

STAT 451 Final Exam NetID (mine is “jgillett” from “jgillett@wisc.edu”): _____

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

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)	1	
Q1	4	
Q2	4	
Q3	16	
Q4	8	
Q5	6	
Q6	6	
Q7	6	
Q8	6	
Q9	18	
Total	75	

1. Consider using k -means on the unsupervised 1D dataset $\{\mathbf{x}\} = \{0, 3, 6, 7, 8\}$ to create $k = 2$ clusters. Suppose the two initial randomly-chosen cluster centroids are $\mathbf{c}_1 = 6$ and $\mathbf{c}_2 = 7$.

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

$\mathbf{c}_1 =$ and $\mathbf{c}_2 =$.

ANSWER:

0, 3, and 6 are closest to $\mathbf{c}_1 = 6$, so the new $\mathbf{c}_1 = \frac{1}{3}(0 + 3 + 6) = 3$.

7 and 8 are closest to $\mathbf{c}_2 = 7$, so the new $\mathbf{c}_2 = \frac{1}{2}(7 + 8) = 7.5$.

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

$\mathbf{c}_1 =$ and $\mathbf{c}_2 =$.

ANSWER:

0 and 3 are closest to $\mathbf{c}_1 = 3$, so the new $\mathbf{c}_1 = \frac{1}{2}(0 + 3) = 1.5$.

6, 7, and 8 are closest to $\mathbf{c}_2 = 7.5$, so the new $\mathbf{c}_2 = \frac{1}{3}(6 + 7 + 8) = 7$.

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

(a) A model has two hyperparameters:

- The first is from the normal distribution with mean 6 and standard deviation 2.
- The second is an integer in the range $[-10000, 10000]$.

☐ Grid search

☐ Random search ANSWER: ☒

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.

(b) A model has three hyperparameters:

- The first takes a floating-point number from the set $\{1, e, \pi\}$.
- The second is a metric, either 'euclidean', 'manhattan', or a user-defined function.
- The third is the number of CPU cores available, from 1 to 3.

☐ Grid search ANSWER: ☒

☐ Random search

Grid search because there are only 18 combinations to try and there's no need to risk random search missing one.

3. Mark each statement true or false.

- (a) If $0 < d < p < D$, then using PCA to transform D -dimensional data $\{\mathbf{x}_i\}$ to p -dimensional data and then using PCA again to transform those p -dimensional data to d -dimensional data gives the same transformed data as using PCA to transform the D -dimensional data directly to d -dimensional data.
- ☐ True **ANSWER:** ●
- ☐ False
- (b) An SVM makes a classification error on $(\mathbf{x}, y = -1)$ when $y(\mathbf{w}\mathbf{x} + b) > 0$.
- ☐ True
- ☐ False **ANSWER:** ●
- $y(\mathbf{w}\mathbf{x} + b) > 0 \implies -1(\mathbf{w}\mathbf{x} + b) > 0 \implies \mathbf{w}\mathbf{x} + b < 0 \implies$ the SVM classifies $\mathbf{x}$ as -1 , which is correct.
- (c) In logistic regression, we predict $\hat{y} = 1$ for $\mathbf{x}$ if $\frac{1}{1+e^{-(\mathbf{w}\mathbf{x}+b)}} > 0$.
- ☐ True
- ☐ False **ANSWER:** ●
- $\frac{1}{1+e^{-(\mathbf{w}\mathbf{x}+b)}} > 0$ for every $\mathbf{x}$. We predict $\hat{y} = 1$ when that value is larger than some threshold $t \in (0, 1)$. (Typically $t = \frac{1}{2}$.)
- (d) We would get the same model in linear regression if we chose parameters to minimize the logarithm of MSE instead of choosing them to minimize MSE itself.
- ☐ True **ANSWER:** ●
- ☐ False
- (e) When splitting a binary decision tree node S containing $\{y\} = \{0, 0, 1, 1, 1\}$, the tree will prefer the split $S_- = \{0, 0, 1, 1\}$ and $S_+ = \{1\}$ to the split $S_- = \{0, 0, 1\}$ and $S_+ = \{1, 1\}$.
- ☐ True
- ☐ False **ANSWER:** ●
- Both have $H(S_+) = 0$, so $H(S_-, S_+) = \frac{|S_-|}{|S|}H(S_-) + \frac{|S_+|}{|S|}H(S_+) = \frac{|S_-|}{|S|}H(S_-) + 0$. $\frac{|S_-|}{|S|}$ and $H(S_-)$ are both lower in the second than in the first. ($H(S_-, S_+) = \frac{4}{5} = 0.8$ for the first and $H(S_-, S_+) = \frac{5}{9} \approx 0.556$ for the second.)
- (f) In a linear support vector machine, $\mathbf{w}$ is normal to all three lines given by $\mathbf{w}\mathbf{x} + b = -1$, $\mathbf{w}\mathbf{x} + b = 0$, and $\mathbf{w}\mathbf{x} + b = 1$.
- ☐ True **ANSWER:** ●
- ☐ False
- (g) Ridge regression can address the problem of ordinary least squares regression overfitting the training data.
- ☐ True **ANSWER:** ●
- ☐ False
- (h) If we run k -means clustering with $k = 2$ on the 1D data $\{0, 2, 3, 4, 5\}$ starting with cluster centers 0 and 5, the final centers are 2 and 4.

