j =$ and threshold $t =$.

ANSWER:

Using feature $j = 3$ and threshold $t = 14.5$ (or any $t \in [14, 15)$) splits S into $S_- = \{(\mathbf{x}, y) \in S | x^{(j)} \leq t\} = \{0, 0, 0\}$ and its complement $S_+ = \{(\mathbf{x}, y) \in S | x^{(j)} > t\} = \{1, 1, 1\}$, each of which has entropy 0.

- (b) Consider a regression decision tree with `max_depth=1` (that is, the root node is split once into two leaves) made from the set of examples $S = \{(\mathbf{x}, y)\}$ where $\mathbf{x} = x_1$:

S

x_1	y
0	10
1	11
2	12
3	13
4	14
5	15

What value does this tree predict for $x_1 = 4.5$?

ANSWER: 14

The best split uses feature $j = 1$ and threshold $t = 2.5$, yielding a left subtree containing the first 3 examples and a right subtree containing the last 3. A prediction with $x_1 = 4.5$ uses the right subtree, whose average y value is $\hat{y} = 14$.

8. Consider finding the linear regression line by hand for the points $\{(\mathbf{x}, y)\} = \{(x, y)\} = \{(1, 2), (2, 4), (3, 3)\}$. Match each mathematical quantity on the left with its value on the right. (Hint: Very little arithmetic is required.)

(a) $\mathbf{X} =$



G: must be 3×2

(b) $\mathbf{X}^T \mathbf{X} =$



B: must be 2×2

(c) $\mathbf{y} =$



D: from data

(A) 3.5

(B) $\begin{bmatrix} 3 & 6 \\ 6 & 14 \end{bmatrix}$

(d) $\mathbf{w} =$



C: must be 2×1

(C) $\begin{bmatrix} 2 \\ 0.5 \end{bmatrix}$

(D) $\begin{bmatrix} 2 \\ 4 \\ 3 \end{bmatrix}$

(e) $\hat{\mathbf{y}} =$



E: must be 3×1 and D is taken by $\mathbf{y}$

(E) $\begin{bmatrix} 2.5 \\ 3.0 \\ 3.5 \end{bmatrix}$

(F) 2

(G) $\begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{bmatrix}$

(f) $f_{\mathbf{w},b}(3) =$



A: must be scalar and 2 is taken by b

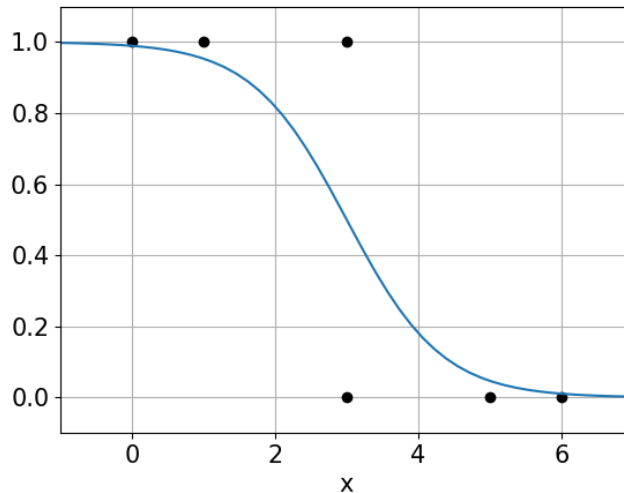
(g) intercept =



F: first element of $\mathbf{w}$

(My answers above are mostly from considering dimensions. With a little more work, the matrix arithmetic can be done to find the answers.)

9. Consider the 6-point data set below.



- (a) Estimate $P(y = 1)$ to the nearest 0.1 at $x = 4$.
 0.0 ☐, 0.1 ☐, 0.2 ☐ **ANSWER: ☒**, 0.3 ☐, 0.4 ☐, 0.5 ☐, 0.6 ☐, 0.7 ☐, 0.8 ☐, 0.9 ☐
- (b) Find the predicted $\hat{y}$ at $x = 4$. -1 ☐, 0 ☐ **ANSWER: ☒**, 1 ☐, None of the above ☐
- (c) For which of the coefficients is the likelihood of the data according logistic regression model the highest? Hint: It is not necessary to estimate any likelihoods.
- ☐ $w \approx 1.5, b \approx 4.5$
 - ☐ $w \approx 1.5, b \approx -4.5$
 - ☐ $w \approx -1.5, b \approx 4.5$ **ANSWER: ☒**
 - ☐ $w \approx -1.5, b \approx -4.5$
 - ☐ None of the above
- (d) Suppose this code was run on X and y , each having N rows and $D = 1$ columns:

```
model = linear_model.LogisticRegression(C=1000)
model.fit(X, y)
```

Which lines set b and w to the parameters specified by the logistic regression model?

- ☐ $b = \text{model.b_}[0]$
 $w = \text{model.w_}[0][0]$
- ☐ $b = \text{model.b_}[1]$
 $w = \text{model.w_}[0][1]$
- ☐ $b = \text{model.intercept_}[0]$
 $w = \text{model.coef_}[0][0]$
- ANSWER: ☒**
- ☐ $b = \text{model.intercept_}[1]$
 $w = \text{model.coef_}[0][1]$