STAT 451 Final Exam

Question	Points	Earned
Q1 (cover)	1	
Q2	8	
Q3	8	
Q4	8	
Q5	8	
Q6	18	
Q7	8	
Q8	8	
Q9	8	
Total	75	

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/460845 **as you work through the exam** so that you do not run out of time with questions unanswered.

2. 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?
   $\mathbf{c}_1 =$ _____ and $\mathbf{c}_2 =$ _____.
   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?
   $\mathbf{c}_1 =$ _____ and $\mathbf{c}_2 =$ _____.
   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 $\mathbf{c}_2 = \frac{1}{2}(10 + 12) = 11$.

3. 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.

4. 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

x	y
1	0
2	1
3	0
4	1

Resample #1

x	y
1	0
2	1
4	1
4	1

Resample #2

x	y
1	0
1	0
3	0
4	1

Resample #3

x	y
1	0
1	0
2	1
2	1

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

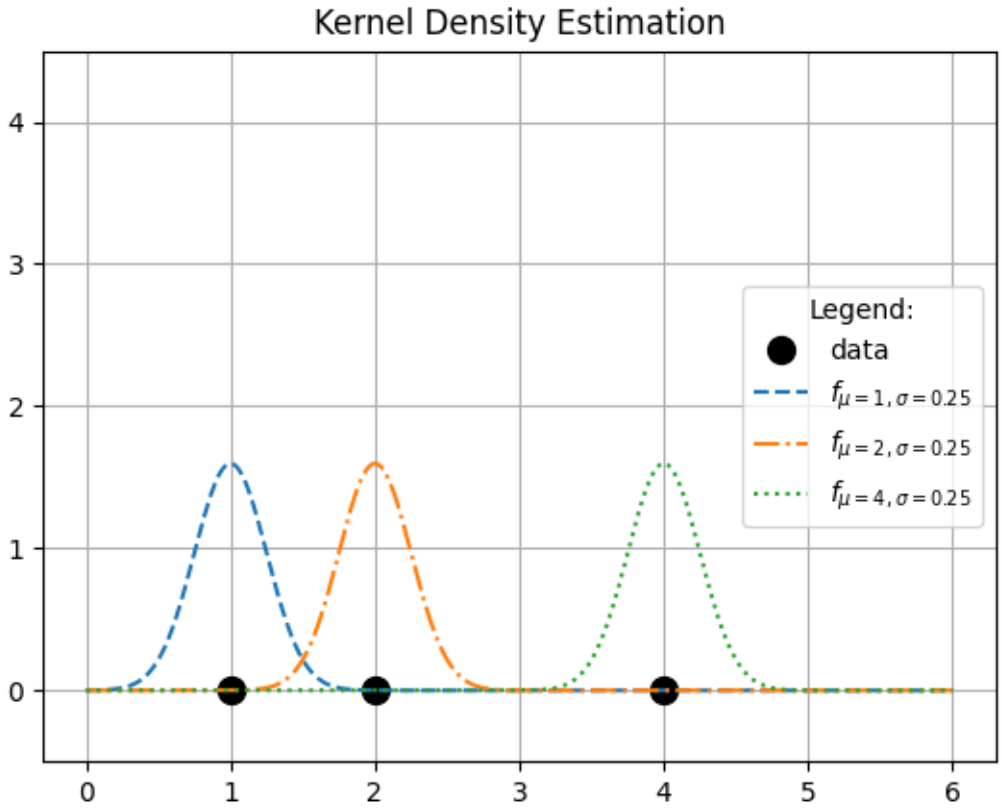
(a) What prediction is made by the tree trained on Resample #1? $\hat{y} =$ _____
ANSWER:
The tree uses threshold $t = 1.5$ and predicts $\hat{y} = 1$.

(b) What prediction is made by the tree trained on Resample #2? $\hat{y} =$ _____
ANSWER:
The tree uses threshold $t = 3.5$ and predicts $\hat{y} = 0$.

(c) What prediction is made by the tree trained on Resample #3? $\hat{y} =$ _____
ANSWER:
The tree uses threshold $t = 1.5$ and predicts $\hat{y} = 1$.

(d) What prediction is made by this bagging classifier? $\hat{y} =$ _____
ANSWER:
The bagging classifier predicts $\hat{y} = 1$, the most frequent of the $B = 3$ predictions.

5. Here is a graph of 1D data $\{\mathbf{x}_i\} = \{x_i\} = \{1, 2, 4\}$ and corresponding Gaussian curves $\{f_{\mu=x_i,\sigma=b}(x)\}$ made with bandwidth $b = 0.25$.

