STAT 451 Midterm Exam

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

NetID: _____

Last name: _____ First name: _____

Mark your lecture with "X":

_____ TuTh 1:00-2:15

_____ TuTh 2:30-3:45

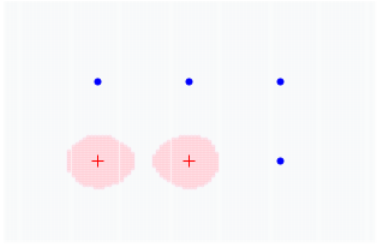
Instructions:

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 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).

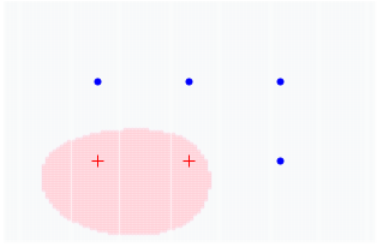
Question	Points	Earned
Q0 (cover)	1	
Q1	7	
Q2	3	
Q3	6	
Q4	6	
Q5	6	
Q6	6	
Q7	6	
Q8	4	
Q9	5	
Total	50	

1. Each graph below shows six training examples for which y is binary along with the decision boundary of a classifier trained on those examples.

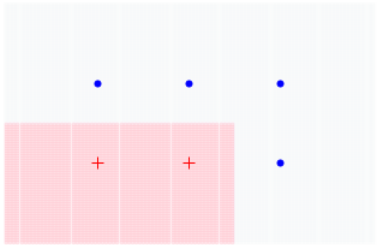
   Match each graph, with the letter corresponding to the classifier (further below) that produced that graph. That is, write one of "a" through "f" in each of the blanks. ANSWER:



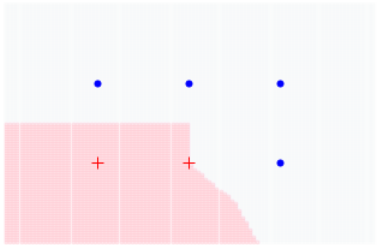
_____ e



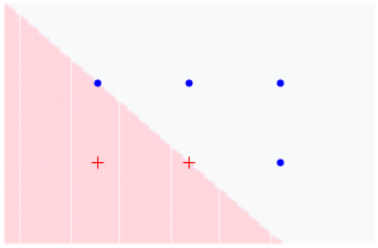
_____ f



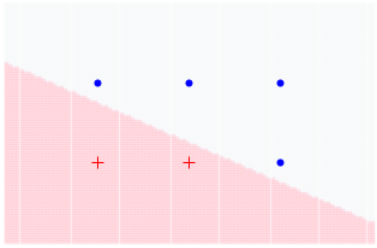
_____ a



_____ b



_____ c



_____ d

Identifying the two bottom plots is difficult. The hard-margin SVM (obtained by a high $C = 1000$) must be the one on the right because the hard-margin SVM does not have points on its decision boundary, as the left plot does.

That means the low $C = 1$ SVM must be the one on the left. At first glance, I thought its road width is close to 0, while decreasing C from 1000 to 1 should yield a wider road, so I was puzzled. Then I realized the road isn't visible, only its center. (In fact, the road on the left plot extends left to the left red plus and right to the top-middle and bottom blue dots.)

(a) `DecisionTreeClassifier(criterion='entropy', max_depth=None, random_state=0)`

(b) `KNeighborsClassifier(n_neighbors=3, metric='euclidean')`

(c) `svm.SVC(kernel='linear', C=1)`

(d) `svm.SVC(kernel='linear', C=1000)`

(e) `svm.SVC(kernel='rbf', C=1, gamma=6)`

(f) `svm.SVC(kernel='rbf', C=1, gamma=1)`

Mark with an "X" below any of the graphs in (a) that could possibly have been produced by `linear_model.LogisticRegression(C=1000)`.

