STAT 451 Midterm Exam NetID: ____________

Last name: ______________________________ First name: ________________________

Instructions:

1. Do not open the exam until I say "go."
2. Put away everything except a pencil or pen, a calculator, and your two one-page (two sides each) notes sheets.
3. Show your work. Correct answers without enough work may receive no credit.
4. If a question is ambiguous, resolve the ambiguity in writing. We will consider grading accordingly.
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. You are welcome to turn your exam in to me before I call time. However, if you are still here in the last five minutes, please remain seated until I've called time (to avoid disturbing peers).

Question	Points	Earned
Q0 (cover)	1	
Q1	9	
Q2	12	
Q3	6	
Q4	12	
Q5	3	
Q6	3	
Q7	4	
Total	50	

1. 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
4	9	1
2	6	0
5	7	0
3	8	1

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

ANSWER:

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

$$\begin{aligned} H(S) &= \sum_{y\in\{0,1\}} P(y)\,[-\log_2 P(y)] \\ &= -f_{ID3}(S)\log_2 f_{ID3}(S) - [1 - f_{ID3}(S)]\log_2 [1 - f_{ID3}(S)] \\ &= -\frac{1}{2}\log_2 \frac{1}{2} - \left(1 - \frac{1}{2}\right)\log_2\left(1 - \frac{1}{2}\right) \\ &= -\frac{1}{2}(-1) - \frac{1}{2}(-1) \\ &= 1 \end{aligned}$$

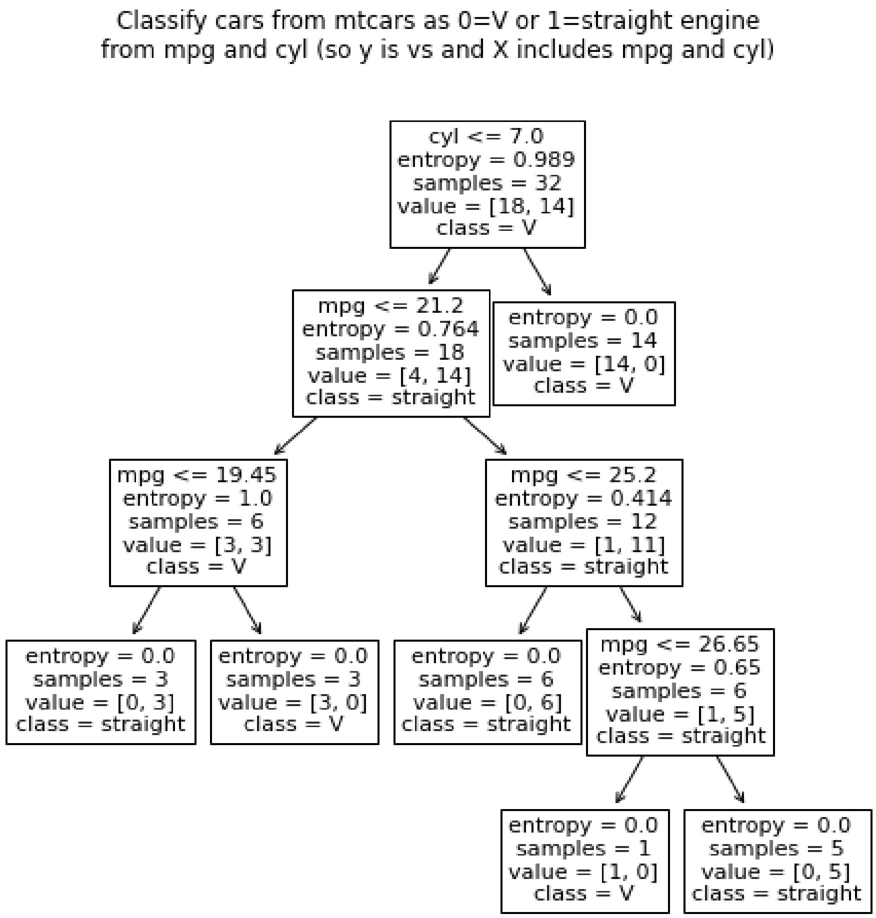
(Or, since a random draw from S amounts to a coin flip, the entropy is 1 bit.)

