

First name Last name (please write clearly so Gradescope’s OCR can read your name):

KEY

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 two one-page (two sides each) notes sheets.
4. For questions with circles in front of the answers, fill in one circle completely. Ok?
 - ☒ Yes
 - ☐ No
5. Show your work. Correct answers without at least a minimal version of the work normally required may receive no credit. For a question with an answer box “,” write only the answer in the box; work should be outside the box.
6. If you continue writing or do not turn in your exam when I say time is up, you risk a penalty. (The alternative, that you get more time than your peers, is unfair.)
7. 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)	0	
Q1	6	
Q2	8	
Q3	10	
Q4	9	
Q5	4	
Q6	8	
Q7	6	
Q8	6	
Q9	6	
Q10	12	
Total	75	

1. Consider kernel density estimation on the data $\{x_i\} = \{0, 1, 2\}$.

(a) Decreasing the bandwidth b toward 0 makes the density estimate at $x = 1$ tend toward:

- ☐ 0
☐ $\frac{1}{3}$
☐ $\frac{1}{2}$
☐ 1
☐ 3
☐ ∞ **ANSWER: ●**

(b) Decreasing the bandwidth b toward 0 makes the density estimate at $x = 1.5$ tend toward:

- ☐ 0 **ANSWER: ●**
☐ $\frac{1}{3}$
☐ $\frac{1}{2}$
☐ 1
☐ 3
☐ ∞

2. Consider the use of bagging applied to classification decision trees of depth 1 (one decision node and two leaf nodes per tree). A training data set, on the left, consists of $\{(\mathbf{x}, y)\} = \{(x, y)\}$ because $\mathbf{x}$ has only one feature, x . It is followed by $B = 3$ bootstrap resamples created by sampling with replacement from the training data.

Training data	Resample #1	Resample #2	Resample #3
$x \mid y$	$x \mid y$	$x \mid y$	$x \mid y$
1 1	1 1	1 1	1 1
2 0	2 0	1 1	1 1
3 1	4 0	3 1	2 0
4 0	4 0	4 0	2 0

Consider making a prediction for $\mathbf{x} = 2$.

(a) What prediction is made by the tree trained on Resample #1? $\hat{y} =$

ANSWER:

The tree uses threshold $t = 1.5$ and predicts $\hat{y} = 0$.

(b) What prediction is made by the tree trained on Resample #2? $\hat{y} =$

ANSWER:

The tree uses threshold $t = 3.5$ and predicts $\hat{y} = 1$.

(c) What prediction is made by the tree trained on Resample #3? $\hat{y} =$

ANSWER:

The tree uses threshold $t = 1.5$ and predicts $\hat{y} = 0$.

(d) What prediction is made by this bagging classifier? $\hat{y} =$

ANSWER:

The bagging classifier predicts $\hat{y} = 0$, the most frequent of the $B = 3$ predictions.

3. For each part, select all that apply. (Part (a), for example, has four bubbles, each of which must be filled or not filled correctly to earn credit for (a).)

(a) Which model(s) can be trained using gradient decent?

- ☐ kNN
- ☐ logistic regression ●
- ☐ linear regression ●
- ☐ decision tree

(b) Which model(s) require looping through the data at least once for training?

- ☐ kNN
- ☐ logistic regression ●
- ☐ linear regression ●
- ☐ decision tree ●
- ☐ SVM ●

(c) Which model(s) have a fixed number of parameters?

- ☐ kNN ●
- ☐ logistic regression ●
- ☐ linear regression ●
- ☐ decision tree
- ☐ SVM ●

(d) Which model(s) can exhibit improved performance by scaling features?

- ☐ kNN ●
- ☐ logistic regression ●
- ☐ linear regression ●
- ☐ decision tree
- ☐ SVM ●

(e) Which model(s) requires the least amount of time to train?

- ☐ kNN ●
- ☐ logistic regression
- ☐ linear regression
- ☐ decision tree
- ☐ SVM

4. Mark each statement true or false.

- (a) In kernel density estimation we use a weighted average of the y 's, where the weights come from Gaussians centered at the x 's, to determine the height of an approximate density curve at each x .

☐ True

☐ False ANSWER: ●; there are no y 's in KDE

- (b) Increasing n (the minimum number of examples in a cluster) or decreasing ϵ (the maximum distance between two examples for them to be considered neighbors) makes DBSCAN find denser clusters.

☐ True ANSWER: ●

☐ False

- (c) When principal component analysis retains some number p of principal components of a D -dimensional data set (where $0 < p < D$), those p components are the subset of p features that explain the most variance in the original data.