_____ top left

_____ top right

_____ center left

_____ center right

_____ bottom left X

_____ bottom right X

The logistic regression model depends on $\mathbf{x}$ only via the linear expression $\mathbf{w}\mathbf{x} + b$, so it has a linear decision boundary. (In fact the bottom right plot is the one for both (d) and logistic regression; but I don't know how to exclude the bottom left plot, since it also has a linear boundary, so we will give credit for either or both of "bottom left" and "bottom right".)

2. Consider three multiple linear regression models, one using ordinary least squares (OLS), one using lasso, and one using ridge regression (each with the `scikit-learn` defaults). Each is trained on a random half of the rows of `mtcars` using y as the `mpg` column and X as the other ten columns. Each is tested on the other half of `mtcars`.

   Answer each of the following questions with one of "OLS", "lasso", or "ridge". (You may use an answer more than once, if you wish.)

   (a) For which model do we expect `MSE_train` to be the smallest? _____ ANSWER: OLS

   (b) For which model do we expect `MSE_test` to be the smallest? _____ ANSWER: ridge

   (c) For which do we expect `np.sum(np.abs(model.coef_) > 0)` to be smallest? _____ ANSWER: lasso

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

x_1	y
1	0
3	1
5	0
7	0

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

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

$$
\begin{aligned}
H(S) &= \sum_{y \in \{0,1\}} \hat{P}(y) \left[-\log_2 \hat{P}(y) \right] \\
&= -f_{ID3}(S) \log_2 f_{ID3}(S) - [1 - f_{ID3}(S)] \log_2 [1 - f_{ID3}(S)] \\
&= -\frac{1}{4} \log_2 \frac{1}{4} - \left(1 - \frac{1}{4}\right) \log_2 \left(1 - \frac{1}{4}\right) \\
&= -\frac{1}{4}(-2) - \frac{3}{4}(-0.4150375) \\
&\approx 0.811
\end{aligned}
$$

(b) There is only one feature, $j = 1$. The (feature, threshold) pair (j, t) that yields the best split for this node is feature $j = 1$ and threshold $t =$ _____.

ANSWER:
Possible thresholds are $t = 2, 4, 6$.
$t = 2$ yields $H(S_-, S_+) = \frac{1}{4}[0] + \frac{3}{4}\left[\frac{2}{3}(-\log_2 \frac{2}{3}) + \frac{1}{3}(-\log(\frac{1}{3})\right] \approx 0.689$.
$t = 4$ yields $H(S_-, S_+) = \frac{2}{4}\left[\frac{1}{2}(-\log_2(\frac{1}{2}) + \frac{1}{2}(-\log_2(\frac{1}{2})\right] + \frac{2}{4}[0] = 0.5$.
$t = 6$ yields the same $H(S_-, S_+)$ as $t = 2$.
The best threshold is the one with the minimum $H(S_-, S_+)$, which is $t = 4$ (or any $t \in (3, 5)$).

4. Consider a logistic regression model with $\mathbf{w} = (-3, 3)$ and $b = 5$.

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

   ANSWER:
   The model is $\hat{P}(y = 1|\mathbf{x}) = \frac{1}{1+e^{-(\mathbf{wx}+b)}} = \frac{1}{1+e^{-((-3,3)\cdot(2,1)+5)}} = \frac{1}{1+e^{-2}} \approx 0.881$.

   (b) For the point $\mathbf{x} = (3, 1)$, I did the arithmetic and found $\hat{P}(y = 1|\mathbf{x}) \approx 0.269$. How do we classify $\mathbf{x} = (3, 1)$ using a decision threshold of 0.5?

   ANSWER:
   $\hat{y} = 0$ because $0.269 < 0.5$.
   (Here's a check of $\hat{P}(y_i = 1|\mathbf{x}) = \frac{1}{1+e^{-(\mathbf{wx}+b)}} = \frac{1}{1+e^{-((-3,3)\cdot(3,1)+5)}} = \frac{1}{1+e^{1}} \approx 0.269$.)

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

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

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

   (b) Does $(\mathbf{x}, y)$ satisfy the SVM constraint?

   ANSWER:
   No. It satisfies the constraint $\iff y(\mathbf{wx} + b) \geq 1 \iff (-1)((-3, 3) \cdot (2, 1) + 5) \geq 1 \iff -2 \geq 1$, which is false.