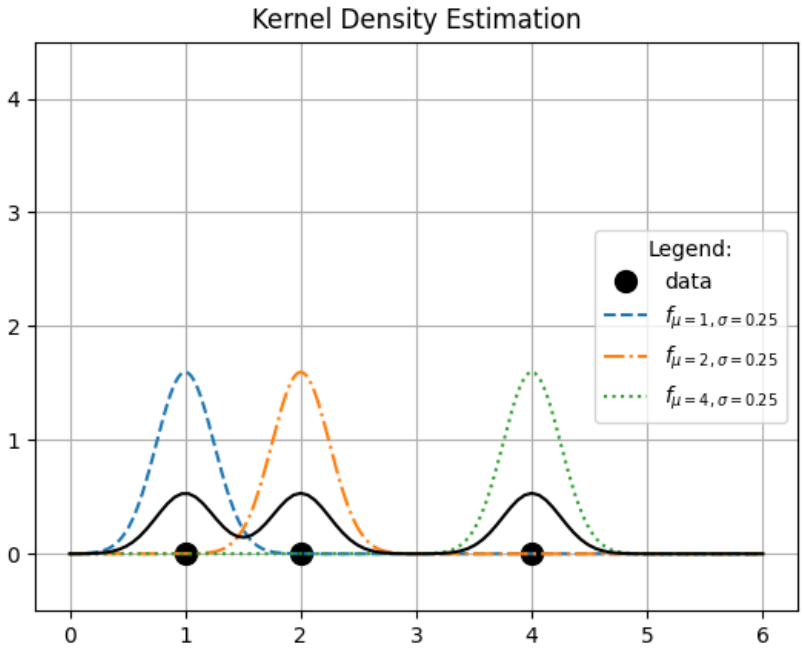


(a) Supposing the data were randomly sampled from some population, use kernel density estimation to estimate the population's probability density $f(x)$ at $x = 1$.
Based on the plot, the estimate is $\hat{f}_{b=0.25}(1) \approx$ _____.
ANSWER:
$\hat{f}_{b=0.25}(1)$ is the average of the three Gaussians at $x = 1$, which is $\approx \frac{1}{3}(1.6+0+0) \approx 0.5$.
(Any answer between 0.4 and 0.7 is also fine.)

(b) Estimate the density at $x = 1.5$.
Based on the plot, the estimate is $\hat{f}_{b=0.25}(1.5) \approx$ _____.
ANSWER:
$\approx \frac{1}{3}(0.2 + 0.2 + 0) \approx 0.13$. (Any answer between 0.05 and 0.2 is also fine.)

(c) On the figure above, draw the estimated density function over the interval $[0, 6]$.

ANSWER:



Anything close is also ok.

6. 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	0
(1, 4)	1	1
(3, −2)	1	1
(3, 0)	0	1

Complete the corresponding confusion matrix:

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

ANSWER:

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

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

	predicted $\hat{y}$	
actual y	0	1
0	2	4
1	3	5

What is the precision on the test data?

ANSWER:

The matrix is

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

so

$$\textit{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} = \frac{5}{5+4} = \frac{5}{9} \approx 0.556$$

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

$$\textit{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} = \frac{5}{5+3} = \frac{5}{8} = .625$$

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

$$\textit{accuracy} = \frac{\text{TP}+\text{TN}}{\text{TN}+\text{FN}+\text{FP}+\text{TP}} = \frac{5+2}{5+2+4+3} = \frac{7}{14} = \frac{1}{2} = .5$$

(e) For a classifier that is randomly guessing with $P(\hat{y} = 1) = \frac{1}{3}$, what is the AUC?

0.5

(f) For a classifier with TPR = 1 and FPR = 0, what is the AUC?