- ☐ True
☐ False ANSWER: ●
 0 and 2 are closest to 0, so the first new center is $\frac{1}{2}(0 + 2) = 1$. 3, 4, and 5 and 8 are closest to 5, so the second new center is $\frac{1}{3}(3 + 4 + 5) = 4$. These centers do not change on the next iteration.
- (i) If we run DBSCAN with $\epsilon = 2$ and $n = 3$ on the 1D data $\{0, 2, 3, 4\}$ starting at 3, the final clusters are $\{0\}$ and $\{2, 3, 4\}$.
☐ True
☐ False ANSWER: ●
 2, 3, and 4 are core examples, so they end up in the first cluster. 0 is a neighbor of core example 2 in the first cluster, so it ends up in that first cluster too.
- (j) The bagging and random forest methods both use resampling with replacement from the training examples.
☐ True ANSWER: ●
☐ False
- (k) Boosting selects a different random subset of features to process on each iteration.
☐ True
☐ False ANSWER: ●
- (l) k -NN classification may benefit from standardizing features.
☐ True ANSWER: ●
☐ False
- (m) Decision tree regression may benefit from standardizing features.
☐ True
☐ False ANSWER: ●
- (n) Simple linear regression (in which $\mathbf{x}$ is a 1D scalar x) may benefit from standardizing the feature x .
☐ True
☐ False ANSWER: ●
- (o) In PCA, the first principal component is the one in the direction of the greatest variability in the data.
☐ True ANSWER: ●
☐ False
- (p) $1 + 1 = 2$.
☐ True ANSWER: ●
 (I replaced a defective question with this one.)
☐ False

4. Which of the following methods are typically used for feature selection or feature extraction?
Choose Yes or No for each method.

(a) Lasso regression

☐ Yes. ANSWER: ●

☐ No.

(b) Kernel trick

☐ Yes.

☐ No. ANSWER: ●

(c) Correlation

☐ Yes. ANSWER: ●

☐ No.

(d) Principal component analysis

☐ Yes. ANSWER: ●

☐ No.

(e) Kernel density estimation

☐ Yes.

☐ No. ANSWER: ●

(f) Gradient descent

☐ Yes.

☐ No. ANSWER: ●

(g) Ridge regression

☐ Yes.