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

   ANSWER:
   The hinge loss is $\max(0, 1 - y(\mathbf{wx} + b)) = \max(0, 1 - (-1)((-3, 3) \cdot (2, 1) + 5)) = \max(0, 3) = 3$.

6. Here are some questions on algorithm selection.

   (a) Consider a classifier already trained on a large number N of examples and a small number D of features. Which classifier is likely to be slowest in making a prediction on an unseen example? (Circle one.)

   i. SVM
   ii. k-NN
   iii. Logistic regression
   iv. Balanced decision tree (a *balanced* tree arises when each split of a node leads to two subnodes of about the same size; it has depth about $\log_2 N + 1$)

   ANSWER: ii, because k-NN must visit all N examples.

   (b) Choose the best match of properties to regression variants.

   Here are the properties:

   _____ Models a probability in $[0, 1]$.
   _____ Tends to do feature selection by setting some of the coefficients in $\mathbf{w}$ to zero.
   _____ Can use gradient descent or stochastic gradient descent to optimize its coefficients.
   _____ Has a greater tendency to overfit training data than the other two linear methods.

   Here are the regression variants:

   i. ridge
   ii. ordinary least squares
   iii. lasso
   iv. logistic regression

   ANSWER: iv, iii, i, ii

7. Consider these training examples:

$\mathbf{x}$	y
(0, 0)	1
(1, 2)	1
(2, 3)	0
(5, 0)	0

   Hint: Draw them.

(a) Find the Manhattan distance between (2, 1) and each of the other points:

x	Manhattan distance from (2, 1) to **x**
(0, 0)	
(1, 2)	
(2, 3)	
(5, 0)	

ANSWER:

x	Manhattan distance from (2, 1) to **x**
(0, 0)	3
(1, 2)	2
(2, 3)	2
(5, 0)	4

(b) How does 3-NN (three nearest neighbors) classifiy (2, 1)?

ANSWER: 1

8. Here are two questions about feature engineering.

   (a) Consider the data 3, 1, 2, which have these summary statistics:

   - population minimum 1
   - population mean 2
   - population median 2
   - population maximum 3
   - population standard deviation $\sqrt{2/3} \approx 0.816$

   Do standardization rescaling on feature `x`:

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

   ANSWER:

`x`	`x_rescaled`
3	1.225
1	-1.225
2	0

   Or, if we use the sample standard deviation $s = 1$ instead of the population standard deviation $\sigma \approx 0.816$, we get this answer (which also received full credit):

`x`	`x_rescaled`
3	1
1	-1
2	0

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

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

X = np.array([[3], [np.nan], [2]])
imp = SimpleImputer(missing_values=np.nan, strategy='median', fill_value=None)
X_transformed = imp.fit_transform(X)
print(X_transformed)
```

   The numbers printed are _____.

   ANSWER:

   The output is an array containing 3, 2.5 (the median of the non-missing values) and 2.

9. Mark each statement "T" if it is true or "F" if it is false.

(a) _____ The hinge loss function of a soft-margin SVM gives a nonzero value for any $\mathbf{x}$ such that $\mathbf{w}\mathbf{x} + b \in (-1, 1)$ (i.e. between -1 and 1).

ANSWER: T = true

(b) _____ Consider training a sequence of N decision trees on $S = \{(\mathbf{x}_i, y_i) : i = 1, \ldots, N$ and $\mathbf{x}_i \neq \mathbf{x}_j$ for all i, j$\}$. Suppose the ith tree is permitted maximum depth i and its accuracy on S is a_i. Then $a_1 < a_2 < a_3 < \ldots < a_N$.

ANSWER: F = false. For example, increasing depth past that required to get nodes with entropy 0 does not improve accuracy.

(c) _____ Increasing k in k-NN regression leads to overfitting the training data.

ANSWER: F = false. Decreasing k toward $k = 1$ leads to overfitting.

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

ANSWER: T = true. Disarding support vectors allows the "road" to widen past those discarded examples.

(e) _____ 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: T = true.