

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:

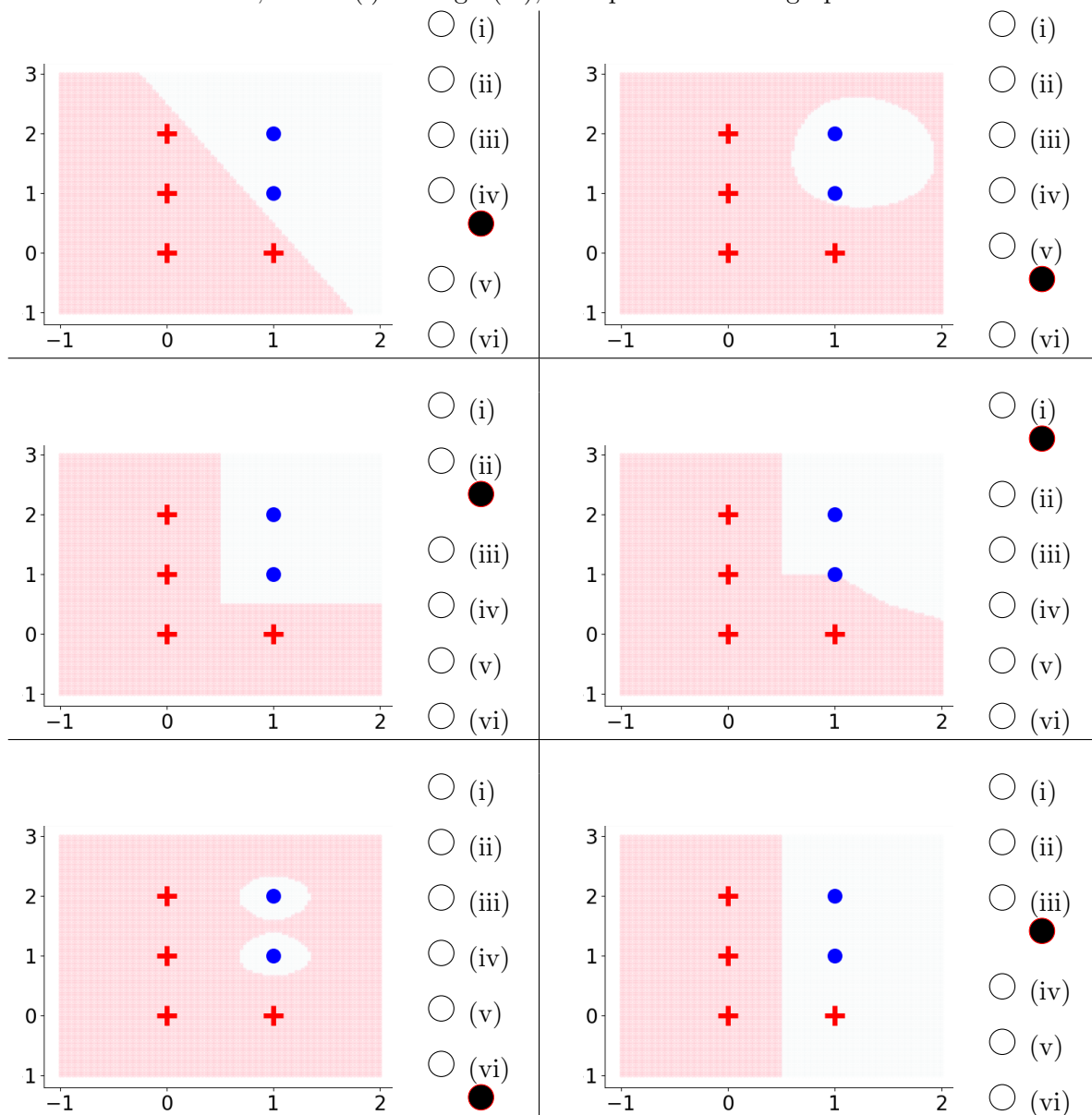
1. Please sit in columns with an empty seat separating columns.
2. Do not open the exam until I say “go.”
3. Put away everything except a pencil or pen, a calculator, and your two one-page (two sides each) notes sheets.
4. For questions with circles in front of the answers, fill in one circle completely. 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 #5 has one part worth 4 points.	Sub-total:	4
Problem parts #3a–f and #8a–b are worth 3 points each.	Sub-total:	24
The other 36 problem parts are worth 2 points each.	Sub-total:	72
	Total:	100

