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):

Indicate your lecture with by filling one circle completely:

○ TuTh 8:00-9:15

○ TuTh 11:00-12:15

Instructions:

1. Please sit in columns with two empty seats 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 “$\boxed{}$,” 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.

Question	Points	Earned
Q0 (cover)	2	
Q1	12	
Q2	8	
Q3	12	
Q4	12	
Q5	14	
Q6	12	
Q7	16	
Q8	12	
Total	100	

1. 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
2	11	12	1
3	6	14	1
0	8	17	0
4	10	15	1
1	7	13	0
5	9	16	1

i. The entropy of this node in bits is $\boxed{}$.

ii. The (feature, threshold) pair (j, t) that yields the best split for this node is feature $j = \boxed{}$ and threshold $t = \boxed{}$.

(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	21
3	22
4	23
5	24

What value does this tree predict for $x_1 = 4.5$? $\hat{y} = \boxed{}$

2. Here are questions about feature engineering.

(a) Consider the data -5, 5, 5, 5, 5, which have these summary statistics:

- minimum -5
- mean 3
- median 5
- maximum 5
- (population) standard deviation 4

Do standardization rescaling on feature `x`:

(input) `x`	(output) `x_rescaled`
-5	
5	
5	
5	
5	

(b) Use one-hot encoding to transform the categorical feature `power_source` into binary features with reasonable names that are in alphabetical order.

(input) `power_source`	(output)
grid	
solar	
generator	
grid	

3. Consider the gradient descent algorithm.

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

   $\mathbf{x}_1 = \boxed{}$.

   (b) Mark each statement as true or false.

   - Gradient descent can fail to converge on a convex function if step size α is such that it gets stuck in a cycle, oscillating between two or several values.
     ○ True
     ○ False
   - For a non-convex function, gradient descent can fail to converge by descending without bound.
     ○ True
     ○ False
   - Gradient descent can fail to converge if it gets stuck in a local minimum.
     ○ True
     ○ False
   - Gradient descent can fail to converge on a convex function if the step size $\alpha > 0$ is too large.
     ○ True
     ○ False

4. Consider 3-NN (three nearest neighbors) using the Minkowski distance with $p = 1$.

(a) Find the distance from $\mathbf{z} = (2, 2)$ to each of the other points $\mathbf{x}$:

$\mathbf{x}$	y	Distance from $\mathbf{z}$ to $\mathbf{x}$
$(-1, -1)$	1	
(0, 1)	0	
(1, 0)	0	
$(2, 3)$	1	

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

$\hat{y} =$

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

$\hat{y} =$

(d) What y value does 3-NN regression predict for $\mathbf{z}$?

$\hat{y} =$

5. 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) $\mathbf{X} =$

Ⓐ Ⓑ Ⓒ Ⓓ Ⓔ Ⓕ Ⓖ

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

Ⓐ Ⓑ Ⓒ Ⓓ Ⓔ Ⓕ Ⓖ

(c) $\mathbf{y} =$

Ⓐ Ⓑ Ⓒ Ⓓ Ⓔ Ⓕ Ⓖ

(d) $\mathbf{w} =$

Ⓐ Ⓑ Ⓒ Ⓓ Ⓔ Ⓕ Ⓖ

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

Ⓐ Ⓑ Ⓒ Ⓓ Ⓔ Ⓕ Ⓖ

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

Ⓐ Ⓑ Ⓒ Ⓓ Ⓔ Ⓕ Ⓖ

(g) intercept $=$

Ⓐ Ⓑ Ⓒ Ⓓ Ⓔ Ⓕ Ⓖ

(A) 3.5

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

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

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

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

(F) 2

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

6. Here are some questions on support vector machines.

(a) Suppose we have a soft-margin SVM for which $\mathbf{w} = (2, 3)$ and $b = -1$. How does the SVM classify $(\mathbf{x} = (1, 1), y = 1)$?

$\boxed{}$.

(b) Suppose we have some training data $\{(\mathbf{x}_i, y_i)\}_{i=1}^N$ (where $\mathbf{x}_i$ is 2D) in matrices `X` and `y`. We have plotted the data with $y = -1$ examples red and $y = 1$ examples blue. Which line of code gives the best model for predicting new examples?

For each question, write the best answer from among these lines labeled "A" thorugh "H".

A: `clf = svm.SVC(kernel='linear', C=1); clf.fit(X, y)`

B: `clf = svm.SVC(kernel='linear', C=1000); clf.fit(X, y)`

C: `clf = svm.SVC(kernel="rbf", C=1, gamma=1); clf.fit(X, y)`

D: `clf = svm.SVC(kernel="rbf", C=1, gamma=10); clf.fit(X, y)`

E: `clf = svm.SVC(kernel="euclidean", C=1, gamma=2); clf.fit(X, y)`

F: `clf = svm.SVC(kernel="euclidean", C=1000, gamma=2); clf.fit(X, y)`

G: `clf = svm.SVC(kernel="gini", C=1); clf.fit(X, y)`

H: `clf = svm.SVC(kernel="gini", C=1000); clf.fit(X, y)`

i. A B C D E F G H
The red points are scattered between $x_1 = 0$ and $x_1 = 2\pi$ and roughly along $x_2 = \sin x_1$, a wave. The blue points are scattered over the same x_1 interval and roughly along $x_2 = \sin x_1 + 1$, a wave 1 higher than the first wave.

ii. A B C D E F G H
The data consist of two clouds of points, one red and one blue, that are linearly-separable except for a few outliers of each color.

iii. A B C D E F G H
The data are mixed red and blue points scattered randomly in the disk $x_1^2 + x_2^2 \leq 1$.