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

ANSWER:

Using feature $j = 2$ and threshold $t = 7.5$ (or any $t \in (7, 8]$) 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:



This tree says a car whose gas mileage (mpg) is 26 and number of engine cylinders (cyl) is 4 has a ____________ engine.

ANSWER: V

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

(a) TRUE / FALSE An SVM makes a classification error on $\mathbf{x}$ when $\mathbf{wx} + b \in (-1, 1)$ (i.e. between -1 and 1).

ANSWER: FALSE. It only makes an error when $\mathbf{x}$ is on the wrong side of $\mathbf{wx} + b = 0$.

(b) TRUE / FALSE In logistic regression, we model $P(y = 1)$ as a linear function of $\mathbf{x}$.

ANSWER: FALSE. We use $P(y_i = 1) = \frac{1}{1+e^{-(\mathbf{wx}+b)}}$, which is a nonlinear function of $\mathbf{x}$.

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

ANSWER: FALSE. If we use mean error, then any line through the centroid of two points would work equally well (because positive and negative errors cancel).

(d) TRUE / FALSE In a decision tree node, an entropy of 1 indicates all the node's examples have the same y value.

ANSWER: FALSE. Entropy 0 indicates a pure node.

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

ANSWER: TRUE. Use $k = 1$.

(f) TRUE / FALSE If we train an SVM on linearly separable data, then discard all training examples which are not support vectors, and then train a new SVM on the remaining examples, the first SVM will classify unseen examples better than the second.

ANSWER: FALSE. The two SVMs are the same.

3. Here are two questions about feature engineering.

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

(input) `weather`	(output)
sunny	
raining	
cloudy	
raining	

ANSWER:

(input) `weather`	(output) `sunny`	`cloudy`	`raining`
sunny	1	0	0
raining	0	0	1
cloudy	0	1	0
raining	0	0	1

(b) Do min-max rescaling on feature `x`:

(input) `x`	(output) `x_rescaled`
3	
1	
2	

ANSWER:

(input) `x`	(output) `x_rescaled`
3	1
1	0
2	0.5

4. Consider the logistic regression model,

$$P(y_i = 1) = \frac{1}{1 + e^{-(\mathbf{wx}+b)}}\,.$$

(a) Logistic regression is named after the log-odds of success, $\ln \frac{p}{1-p}$, where $p = P(y_i = 1)$. Show that this log-odds equals $\mathbf{wx} + b$.

ANSWER:

$$\begin{aligned}
\ln \frac{p}{1-p} &= \ln \frac{\frac{1}{1+e^{-(\mathbf{wx}+b)}}}{1 - \frac{1}{1+e^{-(\mathbf{wx}+b)}}} \\
&= \ln \frac{1}{\left(1 + e^{-(\mathbf{wx}+b)}\right) - 1} \\
&= \ln e^{\mathbf{wx}+b} \\
&= \mathbf{wx} + b, \text{ a linear function of } \mathbf{x}
\end{aligned}$$

(b) Match each function on the left that plays a role in the model with its image on the right. Hint: The *image* of a function is the set of all output values it may produce.

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

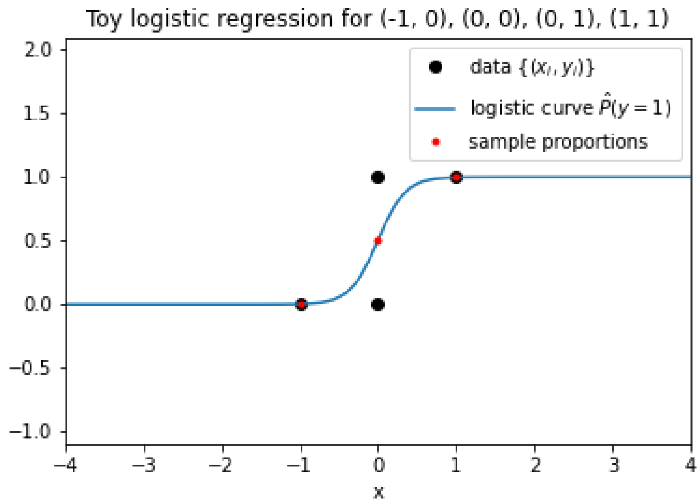
ii. ______ $f_2(t) = \frac{1}{1+e^{-t}}$ for $t \in \mathbb{R}$