1. Each graph shows six 2D examples and the decision boundary of a binary classifier.

(a) Indicate the classifier, one of (i) through (vi), that produced each graph.



The hard-margin linear SVM has a linear boundary, so top-left is (iv).

The decision tree makes decisions of the form $x^{(j)} \leq t$ for some feature j and some threshold t , so its boundary (in 2D) must be defined by vertical and horizontal line segments. Bottom-right has only one split and so must be (iii), leaving center-left with its two splits using (ii).

Center-right is (i). ((i) does not match top-right or bottom-left because, e.g., (2, 1) should be blue circle for (i) because its 3NN are (1, 0), (1, 1), and (1, 2)).

Top-right is (v) and bottom-left is (vi) because increasing γ increases over-fitting.

i. `KNeighborsClassifier(n_neighbors=3, metric='euclidean')`

- ii. `DecisionTreeClassifier(criterion='entropy', max_depth=None, random_state=0)`
- iii. `DecisionTreeClassifier(criterion='entropy', max_depth=1, random_state=0)`
- iv. `svm.SVC(kernel='linear', C=1000)`
- v. `svm.SVC(kernel='rbf', C=1, gamma=1)`
- vi. `svm.SVC(kernel='rbf', C=1, gamma=6)`

(b) Which (feature $j \in \{1, 2, \dots\}$, threshold $t \in \mathbb{R}$) pair was chosen in the first split of (ii)?

feature $j =$ ANSWER: 1, threshold $t =$ ANSWER: 0.5

This must be the best split because it is used by the `max_depth=1` tree on the same data. Arithmetic is not necessary to answer the question, but here are details:

j, t	$H(S_-, S_+) = \frac{ S_- }{ S } H(S_-) + \frac{ S_+ }{ S } H(S_+)$
1, 0.5	$\frac{3}{6}(0) + \frac{3}{6} \left[-(1 - \frac{1}{3}) \log_2(1 - \frac{1}{3}) + -\frac{1}{3} \log_2(\frac{1}{3}) \right] \approx 0.46$
2, 0.5	$\frac{2}{6}(0) + \frac{4}{6}(1) = \frac{2}{3} \approx 0.67$
2, 1.5	$\frac{4}{6} \left(\left[-(1 - \frac{3}{4}) \log_2(1 - \frac{3}{4}) + -\frac{1}{4} \log_2(\frac{1}{4}) \right] \right) + \frac{2}{6}(1) \approx 0.87$

2. Consider the logistic regression model, $P(y_i = 1) = \frac{1}{1 + e^{-(\mathbf{w}\mathbf{x} + b)}}$.

(a) For each function below that plays a role in the model, indicate its image from among these choices:

- A. $\mathbb{Z}$ = integers
- B. $\mathbb{Z}_+$ = positive integers
- C. $\mathbb{R}$ = real numbers
- D. $\mathbb{R}_+$ = positive real numbers
- E. $(0, 1)$ = interval from 0 to 1

Hint: The *image* of a function is the set of all output values it may produce.

i. $f_1(\mathbf{x}) = \mathbf{w}\mathbf{x} + b$ for $\mathbf{x} \in \mathbb{R}^D$:

A ☐, B ☐, C ☐ ANSWER: ☒, D ☐, E ☐

ii. $f_2(t) = \frac{1}{1+e^{-t}}$ for $t \in \mathbb{R}$:

A ☐, B ☐, C ☐, D ☐, E ☐ ANSWER: ☒

iii. $f_3(t) = e^{-t}$ for $t \in \mathbb{R}$:

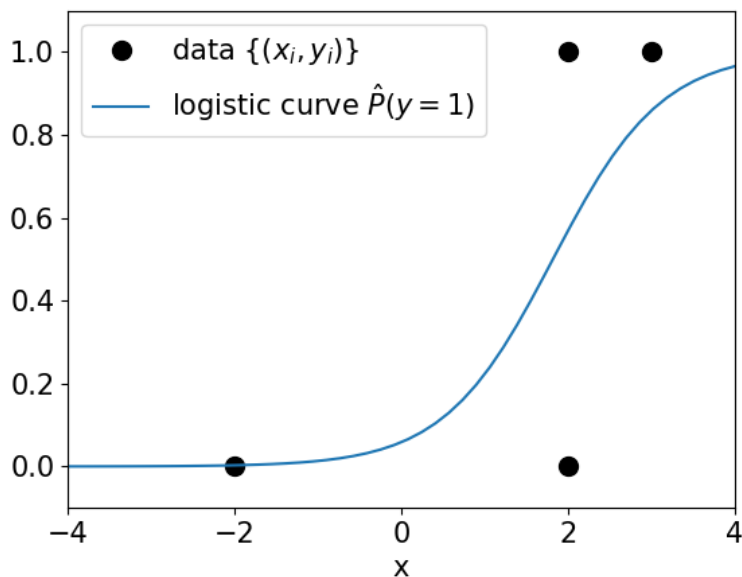
A ☐, B ☐, C ☐, D ☐ ANSWER: ☒, E ☐

(b) For the model with $\mathbf{w} = (-3, 3)$ and $b = 3$, find $\hat{P}(y = 1 | \mathbf{x} = (2, 1))$.

ANSWER:

$$\hat{P}(y = 1|\mathbf{x}) = \frac{1}{1 + e^{-(\mathbf{w}\mathbf{x}+b)}} = \frac{1}{1 + e^{-[(-3,3) \cdot (2,1)+3]}} = \frac{1}{1 + e^0} = \frac{1}{2}.$$

(c) I ran some Python/scikit-learn code to make the model pictured here:



i. For each array below, indicate the line of code that could have produced it from among these choices:

A. `model.fit(X, y)`

B. `model.intercept_`

C. `model.coef_[0]`

D. `model.predict(X)`

E. `model.predict_proba(X)[:, 1]`

F. `model.score(X, y)`

1. `array([0, 1, 1, 1])`:

A ☐, B ☐, C ☐, D ☒, E ☐, F ☐

2. `array([0.003, 0.569, 0.569, 0.859])`

A ☐, B ☐, C ☐, D ☐, E ☒, F ☐

3. `array([1.528])`

A ☐, B ☐, C ☒, D ☐, E ☐, F ☐

4. `array([-2.778])`

☐, B ☒, C ☐, D ☐, E ☐, F ☐

ii. How do we classify a new point at $x = 0.5$ if using a decision threshold of 0.5?

☐ $\hat{y} = 0$ ☒

☐ $\hat{y} \approx 0.05$

☐ $\hat{y} \approx 0.95$

☐ $\hat{y} = 1$

ANSWER:

$\hat{y} = 0$. The graph shows $\hat{P}_{\mathbf{w},b}(y = 1|x = 0.5)$ is between 0 and 0.2 (Python says ≈ 0.12), less than the 0.5 threshold. So we assign $\hat{y} = 0$.

3. Mark each statement True or False.

(a) In linear regression, a reasonable alternative to the cost function *mean squared error* =

$$\frac{1}{N} \sum_{i=1}^N [f_{\mathbf{w},b}(\mathbf{x}_i) - y_i]^2 \text{ is } \textit{mean error} = \frac{1}{N} \sum_{i=1}^N [f_{\mathbf{w},b}(\mathbf{x}_i) - y_i].$$

☐ True

☐ False ☒

(b) A linear SVM with decision boundary $(2, 1, 2) \cdot \mathbf{x} + 2 = 0$ has a smaller margin between $+1$ and -1 support vectors than one with boundary $(6, 0, 8) \cdot \mathbf{x} - 3 = 0$.

