STAT 451 Midterm Exam
(1 point for easily legible writing on this cover sheet.)

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 one-page (two sides) notes sheet.

3. Show your work. Correct answers without at least a minimal version of the work normally required 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 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).

<table>
<thead>
<tr>
<th>Question</th>
<th>Points</th>
<th>Earned</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q0 (cover)</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Q1</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>Q2</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>Q3</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>Q4</td>
<td>19</td>
<td></td>
</tr>
<tr>
<td>Q5</td>
<td>15</td>
<td></td>
</tr>
<tr>
<td>Q6</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>Q7</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>Q8</td>
<td>15</td>
<td></td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>100</strong></td>
<td></td>
</tr>
</tbody>
</table>
1. Consider a logistic regression model with \( w = (3, -2) \) and \( b = 1 \).

(a) (8 points) Find \( \hat{P}(y = 1|x = (1, 2)) \).

ANSWER:

The model is \( \hat{P}(y = 1|x) = \frac{1}{1 + e^{-(wx+b)}} = \frac{1}{1 + e^{-(3, -2)\cdot(1, 2)+1}} = \frac{1}{1 + e^0} = \frac{1}{2} \).

(b) (2 points) For the point \( x = (1, 0) \), I did the arithmetic and found \( \hat{P}(y = 1|x) \approx 0.98 \).

How do we classify this \( x \) using a decision threshold of 0.5?

ANSWER:

\( \hat{y} = 1 \) because 0.98 > 0.5.

(Here’s a check of \( \hat{P}(y_i = 1|x) = \frac{1}{1 + e^{-(wx+b)}} = \frac{1}{1 + e^{-(3, -2)\cdot(1, 0)+1}} = \frac{1}{1 + e^1} \approx 0.98 \).)

2. Consider the gradient descent algorithm.

(a) (2 points) Consider applying gradient descent with step size \( \alpha = 0.1 \) to find the \( x \) that minimizes the function \( f(x) = f ((x^{(1)}, x^{(2)})) = (x^{(1)} - 1)^2 + (x^{(2)} - 2)^2 \) starting from \( x_0 = (0, 0) \). Find the value \( x_1 \) after one iteration.

ANSWER:

\( \nabla f(x) = (2(x^{(1)} - 1), 2(x^{(2)} - 2)) \), which is \( (-2, -4) \) at \( x = (0, 0) \).

Move to \( x_1 = x_0 - \alpha \nabla f(x_0) = (0, 0) - (-0.2, -0.4) = (0.2, 0.4) \).

(b) (8 points) Mark each statement as true (T) or false (F).

_____ Gradient descent can fail to converge on a convex function if step size \( \alpha \) is such that we get stuck in a cycle, oscillating between two or several values. ANSWER: T

_____ For a non-convex function, gradient descent can fail to converge by descending without bound. ANSWER: T

_____ Gradient descent can fail to converge on a convex function if it gets stuck in a local minimum. ANSWER: F

_____ Gradient descent can fail to converge on a convex function if the step size \( \alpha > 0 \) is too small. ANSWER: F
3. Each graph below shows six training examples for which $y$ is binary along with the decision boundary of a classifier trained on those examples.

Alas, the printed version of the exam revealed the answers to this question because of my editing mistake.

- (8 points) Match each graph, with the letter corresponding to the classifier (further below) that produced that graph. That is, write one of “a” through “d” in each of the blanks. **ANSWER:**

- The hard-margin linear SVM has a linear boundary, so top-left is (c).
- 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 lines, so bottom-left is (a).
- Both SVM with RBF kernel and 3-NN allow a nonlinear boundary. To separate them, note that 3-NN classifies a point just right of the right-most blue dot as red (because its 3 nearest neighbors are blue, red, red), so bottom-right must be (b). That leaves top-right as (d).

(a) `DecisionTreeClassifier(criterion='entropy', max_depth=None, random_state=0)`
(b) `KNeighborsClassifier(n_neighbors=3, metric='euclidean')`
(c) `svm.SVC(kernel='linear', C=1000)`
(d) `svm.SVC(kernel='rbf', C=1, gamma=1)`

- (2 points) Mark with an “X” below any of the graphs above that could have been produced by

Alas, the printed version of the exam had “`KNeighborsClassifier(n_neighbors=3, metric='euclidean')`” instead of the logistic regression I intended because of my editing mistake. The correct answer to the printed version was “bottom right”.

```
linear_model.LogisticRegression(C=1000).
```

- top left X
- top right
4. Here are questions about feature engineering.

(a) (8 points) Use one-hot encoding to transform the categorical feature animal into binary features with reasonable names.

<table>
<thead>
<tr>
<th>(input)</th>
<th>(output)</th>
</tr>
</thead>
<tbody>
<tr>
<td>animal</td>
<td></td>
</tr>
<tr>
<td>dog</td>
<td></td>
</tr>
<tr>
<td>bird</td>
<td></td>
</tr>
<tr>
<td>dog</td>
<td></td>
</tr>
<tr>
<td>lamb</td>
<td></td>
</tr>
</tbody>
</table>

ANSWER:

<table>
<thead>
<tr>
<th>(input)</th>
<th>(output)</th>
</tr>
</thead>
<tbody>
<tr>
<td>animal</td>
<td>bird</td>
</tr>
<tr>
<td></td>
<td>dog</td>
</tr>
<tr>
<td></td>
<td>lamb</td>
</tr>
</tbody>
</table>

| dog     | 0 1 0   |
| bird    | 1 0 0   |
| dog     | 0 1 0   |
| lamb    | 0 0 1   |