7. Mark each statement True or False.

(a) In linear regression, a reasonable alternative to the cost function *mean squared error* $= \frac{1}{N}\sum_{i=1}^{N}[f_{\mathbf{w},b}(\mathbf{x}_i) - y_i]^2$ is *sum of squared error* $= \sum_{i=1}^{N}[f_{\mathbf{w},b}(\mathbf{x}_i) - y_i]^2$.

○ True
○ False

(b) For the soft-margin SVM with decision boundary $\mathbf{w}\mathbf{x} + b = 0$ where $\mathbf{w} = (1, 2)$ and $b = 3$, the example $(\mathbf{x}, y) = ((4, 5), -1)$ has hinge loss 18.

○ True
○ False

(c) For training data $\{(\mathbf{x}, y)\}$ such that $\mathbf{x}_i \neq \mathbf{x}_j$ for all i and j, we can build a 3NN model that classifies the training examples without error.

○ True
○ False

(d) If we train a hard-margin linear SVM on linearly separable data, then discard training examples which are support vectors, and then train a new SVM on the remaining examples, the first SVM will have a wider "road" than the second.

○ True
○ False

(e) A linear SVM with decision boundary $(1, 2, 2) \cdot \mathbf{x} - 2 = 0$ has a smaller margin between $+1$ and -1 support vectors than one with boundary $(1, 4, 8) \cdot \mathbf{x} + 3 = 0$.

○ True
○ False

(f) Every decision tree regression function is a step function.

Hint: A *step function* is a function that is constant over each of one or more intervals.

○ True
○ False

(g) Every k-NN regression function is a step function.

Hint: A *step function* is a function that is constant over each of one or more intervals.

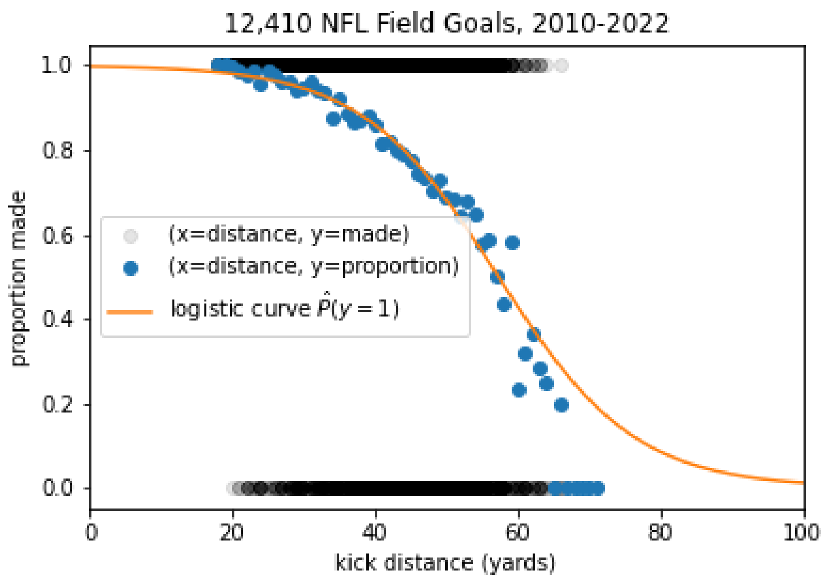
○ True
○ False

(h) In logistic regression, we use the natural log function to facilitate finding a closed-form expression for the coefficients $\mathbf{w}$ and b in terms of the data.

○ True
○ False

8. Consider a logistic regression model with $\mathbf{w} = (1, 2)$ and $b = 0$.

(a) From the logistic regression model represented in the figure, estimate the likelihood of an NFL field goal kicker making two field goals in a row, one from 20 yards and one from 60. We may suppose these attempts are independent and make other reasonable simplifiying assumptions.



P(NFL kicker makes the two field goals) $\approx$ $\boxed{}$. (Give an estimate.)

(b) In calculating the coefficients for a logistic regression model, why do we minimize negative log likelihood instead of maximizing likelihood? Mark each statement as a true or false.

i. A product of probabilities can overflow in fixed-precision computer arithmetic.
   - ◯ True
   - ◯ False

ii. A product of probabilities can underflow in fixed-precision computer arithmetic.
   - ◯ True
   - ◯ False

iii. The natural log of a product is naturally expressed as a sum, and differentiating a sum is easier than differentiating a product.
   - ◯ True
   - ◯ False

iv. The natural log is strictly increasing, so maximizing the likelihood is the same as minimizing the negative log likelihood.
   - ◯ True
   - ◯ False