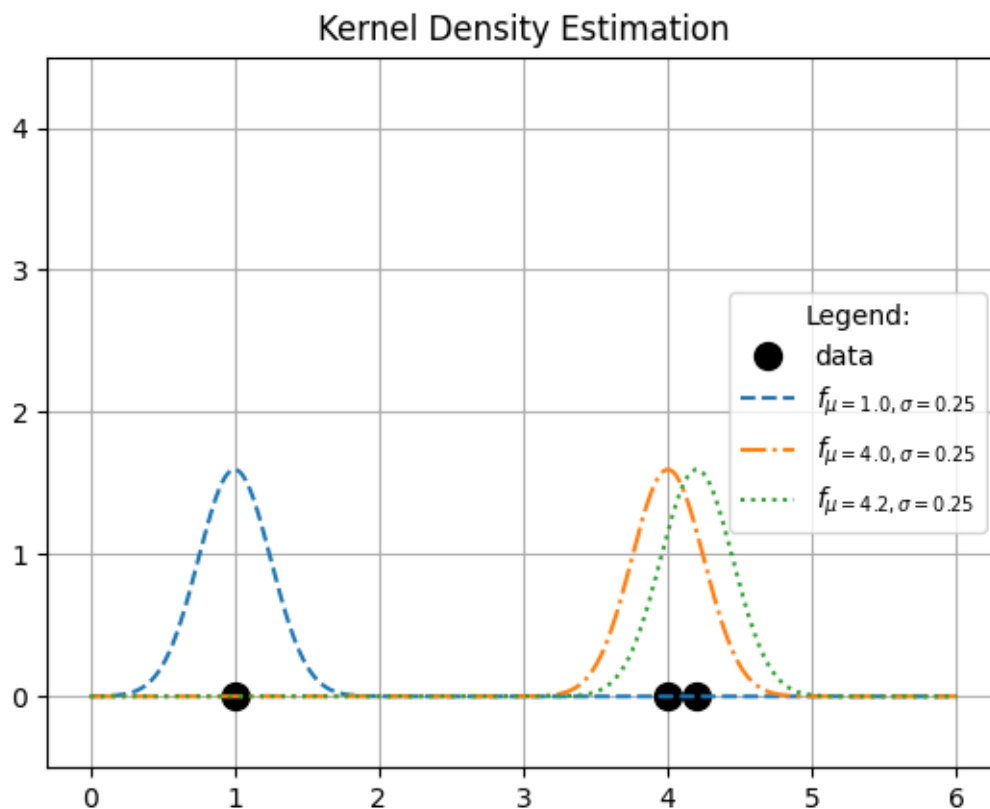
☐ No. ANSWER: ●

(h) Permutation feature importance

☐ Yes. ANSWER: ●

☐ No.

5. Here is a graph of 1D data $\{\mathbf{x}_i\} = \{x_i\} = \{1, 4, 4.2\}$ 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 $x = 4.1$.

Based on the plot, the estimate is $\hat{f}_{b=0.25}(4.1) \approx$

ANSWER:

$\hat{f}_{b=0.25}(4.1)$ is the average of the three Gaussians at $x = 1$, which is $\approx \frac{1}{3}(0 + 1.5 + 1.5) \approx 1$.
(Any answer between 0.7 and 1.3 is also fine.)

- (b) Estimate the density at $x = 2$.

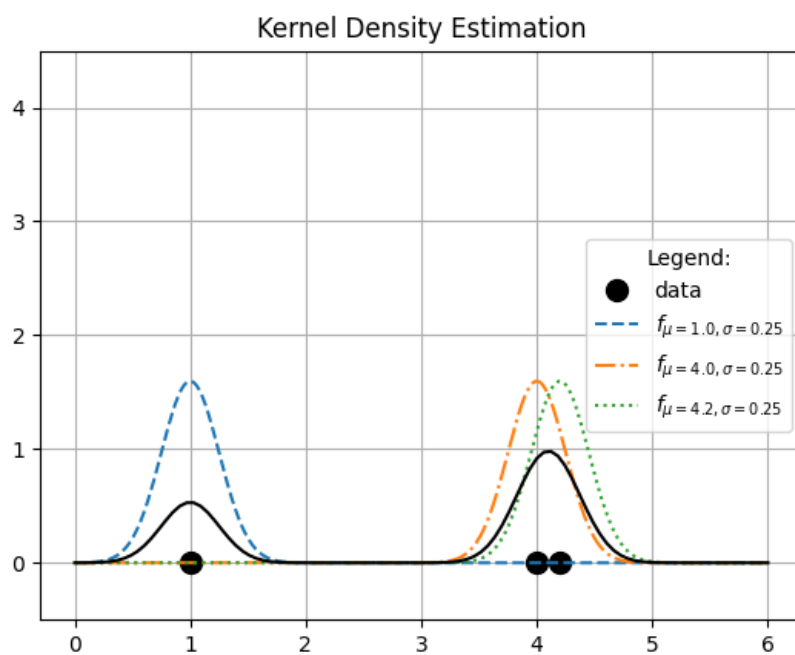
Based on the plot, the estimate is $\hat{f}_{b=0.25}(2) \approx$