(b) (6 points) Do min-max rescaling on feature x:

<table>
<thead>
<tr>
<th>(input)</th>
<th>(output)</th>
</tr>
</thead>
<tbody>
<tr>
<td>x</td>
<td>x_rescaled</td>
</tr>
<tr>
<td>11</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td></td>
</tr>
</tbody>
</table>

ANSWER:

<table>
<thead>
<tr>
<th>(input)</th>
<th>(output)</th>
</tr>
</thead>
<tbody>
<tr>
<td>x</td>
<td>x_rescaled</td>
</tr>
<tr>
<td>11</td>
<td>1/4</td>
</tr>
<tr>
<td>10</td>
<td>0</td>
</tr>
<tr>
<td>14</td>
<td>1</td>
</tr>
</tbody>
</table>

(c) (5 points) What numbers are printed by this code? (I am not worried about the exact python structure in which they are contained.)

```python
import numpy as np
from sklearn.impute import SimpleImputer

X = np.array([1, 3, np.nan, 2]).reshape(-1, 1)
imp = SimpleImputer(missing_values=np.nan, strategy='mean', fill_value=None)
X_transformed = imp.fit_transform(X)
print(X_transformed)
```
The numbers printed are ________________________________.

ANSWER:
The output is an array containing 1, 3, 2, (the mean of the non-missing values) and 2.
5. Consider 3-NN (three nearest neighbors) with the Minkowski distance using \( p = 2 \).

(a) (8 points) Find the distance from \( z = (3, 4) \) to each of the other points \( x \):

\[
\begin{array}{c|c|c}
\text{Point} & y & \text{Distance from } z \text{ to } x \\
(0, 0) & 0 & \text{ANSWER: 5} \\
(1, 4) & 1 & \text{ANSWER: 2} \\
(3, -2) & 1 & \text{ANSWER: 6} \\
(3, 0) & 0 & \text{ANSWER: 4} \\
\end{array}
\]

(b) (3 points) How does 3-NN classify \( z \)?

ANSWER: 0

(c) (4 points) How does weighted 3-NN classify \( z \)?

ANSWER: 1

6. Consider these training examples (with 1D \( x = x \)):

\[
\begin{array}{c|c}
x & y \\
0 & 2 \\
1 & 1 \\
2 & 0 \\
\end{array}
\]

(a) (5 points) Draw the linear regression model fitted to these data over the whole \( x \) domain \((-1, 3)\) of the graph on the left, below.

(b) (5 points) Draw the 2-NN regression model fitted to these data over the whole \( x \) domain \((-1, 3)\) of the graph on the right, above.
ANSWER:
7. In linear regression we minimize the mean squared error (MSE).

(a) (6 points) Find the MSE for the points (0, 0) and (2, 0) relative to the line \( \hat{y} = f_{w,b}(x) = wx + b \), where \( w = 1 \) and \( b = 0 \).

\[ \text{MSE} = \frac{1}{N} \sum_{i=1}^{N} [f_{w,b}(x_i) - y_i]^2 \]
\[ = \frac{1}{2} \sum_{i=1}^{2} [(1x + 0) - y_i]^2 \]
\[ = \frac{1}{2} \left( [(1 \cdot 0 + 0) - 0]^2 + [(1 \cdot 2 + 0) - 0]^2 \right) \]
\[ = 2 \]

(b) (4 points) For the best-fitting line for these data, \( w = \) _____ and \( b = \) _____.

ANSWER:
Since there are only two points, the best line goes through the two points (without regard for the regression machinery). It has slope \( w = 0 \) and intercept \( b = 0 \).

8. Consider a decision tree node containing the following examples \{\((x, y)\)\}, where \( x \) has only one feature, \( x_1 \).

\[
\begin{array}{c|c}
  x_1 & y \\
  \hline
  2 & 0 \\
  4 & 1 \\
  6 & 1 \\
  8 & 0 \\
\end{array}
\]

(a) (3 points) What is the information content associated with drawing \( y = 0 \) from this node?

ANSWER: \( I(0) = -\log_2 P(0) = -\log_2 \frac{1}{2} = 1 \).
(b) (6 points) The entropy of this node in bits is _______.

**ANSWER:**

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

$$H(S) = \sum_{y \in \{0,1\}} \hat{P}(y) \left[ - \log_2 \hat{P}(y) \right]$$

$$= -f_{ID3}(S) \log_2 f_{ID3}(S) - [1 - f_{ID3}(S)] \log_2 [1 - f_{ID3}(S)]$$

$$= -\frac{1}{2} \log_2 \frac{1}{2} - \left(1 - \frac{1}{2}\right) \log_2 \left(1 - \frac{1}{2}\right)$$

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

$$= 1$$

(c) (6 points) A (feature, threshold) pair $(j, t)$ that yields a best split for this node is feature $j = 1$ and threshold $t = _______.

**ANSWER:**

There is only one feature, $j = 1$. Possible thresholds are $t = 3, 5, 7$.

- $t = 3$ yields $H(S_-, S_+) = \frac{1}{3}[0] + \frac{2}{3}\left[\frac{2}{3}(- \log_2 \frac{2}{3}) + \frac{1}{3}(- \log(\frac{1}{3})\right] \approx 0.689$.

- $t = 5$ yields $H(S_-, S_+) = \frac{1}{2}[1] + \frac{1}{2}[1] = 1$.

- $t = 7$ gives the same answer as $t = 3$ by symmetry.

A best threshold is the one with the minimum $H(S_-, S_+)$, which is $t = 3$ (or any $t \in (2, 4)$) or $t = 7$ (or any $t \in (6, 8)$).