STAT 451 Midterm Exam

Question	Points	Earned
Q1 (cover)	1	
Q2	14	
Q3	10	
Q4	12	
Q5	9	
Q6	8	
Q7	9	
Q8	4	
Q9	33	
Total	100	

1. If a question is ambiguous, resolve the ambiguity in writing. We will consider grading accordingly. e.g.

   - In #10, I think "average" refers to the population mean μ (not the sample mean $\bar{X}$).
   - In #13b, I think ...

   Please answer this question with a period (.) if you have no other comment, so that Canvas will think you answered it and give you its 1 point. Do not write unnecessary comments.

Answer each question at https://canvas.wisc.edu/courses/355816/quizzes/456707 **as you work through the exam** so that you do not run out of time with questions unanswered.

2. Consider finding the linear regression line by hand for the points $(0,0),(1,2),(2,1)$. Match each mathematical quantity on the left with its value on the right.

(a) $\mathbf{w}$ = _____ C: must be 2×1

(b) $\hat{\mathbf{y}}$ = _____ E: must be 3×1 and D is taken by $\mathbf{y}$

(c) $f_{\mathbf{w},b}(3)$ = _____ A: must be scalar and 0.5 is taken by b

(d) $\mathbf{X}$ = _____ G: must by 3×2

(e) intercept = _____ F: first element of $\mathbf{w}$

(f) $\mathbf{y}$ = _____ D: from data

(g) $\mathbf{X}^T\mathbf{X}$ = _____ B: must be 2×2

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

(A) 2

(B) $\begin{bmatrix} 3 & 3 \\ 3 & 5 \end{bmatrix}$

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

(D) $\begin{bmatrix} 0 \\ 2 \\ 1 \end{bmatrix}$

(E) $\begin{bmatrix} 0.5 \\ 1 \\ 1.5 \end{bmatrix}$

(F) 0.5

(G) $\begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \end{bmatrix}$

3. 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	ANSWER: -2
5	ANSWER: $\frac{1}{2}$

Or, if we mistakenly use the sample standard deviation $s = \sqrt{5} \approx 2.24$ instead of the population standard deviation $\sigma = 2$, we get this answer (which should receive only a tiny grading penalty, like 0.25 point):

`x`	`x_rescaled`
0	-1.79
5	0.45

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

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

ANSWER:

(input)	(output)		
`activity`	`swim`	`bike`	`run`
swim	1	0	0
bike	0	1	0
run	0	0	1
bike	0	1	0

What is the sum of your table's binary values? (I ask this for the sake of a Canvas question whose input is easy to enter.)

ANSWER:

4

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

S

x_1	x_2	y
2	7	0
1	4	1
0	6	1
3	5	0

(a) The entropy of this node in bits is _____.

ANSWER:

The node's y values are 0, 1, 1, 0, so $f_{ID3}(S) = \frac{1}{|S|}\sum_{(\mathbf{x},y)\in S} y = \frac{1}{4}(0+1+1+0) = \frac{1}{2}$.