ANSWER:

$\approx \frac{1}{3}(0 + 0 + 0) \approx 0$. (Any answer between 0 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 use of gradient boosting to train a regression model on the following data:

x	y
1	2
2	1
3	3

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

$\hat{y} =$

ANSWER:

$\hat{y} = 2$ because $f_0(x)$ is a constant model given by $f_0(x) = \frac{1}{N} \sum_{i=1}^N y_i = \frac{1}{3}(2 + 1 + 3) = 2$.

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

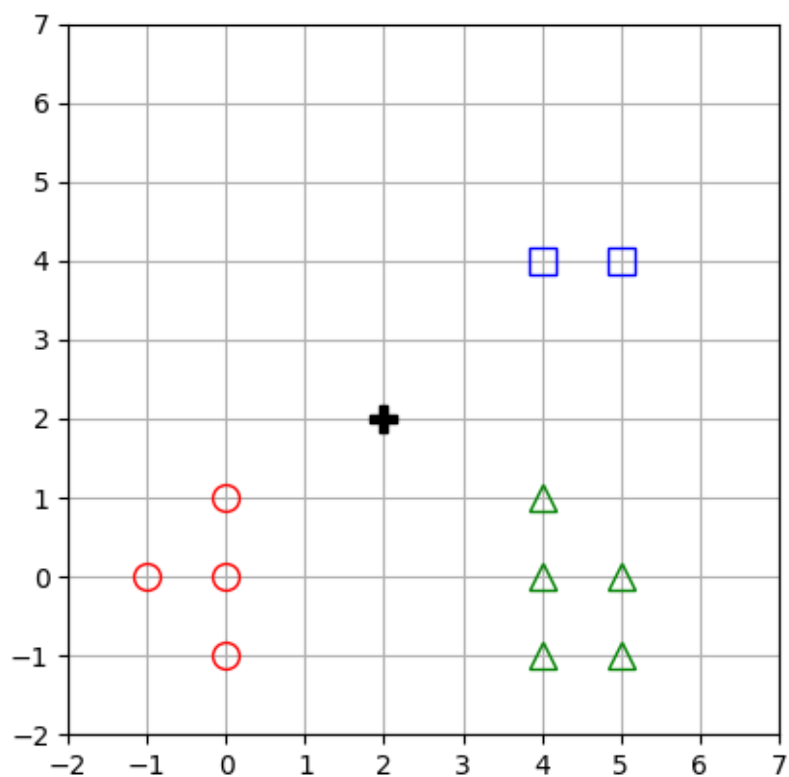
x	y

ANSWER:

x	y
1	0
2	-1
3	1

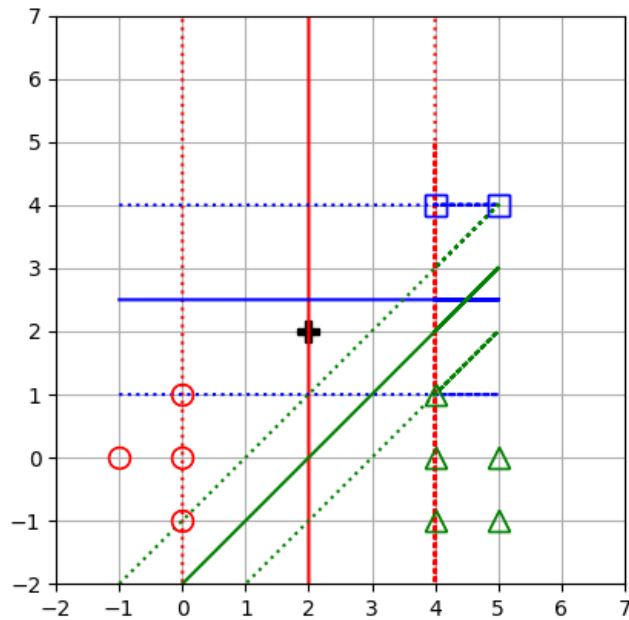
f_1 is trained on the original data after replacing the y values with the residuals with respect to the first model, that is with $\{e_i = y_i - f_0(x_i)\}$. Since $f_0(x) = 2$ for all x , $\{e_i\} = \{2 - 2, 1 - 2, 3 - 2\} = \{0, -1, 1\}$.