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

1. $[0, 1]$, the interval from 0 to 1
2. $\mathbb{R}_+$, the positive real numbers
3. $\mathbb{R}$, the real numbers

ANSWER: (i) = 3, (ii) = 1, (iii) = 2

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



i. Match each code line on the left, with its output on the right.

_______ `model.intercept_`	1. array([0, 0, 0, 1])
_______ `model.coef_[0]`	2. array([0.003, 0.5, 0.5, 0.997])
_______ `model.predict(X)`	3. array([5.832])
_______ `model.predict_proba(X)[:, 1]`	4. array([0.])

ANSWER: 4, 3, 1, 2

ii. How do we classify a new point at $x = -0.5$ if using a decision threshold of 0.7?

_______ $\hat{y} = 0$

_______ $\hat{y} \approx 0.05$

_______ $\hat{y} \approx 0.95$

_______ $\hat{y} = 1$

ANSWER:
$\hat{y} = 0$. The graph shows $\hat{P}_{\mathbf{w},b}(y = 1|x = -0.5)$ is close to 0 (Python says ≈ 0.05), way less than the 0.7 threshold. So we assign $y = 0$. ($\hat{y} = 0.05$ and $\hat{y} = 0.95$ are not possible $\hat{y}$ labels.)

5. e.g. Consider using $\mathbf{w} = \left(X^T X\right)^{-1} X^T \mathbf{y}$ to find the line fitting the points $(0, -1)$ and $(2, 3)$.

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

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

(You should not continue the computation to find the line, which is $y = 2x - 1$.[1])

[1] What did one regression coefficient say to the other?
(This question was in a footnote in the notes, but I forgot to discuss it. You may answer if you wish, for 0 points.)

(a) ________ I'm partial to you.

(b) ________ We do not have a sense of humor we're aware of.

ANSWER: $X = \begin{bmatrix} 1 & 0 \\ 1 & 2 \end{bmatrix}$, $\mathbf{y} = \begin{bmatrix} -1 \\ 3 \end{bmatrix}$

6. Our hard-margin SVM used the constraints $\begin{cases} \mathbf{w}\mathbf{x}_i + b \geq 1 & \text{if } y_i = +1 \\ \mathbf{w}\mathbf{x}_i + b \leq -1 & \text{if } y_i = -1 \end{cases}$, for $i = 1, \ldots, N$.

   Consider a new model, SVM_{new}, that uses the constraints $\begin{cases} \mathbf{w}\mathbf{x}_i + b \geq 0 & \text{if } y_i = +1 \\ \mathbf{w}\mathbf{x}_i + b < 0 & \text{if } y_i = -1 \end{cases}$.

   Changing from the hard-margin SVM to SVM_{new} would do what to the margin?

   _______ Increase it.
   _______ Decrease it.
   _______ Leave it unchanged.
   _______ We cannot say without more information.

   ANSWER: Decrease it (to 0).

7. Consider a database consisting of these three examples:

name $\mathbf{x}$	age y
`Karolin`	20
`Kathrin`	30
`Kerstin`	40

   We want to estimate `Kathryn`'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 `Kathryn` and each of the other three names.

name	Hamming distance to `Kathryn`
`Karolin`	
`Kathrin`	
`Kerstin`	

   ANSWER:

name	Hamming distance to `Kathryn`
`Karolin`	4
`Kathrin`	1
`Kerstin`	5

   (b) Use 2-NN (two nearest neighbors) regression to estimate `Kathryn`'s age from her name.
   `Kathryn`'s age is about _______.
   ANSWER:
   `Kathryn`'s two nearest neighbors are `Kathrin` and `Karolin`. We estimate `Kathryn`'s age as the average of those neighbors' ages, that is, $\frac{1}{2}(20 + 30) = 25$.

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