$H(S) = -\frac{1}{2}(-1) - \frac{1}{2}(-1) = 1$

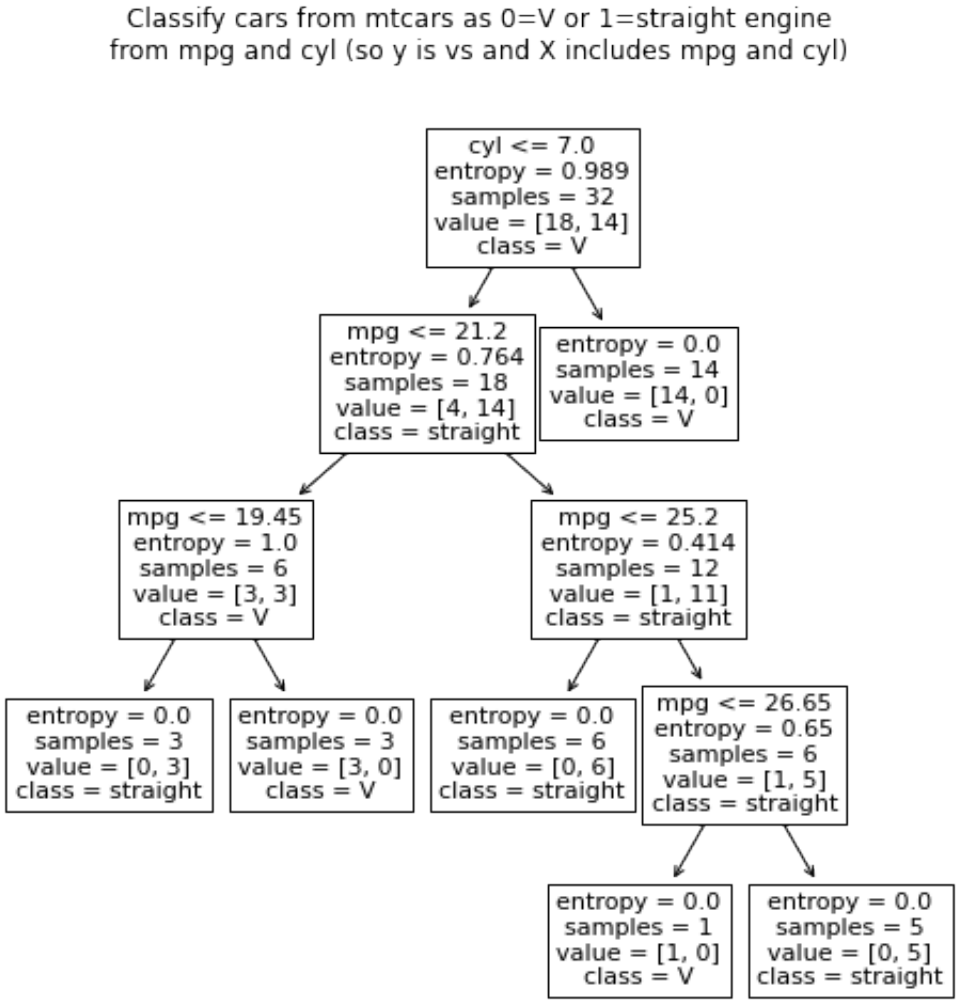
(Or, since a random draw from S amounts to a coin flip, the entropy is 1 bit.)

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

ANSWER:

Using feature $j = 1$ and threshold $t = 1.5$ (or any $t \in [1, 2)$) splits S into $S_- = \{(\mathbf{x}, y) \in S | x^{(j)} < t\}$ and its complement $S_+ = \{(\mathbf{x}, y) \in S | x^{(j)} \geq t\}$, each of which has entropy 0.

(c) Now consider this tree:



How does this tree classify a car whose gas mileage (mpg) is 19 and number of engine cylinders (cyl) is 6?

This car has a _____ engine.

ANSWER: straight

5. 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 $\mathbf{x}$?
   ANSWER:
   $\mathbf{wx} + b = (-6, 3) \cdot (2, 1) + 5 = -4 < 0$, so $\hat{y} = -1$.

   (b) Does $(\mathbf{x}, y)$ satisfy the SVM constraint? (Answer Yes or No.)
   ANSWER:
   Yes. It satisfies the constraint $\iff$ $y(\mathbf{wx} + b) \geq 1$ $\iff$ $(-1)((-6, 3) \cdot (2, 1) + 5) \geq 1$ $\iff$ $14 \geq 1$, which is true.

   (c) What is the hinge loss associated with $(\mathbf{x}, y)$?
   ANSWER:
   It satisfies the constraint, so the hinge loss is 0.
   Or, ignoring the constraint, the hinge loss is $\max(0, 1 - y(\mathbf{wx} + b)) = \max(0, 1 - (-1)((-6, 3) \cdot (2, 1) + 5)) = \max(0, -3) = 0$.

