(1 point for easily legible writing on this cover sheet.)

NetID: ____________________________ (mine is `jgillett` from `jgillett@wisc.edu`)

Last name: ___________________________________ First name: ____________________________

First name and last name: ____ (I am asking for these separately above and together here to use with optical character recognition software.)

Mark your lecture with "X":

_______ TuTh 1:00-2:15

STAT 451 Midterm Exam 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 one-page (two sides) notes sheet.
4. Show your work. Correct answers without at least a minimal version of the work normally required may receive no credit.
5. The exam ends when I call time. If you continue writing after I call time, you risk a penalty. (The alternative, that you get more time than your peers, is unfair.)
6. 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)	1	
Q1	10	
Q2	5	
Q3	15	
Q4	15	
Q5	10	
Q6	12	
Q7	10	
Q8	22	
Total	100	

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

   (a) Use the model to estimate probability that y is 1, given that $\mathbf{x}$ is $(-3, 1)$).

   (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 reason by circling the appropriate choice.

      i. TRUE / FALSE A product of probabilities can overflow in fixed-precision computer arithmetic.

      ii. TRUE / FALSE A product of probabilities can underflow in fixed-precision computer arithmetic.

      iii. TRUE / FALSE The natural log of a product is naturally expressed as a sum, and differentiating a sum is easier than differentiating a product.

      iv. TRUE / FALSE The natural log is strictly increasing, so maximizing the likelihood is the same as minimizing the negative log likelihood.

      v. TRUE / FALSE Using the natural log facilitates finding a closed-form expression for the coefficients in terms of the data.

2. 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 $\mathbf{x}_0 = (1, 2)$. Find the value $\mathbf{x}_1$ after one iteration.

3. Suppose we have a soft-margin SVM for which $\mathbf{w} = (-6, 3)$ and $b = 5$. Consider the example $(\mathbf{x} = (2, 1), y = -1)$.

   (a) How does the SVM classify $\mathbf{x}$?

   (b) Does $(\mathbf{x}, y)$ satisfy the SVM constraint? (Answer Yes or No.)

   (c) What is the hinge loss associated with $(\mathbf{x}, y)$?

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

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

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

(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	23
5	24

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

5. Here are questions about feature engineering.

   (a) Consider the data 0, 5, 5, 5, 5, which have these summary statistics:

   - minimum 0
   - mean 4
   - median 5
   - maximum 5
   - (population) standard deviation 2

   Do standardization rescaling on feature `x`:

(input) `x`	(output) `x_rescaled`
0	
5	
5	
5	
5	

   (b) Use one-hot encoding to transform the categorical feature `activity` into binary features with reasonable names.

(input) `activity`	(output)
swim	
bike	
run	
bike	

6. Consider 3-NN (three nearest neighbors) with the negative cosine similarity distance.

   (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 $\mathbf{x}$
$(-4, 3)$	1	
(0, 1)	0	
(1, 0)	0	
$(6, 8)$	1	

   Hint: Notice the "negative" in "negative cosine similarity."

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

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

7. In linear regression we minimize the *mean squared error* (MSE).

   (a) Find the MSE for the points $(0, 0)$ and $(1, 2)$ relative to the line $\hat{y} = f_{\mathbf{w},b}(\mathbf{x}) = \mathbf{w}\mathbf{x} + b$, where $\mathbf{w} = 2$ and $b = 3$.

   (b) For the best-fitting line for these data, $\mathbf{w} =$ ______ and $b =$ ______.

8. Mark each statement true or false by circling the appropriate choice.

   (a) TRUE / FALSE A support vector machine on 2D $\mathbf{x}$ with decision boundary $(3, 4)\cdot\mathbf{x}+5 = 0$ has a smaller margin between $+1$ and -1 support vectors than one with boundary $(5, 12) \cdot \mathbf{x} + 13 = 0$.

   (b) TRUE / FALSE Logistic regression sends an example $\mathbf{x}$ through a linear function to get a real number, which it sends through an exponential function to get a positive number, which it sends through a logistic function to predict a probability between 0 and 1. (Then it optionally compares that probability to a threshold to predict $\hat{y} = 0$ or $\hat{y} = 1$.)

   (c) TRUE / FALSE The linear regression model is not sensitive to the signs of labels $\{y_i\}$ because it minimizes mean *squared* error.

   (d) TRUE / FALSE A decision tree node containing the examples $\{(\mathbf{x}, y)\} =$
   $\{((1, 2), 0),((3, 4), 1),((5, 6), 0)\}$
   has lower entropy than the one containing the examples
   $\{((7, 8), 0),(9, 10), 1)\}$.

   (e) TRUE / FALSE Weighted kNN evaluates a new example $\mathbf{x}$ using $\mathbf{x}$'s k nearest neighbors weighted with weights proportional to the corresponding distances to $\mathbf{x}$.

   (f) TRUE / FALSE If we train an SVM on linearly separable data, then discard all support vectors, and then train a new SVM on the remaining examples, the first SVM will have a larger margin between $+1$ and -1 examples than the second.

   (g) TRUE / FALSE The hinge loss function of a soft-margin SVM gives a nonzero value for any $\mathbf{x}$ such that $\mathbf{wx} + b \geq 0$.

   (h) TRUE / FALSE If stochastic gradient descent (SGD) and gradient descent (GD) are both run in the same amount of time on a well-behaved function of some parameters over a large N of training examples each having a small D of features, SGD can use more iterations with a smaller learning rate α than GD.

   (i) TRUE / FALSE $1 + 1 = 2$

   (j) TRUE / FALSE Ridge regression tends to set most coefficients to zero.

   (k) TRUE / FALSE Training data are used to set parameters, validation data are used to choose hyperparameter settings and/or models, and test data are used to evaluate the chosen model.