☐ True

☐ False ANSWER: ☒

The margin for the first SVM is $\frac{2}{\|\mathbf{w}\|} = \frac{2}{\sqrt{2^2+1^2+2^2}} = \frac{2}{3}$, while the margin for the second is $\frac{2}{\|\mathbf{w}\|} = \frac{2}{\sqrt{6^2+0^2+8^2}} = \frac{2}{10}$.

(c) The values for $\mathbf{w}$ and b that minimize negative log-likelihood in the logistic regression

$$\text{model also minimize the mean squared error } \frac{1}{N} \sum_{i=1}^N [P_{\mathbf{w},b}(y_i = 1)(\mathbf{x}_i) - y_i]^2.$$

☐ True

☐ False ANSWER: ☒

A predicted probability $\hat{P}_{\mathbf{w},b}(y_i = 1)(\mathbf{x}_i) \in \{0, 1\}$, while a label $y_i \in [0, 1]$. The logistic curve does not fit the data points.

(d) The entropy of a decision tree node containing the set of examples

x_1	x_2	y
2	6	0
5	7	0
3	8	1

is ≈ 0.92 .

☐ True ANSWER: ☒

The node's y values are 0, 0, 1, so $f_{ID3}(S) = P(y) = \frac{1}{|S|} \sum_{(\mathbf{x},y) \in S} y = \frac{1}{3}(0+0+1) = \frac{1}{3}$.

$$\begin{aligned}
 H(S) &= \sum_{y \in \{0,1\}} P(y) [-\log_2 P(y)] \\
 &= -\left(1 - \frac{1}{3}\right) \log_2 \left(1 - \frac{1}{3}\right) - \frac{1}{3} \log_2 \frac{1}{3} \\
 &\approx 0.92
 \end{aligned}$$

☐ False

- (e) While gradient descent's computation speed depends on the number of features D , stochastic gradient descent's computation speed does not depend on D .

☐ True

☐ False ANSWER: ●

Even if SGD uses only one example for each iteration, it can still take longer to process a high- D example than a low- D example.

- (f) For a logistic regression model on data with one feature, the midpoint of the logistic curve is always between the feature minimum and the feature maximum.

☐ True

☐ False ANSWER: ●

Consider, e.g., a data set whose "sample proportions" are all small.

4. Consider using $\mathbf{w} = (X^T X)^{-1} X^T \mathbf{y}$ to find the line fitting the points $((0,0), 3)$, $((1,2), 8)$, and $((3,4), 14)$.

Fill in these matrices to get started on using $\mathbf{w} = (X^T X)^{-1} X^T \mathbf{y}$ to find the line.

$$X = \begin{bmatrix} & & & \\ & & & \\ & & & \end{bmatrix}, \mathbf{y} = \begin{bmatrix} & & & \\ & & & \\ & & & \end{bmatrix}, X^T = \begin{bmatrix} & & & \\ & & & \\ & & & \end{bmatrix}$$

(You should not continue the computation to find the line.)

$$\text{ANSWER: } X = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 2 \\ 1 & 3 & 4 \end{bmatrix}, \mathbf{y} = \begin{bmatrix} 3 \\ 8 \\ 14 \end{bmatrix}, X^T = \begin{bmatrix} 1 & 1 & 1 \\ 0 & 1 & 3 \\ 0 & 2 & 4 \end{bmatrix}$$

(The line is $y = \mathbf{w}\mathbf{x} + b = (1, 2) \cdot \mathbf{x} + 3$.)

5. Consider applying gradient descent with step size $\alpha = 0.5$ to find the $\mathbf{x}$ that minimizes the function $f(\mathbf{x}) = f(x^{(1)}, x^{(2)}) = (x^{(1)} - 1)^2 + (x^{(2)} + 2)^2$ starting from $\mathbf{x}_0 = (1, 2)$. Find the value $\mathbf{x}_1$ after one iteration.