6. 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 $(-3, 1))$.
   ANSWER:
   $\hat{P}(y = 1|\mathbf{x}) = \dfrac{1}{1 + e^{-(\mathbf{wx}+b)}} = \dfrac{1}{1 + e^{-((1,2)\cdot(-3,1)+0)}} = \dfrac{1}{1 + e^{1}} \approx 0.269.$

   (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 FALSE A product of probabilities can overflow in fixed-precision computer arithmetic.
      ii. TRUE / FALSE TRUE A product of probabilities can underflow in fixed-precision computer arithmetic.
      iii. TRUE / FALSE TRUE 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 TRUE The natural log is strictly increasing, so maximizing the likelihood is the same as minimizing the negative log likelihood.
      v. TRUE / FALSE FALSE Using the natural log facilitates finding a closed-form expression for the coefficients in terms of the data.

7. 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	ANSWER: 0
$(0, 1)$	0	ANSWER: -0.8
$(1, 0)$	0	ANSWER: -0.6
$(6, 8)$	1	ANSWER: -1

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

(b) How does 3-NN classifiy $\mathbf{z}$?
ANSWER: 0

(c) What y value does 3-NN regression predict for $\mathbf{z}$?
ANSWER: $\hat{y} = \frac{1}{3}(0 + 0 + 1) = \frac{1}{3}$

8. Consider applying gradient descent with step size $\alpha = 0.5$ to find the $\mathbf{x}$ that minimizes the function $f(\mathbf{x}) = f\left((x^{(1)}, x^{(2)})\right) = \left(x^{(1)} - 3\right)^2 + \left(x^{(2)} - 4\right)^2$ starting from $\mathbf{x}_0 = (1, 2)$. Find the value $\mathbf{x}_1$ after one iteration.

ANSWER:

$\nabla f(\mathbf{x}) = \left(2(x^{(1)} - 3), 2(x^{(2)} - 4)\right)$, which is $(-4, -4)$ at $\mathbf{x} = (1, 2)$.

Move to $\mathbf{x}_1 = \mathbf{x}_0 - \alpha \nabla f(\mathbf{x}_0) = (1, 2) - 0.5(-4, -4) = (3, 4)$.

9. 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 $(5,12)\cdot\mathbf{x}+13=0$.

ANSWER: FALSE.

The margin for the first SVM is $\frac{2}{||\mathbf{w}||}=\frac{2}{\sqrt{3^2+4^2}}=\frac{2}{5}=0.4$, while the margin for the second is $\frac{2}{||\mathbf{w}||}=\frac{2}{\sqrt{5^2+12^2}}=\frac{2}{13}\approx 0.15$.

(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 $\hat{y}=1$.)

ANSWER:

FALSE. The logisitc function itself includes the exponential: $\sigma(t)=\frac{1}{1+e^{-t}}$. There is no other exponential function involved.

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

ANSWER: FALSE.

(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
$\{((7,8),0),(9,10),1)\}$.

ANSWER: TRUE.

The first has entropy $-\frac{1}{3}\log_2\frac{1}{3}-(1-\frac{1}{3})\log_2(1-\frac{1}{3})\approx 0.92$, while the second has entropy 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 $\mathbf{x}$.

ANSWER: FALSE. The weights are inversely proportion to the distances.

(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.

ANSWER: FALSE. The first will have a smaller margin than the second (unless there are no $+1$ or -1 examples remaining, in which case there is no second SVM).

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

ANSWER: FALSE.

It is zero when $y(\mathbf{w}\mathbf{x}+b)\geq 1$, so with $y=1$, it is zero when $\mathbf{w}\mathbf{x}+b\geq 1$.

(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.

ANSWER: TRUE

(i) TRUE / FALSE A loss function measures the error associated with one example while a cost function measures the error associated with a group of examples.

ANSWER: TRUE

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

ANSWER: FALSE

(Lasso 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.

ANSWER: TRUE