7. Consider a one-vs.-rest hard-margin linear 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 **ANSWER:** ☒

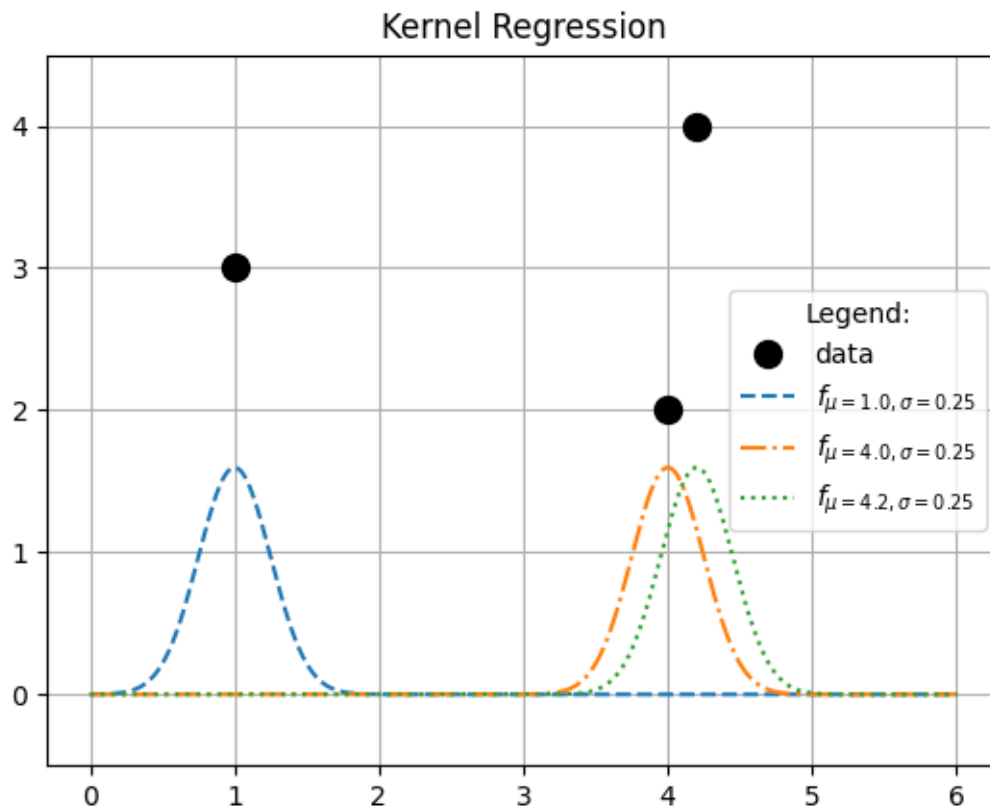
From the graph in (a), note that:

- The blue boundary indicates that “+” is not a square by the square-vs.-rest classifier.
- The green boundary indicates that “+” is not triangle by the triangle-vs.-rest classifier.
- The red boundary indicates that “+” is could be a circle or not by the circle-vs.-rest classifier.

So the “+” is a circle.

- ☐ square
- ☐ triangle

8. Here is a graph of the data set $\{(\mathbf{x}_i, y_i)\} = \{(x_i, y_i)\} = \{(1, 3), (4, 2), (4.2, 4)\}$ (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 $x = 4.1$.

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

$w_i = \frac{k\left(\frac{x-x_i}{b}\right)}{\sum_{j=1}^N k\left(\frac{x-x_j}{b}\right)}$, which is the height of the i th density curve at x divided by the sum of the heights of all the curves at x .