☐ True

☐ False ANSWER: ●; while each principal component is a linear combination of features, it is not typically a feature.

5. For each situation, indicate which hyperparameter search strategy, G = grid search or R = random search, is more likely to be successful. Suppose computation time is limited.

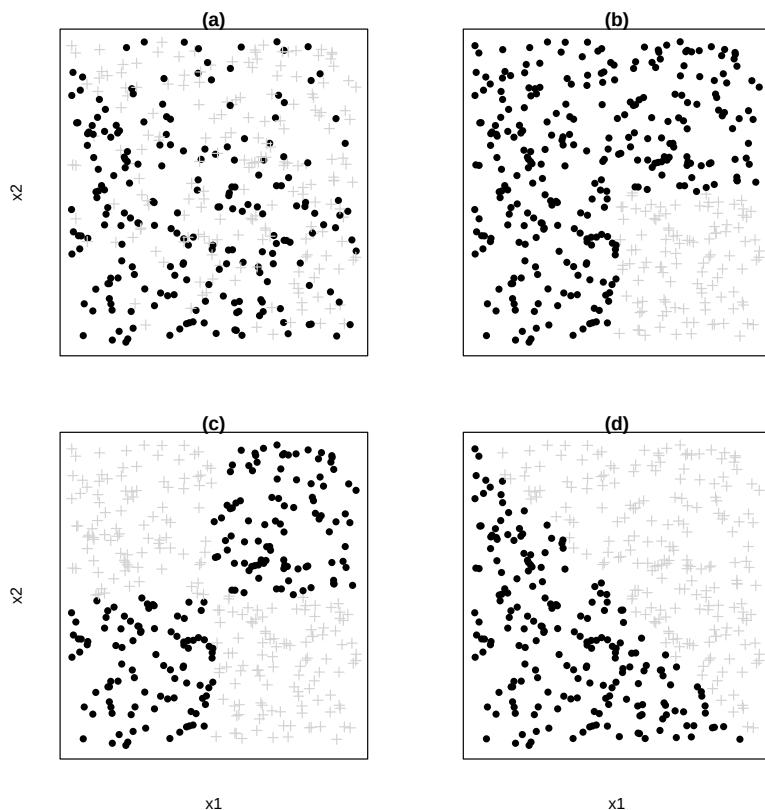
- (a) _____ A model has two hyperparameters. The first takes one of two string values and the other takes one of three numeric values.

G = grid search because there are only 6 combinations to try and there's no need to risk random search missing one.

- (b) _____ A model has two hyperparameters. The first takes a floating-point number in the interval $[0, 1]$ while the second takes an integer in the range $[0, 100000]$.

R = random search because there are infinitely many combinations and we cannot articulate very many or hope we are guessing good combinations for a grid search.

6. Match the four data sets to the four classification models to produce the best overall accuracy. Use each model once. The features are x_1 and x_2 . The points are colored by class: $y = 1$ is a gray “+” and $y = 0$ is a black dot.



- (a) ☐ kNN
☐ logistic regression
☐ SVM with your choice of linear or RBF kernel ☒; use RBF kernel for nonlinear data
☐ decision tree with maximum depth 3
- (b) ☐ kNN
☐ logistic regression
☐ SVM with your choice of linear or RBF kernel
☐ decision tree with maximum depth 3 ☒
- (c) ☐ kNN ☒; note that no decision tree split can reduce entropy and the data are not linearly separable
☐ logistic regression
☐ SVM with your choice of linear or RBF kernel
☐ decision tree with maximum depth 3
- (d) ☐ kNN
☐ logistic regression ☒; note that decision tree would require depth much higher than 3
☐ SVM with your choice of linear or RBF kernel

☐ decision tree with maximum depth 3

7. Suppose we use a stacking model to do multiclass classification such that:

- The data are:

$\mathbf{x}$	y
(0, 2)	1
(1, 1)	0
(2, 3)	2
(4, 1)	0
(2, 1)	1
(3, 2)	2

- The first base model is (multinomial) logistic regression.
- The second base model is k -NN.
- The meta-model is a decision tree.

(a) The first model is trained on how many features?

- ☐ 0
☐ 1
☒ 2 ANSWER: ● (It is trained on the $\mathbf{x}$'s.)
☐ 3
☐ 4
☐ 5
☐ 6

(b) The second model is trained on how many features?

- ☐ 0
☐ 1
☒ 2 ANSWER: ● (It is trained on the $\mathbf{x}$'s.)
☐ 3
☐ 4
☐ 5
☐ 6

(c) The third model is trained on how many features?