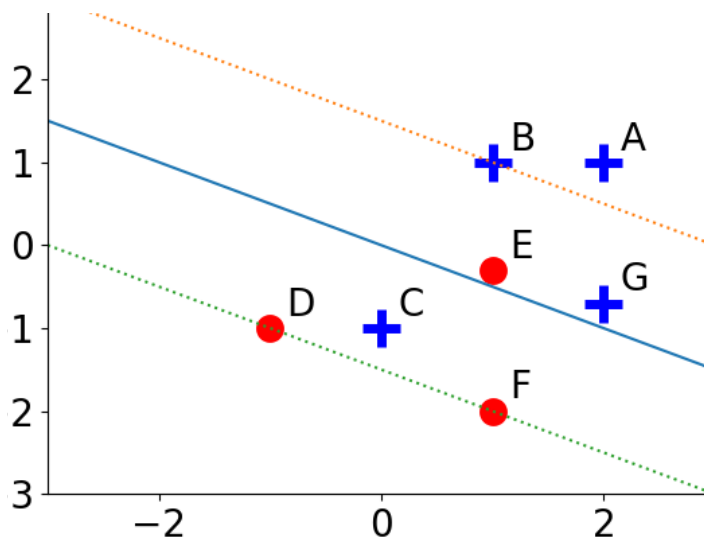
$\mathbf{x}_1 =$

ANSWER:

$\nabla f(\mathbf{x}) = (2(x^{(1)} - 1), 2(x^{(2)} + 2))$, which is $(0, 8)$ at $\mathbf{x} = (1, 2)$.

Move to $\mathbf{x}_1 = \mathbf{x}_0 - \alpha \nabla f(\mathbf{x}_0) = (1, 2) - 0.5(0, 8) = (1, -2)$.

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 C from D, E, F.

Hint: For a “Select each example ...” question, you may select none, one, or more than one of the answers.

7. In each situation, indicate the frequency (never, sometimes, or always) with which we will obtain 100% accuracy on training data that contain no pairs of examples with identical feature vectors. Suppose any hyperparameters are set to optimum values for training performance.

- | | | | |
|---|---|--|--|
| (a) Decision tree with $N = 1$ example | always ☒ | sometimes ☐ | never ☐ |
| (b) Decision tree with $N = 5$ examples | always ☐ | sometimes ☒ | never ☐ |
| (c) Hard-margin linear SVM on linearly-separable data | always ☒ | sometimes ☐ | never ☐ |
| (d) Hard-margin linear SVM on non-linearly-separable data | always ☐ | sometimes ☐ | never ☒ |
| (e) Soft-margin linear SVM on non-linearly-separable data | always ☐ | sometimes ☐ | never ☒ |
| (f) SVM with RBF kernel on non-linearly-separable data | always ☒ | sometimes ☐ | never ☐ |
| (g) k -NN with $k = 1$ | always ☒ | sometimes ☐ | never ☐ |
| (h) k -NN with $k = 3$ | always ☐ | sometimes ☒ | never ☐ |
| (i) Logistic regression | always ☐ | sometimes ☒ | never ☐ |

8. Consider a database consisting of these three examples:

name $\mathbf{x}$	age y
Leili	20
Layla	22
Kayla	24

We want to estimate Lalla’s age from her name, supposing her name is a corrupted version of one of the names in the database. (It was corrupted, e.g., by a typographical error.)

- (a) Find the Hamming distance between Lalla and each of the other three names.

name	Hamming distance to Lalla
Leili	
Layla	
Kayla	

ANSWER:

name	Hamming distance to Lalla
Leili	3
Layla	1
Kayla	2

- (b) Use 2-NN (two nearest neighbors) regression to estimate Lalla's age from her name.

Lalla's age is about years old.

Her 2NN are Layla and Kayla, so we estimate her age as their average age, $\frac{22+24}{2} = 23$.

9. For 0 points, and only if you wish, write something here to make your graders smile.