1

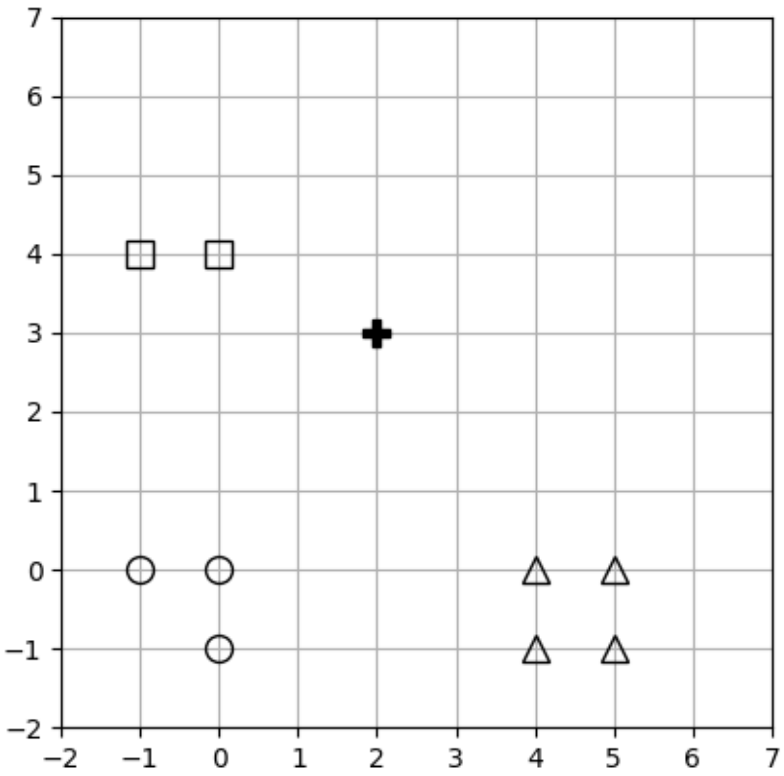
(g) For each situation, indicate whether P = precision or R = recall should be optimized:

i. _____ A bank is doing fraud detection where a fraudulent transaction ("positive") that is missed is expensive but a valid transaction labeled fraudulent is inexpensive.
R: We want low FN and high TP, so we want high recall.

ii. _____ A doctor is screening patients for a disease in which an ill patient ("positive") infects others and dies if the disease is not diagnosed.
R: We want high TP (sick people diagnosed as sick) and low FN (sick people undiagnosed), so we want high recall. (FP are relatively less harmful here.)

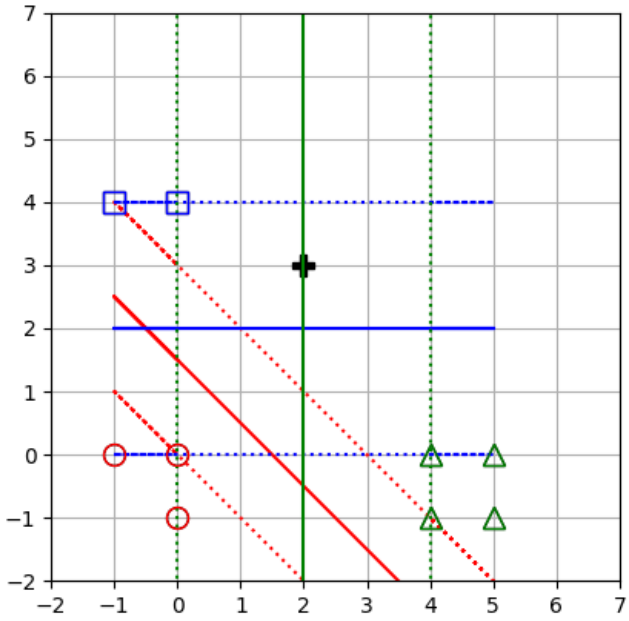
iii. _____ A marketing campaign invests considerable expense in a prospective customer when it classifies that customer as likely to make a purchase ("positive").
P: We want high TP and low FP, so we want high precision. (FN are just lost opportunities, which is ok with a limited marketing budget.)

7. Consider a one-vs.-rest SVM classifier trained on the following data depicted by circles, squares, and triangles:



(a) On the graph above, draw the three binary classifiers required by this method.

ANSWER:



(b) How does this classifier classify the point indicated by "+"?

_____ circle

_____ square

_____ triangle

ANSWER:

_____ circle

_X_ square

_____ triangle

From the graph in (a), note that:

- The blue boundary indicates that "+" is a square by the square-vs.-rest classifier (with a distance of 1 from the boundary).
- The green boundary indicates that "+" could be a triangle or not by the triangle-vs.-rest classifier (with a distance of 0 from the boundary).
- The red boundary indicates that "+" is not a circle by the circle-vs.-rest classifier (with a distance of about 2.5 on the wrong side of the boundary).

So the "+" is a square.

(c) Which category is ranked second by this classifier's decision method for the "+"?