At $x = 4.1$, the weights are $\approx (0, 1.5, 1.5)/(0 + 1.5 + 1.5) \approx (0, 0.5, 0.5)$, so $\hat{y} \approx 0 \times 3 + 0.5 \times 2 + 0.5 \times 4 \approx 3$.

Note that the answer does not depend on the approximate density value being exactly 1.5. The answer is 3 for any choice of first density near zero with second and third densities equal and nonzero.

(b) Estimate $y = f(x)$ for $x = 1$.

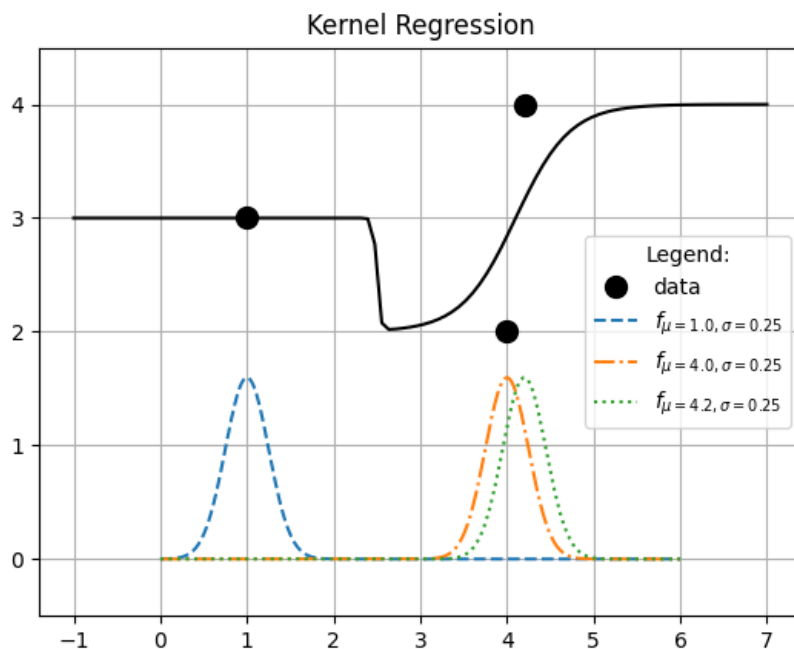
Based on the plot, the estimate is $\hat{y} \approx$.

ANSWER:

We can see at $x = 1$ that $w_1 \approx 1$, $w_2 \approx 0$, and $w_3 \approx 0$, so $\hat{y} \approx 1 \cdot 3 + 0 \cdot 2 + 0 \cdot 4 = 3$.

(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 $x \in (-\infty, \approx 2.5)$
- increasing from ≈ 2 to ≈ 4 over $(\approx 2.5, \infty)$
- ≈ 3 at $x = 4.1$
- increasing toward 4 over $(4.1, \infty)$

9. 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, 2)	1	1
(1, 1)	0	0
(2, 3)	0	0
(4, 1)	0	1

Complete the corresponding confusion matrix:

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

ANSWER:

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

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

$$precision = \frac{TP}{TP + FP} = \frac{4}{4 + 2} = \frac{2}{3} \approx 0.667$$

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

$$recall = \frac{TP}{TP + FN} = \frac{4}{4 + 1} = \frac{4}{5} = 0.8$$

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

$$accuracy = \frac{TP + TN}{TN + FN + FP + TP} = \frac{4 + 3}{3 + 1 + 2 + 4} = \frac{7}{10} = 0.7$$

- (e) Consider a classifier for which we can have either false positive rate 0 with true positive rate 0.5 or false positive rate 0.5 with true positive rate 1. What is the AUC?

The classifier's ROC curve goes through $(\text{FPR}, \text{TPR}) = (0, 0.5)$ and $(0.5, 1)$. The area under it is $7/8 = 0.875$.

- (f) For a classifier with $\text{TPR} = 0$ and $\text{FPR} = 1$, what is the AUC?

0

- (g) For each situation, indicate whether precision or 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.

☐ Precision

☒ Recall ANSWER: ●

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.

☐ Precision

☒ Recall ANSWER: ●

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").

☒ Precision ANSWER: ●

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.)

☐ Recall