- ☐ 0
☐ 1
☐ 2
☐ 3
☐ 4
☐ 5
☒ 6 ANSWER: ● (Its training examples are concatenated output probability vectors from the two base models, each of which has length 3, for a total length of 6.)

8. Consider using k -means on the unsupervised 1D dataset $\{\mathbf{x}\} = \{1, 3, 5, 10, 12\}$ to create $k = 2$ clusters. Suppose the two initial randomly-chosen cluster centroids are $\mathbf{c}_1 = 3$ and $\mathbf{c}_2 = 5$.

(a) What are the centroids after the first iteration of k -means?

$\mathbf{c}_1 = \underline{\hspace{2cm}}$ and $\mathbf{c}_2 = \underline{\hspace{2cm}}$.

ANSWER:

1 and 3 are closest to $\mathbf{c}_1 = 3$, so the new $\mathbf{c}_1 = \frac{1}{2}(1 + 3) = 2$.

5, 10, and 12 are closest to $\mathbf{c}_2 = 5$, so the new $\mathbf{c}_2 = \frac{1}{3}(5 + 10 + 12) = 9$.

(b) What are the centroids after the second iteration?

$\mathbf{c}_1 = \underline{\hspace{2cm}}$ and $\mathbf{c}_2 = \underline{\hspace{2cm}}$.

ANSWER:

1, 3, and 5 are closest to $\mathbf{c}_1 = 2$, so the new $\mathbf{c}_1 = \frac{1}{3}(1 + 3 + 5) = 3$.

10 and 12 are closest to $\mathbf{c}_2 = 9$, so the new $\mathbf{c}_2 = \frac{1}{2}(10 + 12) = 11$.

9. Consider the use of gradient boosting to train a regression model on the following data:

x	y
1	7
2	6
3	5

(a) What prediction $\hat{y} = f_0(x)$ is given by the first model at $x = 2$?

$\hat{y} =$

ANSWER:

$\hat{y} = 6$ because $f_0(x)$ is a constant model given by $f_0(x) = \frac{1}{N} \sum_{i=1}^N y_i = \frac{1}{3}(7 + 6 + 5) = 6$.

(b) What data are used to train the second model $\hat{y} = f_1(x)$?

x	y

ANSWER:

x	y
1	1
2	0
3	-1

f_1 is trained on the original data after replacing the y values with the residuals with respect to the first model, that is with $\{e_i = y_i - f_0(x_i)\}$. Since $f_0(x) = 6$ for all x , $\{e_i\} = \{7 - 6, 6 - 6, 5 - 6\} = \{1, 0, -1\}$.

10. Consider the following questions about model assessment.

- (a) Consider a classifier trained on examples $(\mathbf{x}, y)$ in the first two columns of the table below that makes the predictions on training data in the third column.

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

Complete the corresponding confusion matrix:

actual y	predicted $\hat{y}$	
	0	1
0		
1		

ANSWER:

actual y	predicted $\hat{y}$	
	0	1
0	1	1
1	1	1

- (b) The classifier is evaluated on unseen test data yielding this confusion matrix:

actual y	predicted $\hat{y}$	
	0	1
0	3	2
1	2	3

What is the precision on the test data?

ANSWER: The matrix is

actual y	predicted $\hat{y}$	
	0	1
0	TN	FP
1	FN	TP

 so $precision = \frac{TP}{TP + FP} = \frac{3}{3 + 2} = \frac{3}{5} = 0.6$

- (c) What is the recall on the test data?

ANSWER:

$$recall = \frac{TP}{TP + FN} = \frac{3}{3 + 2} = \frac{3}{5} = 0.6$$

- (d) What is the accuracy on the test data?

ANSWER:

$$accuracy = \frac{TP + TN}{TN + FN + FP + TP} = \frac{3 + 3}{3 + 2 + 2 + 3} = \frac{6}{10} = 0.6$$

- (e) For each situation, indicate whether precision or recall should be optimized:

- i. A smoke detector sounds in the presense of smoke (“positive”), helping people survive a fire in their home. Hearing a false alarm (e.g. because toast is over-cooked) is not expensive.

☐ Precision

☒ Recall ANSWER: ●

We want low FN and high TP, so we want high recall.

- ii. A doctor is screening patients for a disease that is an inconvenience but not a serious risk to health and whose treatment is expensive.

☐ Precision **ANSWER: ●**

We want low FP to avoid wasting money on a patient who does not need the treatment.

☐ Recall