_____ circle

_____ square

_____ triangle

ANSWER:

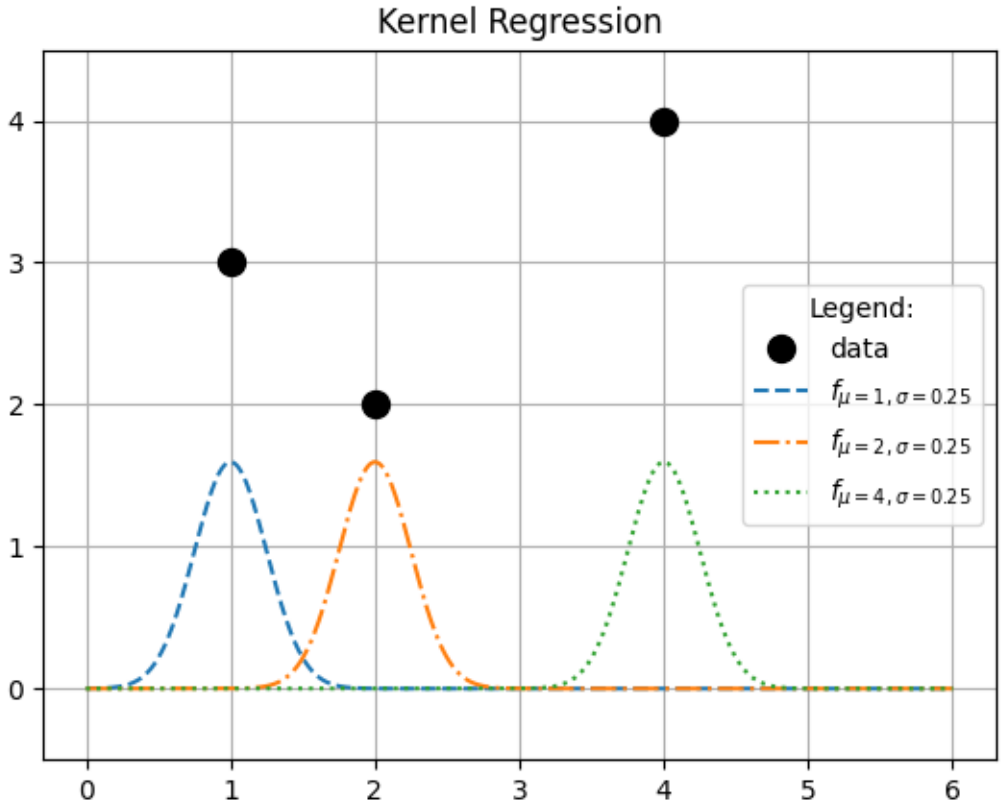
_____ circle

_____ square

_X_ triangle

Since the "+" is on the boundary for triangle but on the "not circle" side for circle, the second choice would be triangle.

8. Here is a graph of the data set $\{(\mathbf{x}_i, y_i)\} = \{(x_i, y_i)\} = \{(1,3),(2,2),(4,4)\}$ (here each $\mathbf{x}_i$ is a 1D x_i) along with corresponding Gaussian curves $\{f_{\mu=x_i,\sigma=b}(x)\}$ made with bandwidth $b = 0.25$:

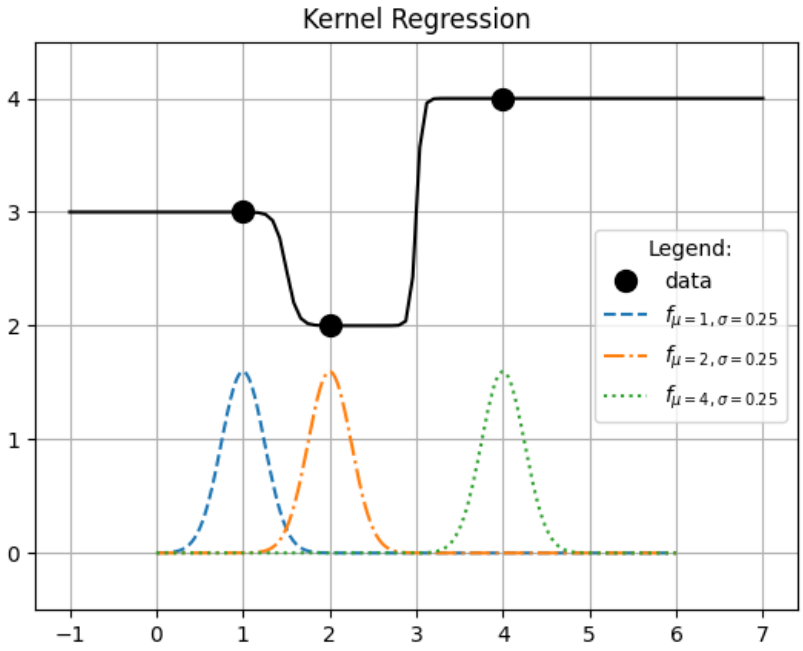


(a) Use kernel regression to estimate $y = f(x)$ for $x = 1$.
Based on the plot, the estimate is $\hat{y} \approx$ _____.
ANSWER:

$\hat{y}$ at x is a weighted average of the y values $(3,2,4)$ given by $f(x) = \sum_{i=1}^{N} w_i y_i$, where $w_i = \dfrac{k\left(\frac{x-x_i}{b}\right)}{\sum_{j=1}^{N} k\left(\frac{x-x_j}{b}\right)}$, which is the height of the ith density curve at x divided by the sum of the heights of all the curves at x.
At $x = 1$, the weights are $\approx (1.6, 0, 0)/(1.6{+}0{+}0) \approx (1, 0, 0)$, so $\hat{y} \approx 1{\times}3{+}0{\times}2{+}0{\times}4 \approx 3$.

(b) Estimate $y = f(x)$ for $x = 1.5$.
Based on the plot, the estimate is $\hat{y} \approx$ _____.
ANSWER:
We can see at $x = 1.5$ that $w_3 \approx 0$ and $w_1 = w_2$, so $w_1 = w_2 \approx \frac{1}{2}$, so $\hat{y} \approx \frac{1}{2}(3) + \frac{1}{2}(2) \approx 2.5$.

(c) On the figure above, draw the estimated regression function over the interval $[0, 6]$.
ANSWER:



Anything close is ok, where "close" includes having $\hat{y}$:

- ≈ 3 over $x \in (-\infty, 1)$
- decreasing from ≈ 3 to ≈ 2 over $(1, 2)$
- increasing from ≈ 2 to ≈ 4 over $(2, 4)$
- ≈ 4 over $(4, \infty)$

9. The next two questions are about principal component analysis (PCA).

(a) Consider the following code and its output:

```
rng = np.random.default_rng(seed=0)
(n_rows, n_cols) = (10, 4)
X = rng.normal(loc=0, scale=1, size=n_rows*n_cols).reshape((n_rows, n_cols))
pca = PCA(n_components=n_cols, random_state=0)
pca.fit(X=X)
with np.printoptions(precision=3):
print(f'pca.components_=\n{pca.components_}')
print(f'pca.explained_variance_={pca.explained_variance_}')
print(f'pca.explained_variance_ratio_={pca.explained_variance_ratio_}')
print(f'pca.noise_variance_={pca.noise_variance_}')
print(f'pca.mean_={pca.mean_}')
print(f'pca.singular_values_={pca.singular_values_}')
```

Output:

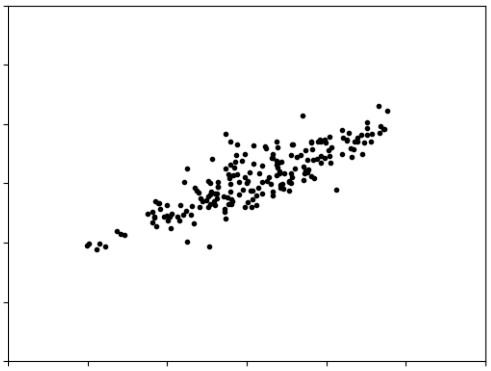
```
pca.components_=
[[-0.219 -0.091 -0.752 -0.615]
  [ 0.854  0.439 -0.085 -0.265]
  [-0.41   0.882 -0.138  0.184]
  [-0.232  0.142  0.639 -0.72 ]]
pca.explained_variance_=[1.237 0.733 0.388 0.109]
pca.explained_variance_ratio_=[0.501 0.297 0.157 0.044]
pca.noise_variance_=0.0
pca.mean_=[-0.448  0.052 -0.093  0.247]
pca.singular_values_=[3.336 2.569 1.869 0.988]
```

What is the minimum number of principal components we must retain to account for 90% of the variability in the data? _____

ANSWER:

3, since we can see that the sum of the first two elements in `pca.explained_variance_ratio_` is less than 90% but the sum of the first three is more than 90%.

(b) Suppose PCA is run on the data in the plot. Draw arrows on the plot representing the first two principal components. (There is more than one correct answer.)



ANSWER:

The first should be in the direction of the greatest variability in the data. The second should be perpendicular to the first (in the direction of the second greatest variability in the data).