





Michigan State University  
CMSE381 - Data Science

Feb 10, 2023  
Dr. Xie

## CMSE381 - Midterm #1

First name (please write as legibly as possible within the boxes)

N i c o l a i

Last name

B a i r d

NetID

*I will adhere to the Spartan Code of Honor in completing this assignment.*

Signed: Nicola Baird

1. Do not open this test booklet until you are directed to do so.
2. You will have class time (2:40-4:00pm) to complete the exam.
3. This exam is closed book, but you can use the cheatsheet provided by the instructor.
4. Throughout the test, show your work so that your reasoning is clear. Otherwise no credit will be given. BOX your answers. Partial credit will be given where warranted.
5. Do not spend too much time on any one problem. Read them all through first and attack them in the order that allows you to make the most progress. Good luck :P



1. (15 points)

(a) Logistic regression is used for regression.

☒ TRUE ☐ FALSE

(b) Any model will never have training error below the irreducible error.

☒ TRUE ☐ FALSE

(c) Increasing your model flexibility always results in a better model.

☐ TRUE ☒ FALSE

(d) A logistic regression model is set up so that the odds are linear.

☐ TRUE ☒ FALSE

(e) Circle all of the following that would represent a qualitative variable.

Age      Year      ☒ Dog\_breed      ☒ Country\_of\_origin

☒ Student\_(True/False)      Weight      Speed      MPG

(f) What equation would you use to evaluate the result of a regression model?

$$\frac{1}{n} \sum_{i=1}^n (y_i - \hat{f}(x_i))^2 \quad \text{mean squared error}$$

(g) What equation would you use to evaluate the result of a classification model?

$$\frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}}$$



2. (15 Points) I'm building a model to predict amount of a given brand of dog food eaten by a collection of dogs. I have 100 dogs eat this dog food, and I collect information on their height, weight, breed, and whether they live with another dog in the house.

(a) List all input variables and specify whether they are quantitative or qualitative.

quantitative: height, weight  
qualitative: breed, live w/ other dog

(b) Say our dog breeds sampled are Huskies, Terriers, and Spaniels. Write down your prediction model.

Huskies:  $[0, 0]$   
Terriers:  $[1, 0]$   
Spaniels:  $[0, 1]$

(c) We found that weight is a key predictor, and its relationship with the response is beyond linear. What extension can you come up with? Write down your updated model.

$$\hat{Y} = \beta_0 + \text{weight}(X)$$

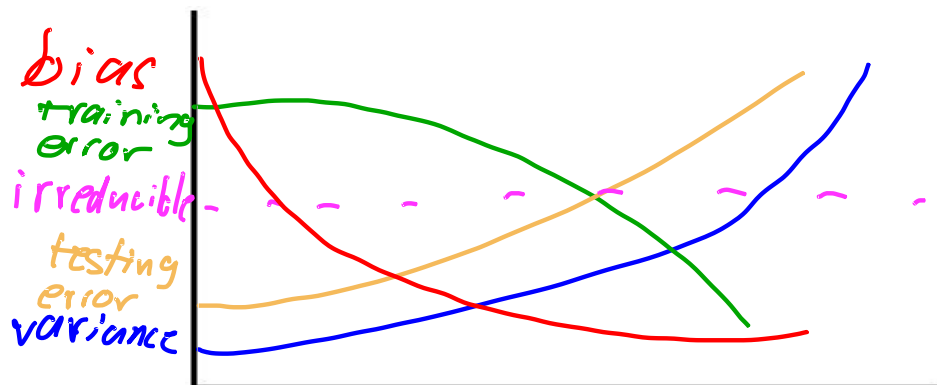


3. (15 points)

(a) What is bias-variance tradeoff? Explain the meaning of each term in the formula.

$E(\text{test msc at } x_i) = \text{Var}(x_i) + \text{bias}(x_i)^2 + \text{var}(\text{irreducible error})$   
 $\text{Var}(x_i)$ : variability of model at  $x_i$ .  
 $\text{bias}(x_i)^2$ : squared bias of model at  $x_i$ .  
 $\text{var}(\text{irreducible})$ : The irreducible error that will exist in all models.

(b) Provide a sketch of typical (squared) bias, variance, training error, test error, and Bayes (or irreducible) error curves, on a single plot, as we go from less flexible statistical learning methods towards more flexible approaches. The  $x$ -axis should represent the amount of flexibility in the method, and the  $y$ -axis should represent the values for each curve. There should be five curves. Make sure to label each one.



(c) Explain why the (i) training error and (ii) testing error lines in your drawing have the shape displayed.

Training error decreases as the amount of flexibility increases because training error is mostly a result of overfitting, meaning that training error is a product of model bias. Testing error increases with flexibility because a less precise model cannot be trusted as much to be tested accurately.



4. (10 points) The data for this example come from a study by Stamey et al. (1989). The data comes from a number of clinical measures in men who were about to receive a radical prostatectomy. The goal is to predict the log of PSA ( $\text{lpsa}$ ) (a prostate-specific antigen measured in nanograms of PSA per milliliter of blood (ng/mL)) from a number of measurements. The variables are log cancer volume ( $\text{lcavol}$ ), log prostate weight ( $\text{lweight}$ ), age, log of the amount of benign prostatic hyperplasia ( $\text{lbph}$ ), seminal vesicle invasion ( $\text{svi}$ ), log of capsular penetration ( $\text{lcp}$ ), Gleason score ( $\text{gleason}$ ), and percent of Gleason scores 4 or 5 ( $\text{pgg45}$ ).

(a) Is this a case of regression or classification? Why?

*This is a case of regression because we are trying to figure out the value of a variable, not its type.*

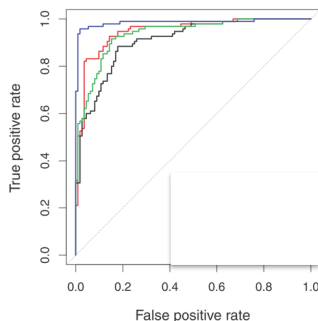
(b) A linear model is fit to the data set, and the following table was returned.

Term	Coefficient	Std. Error	t-Score	p-value
Intercept	2.46	0.09	27.6	<0.00001
lcavol	0.68	0.13	5.37	<0.00001
lweight	0.26	0.1	2.75	0.00596
age	-0.14	0.1	-1.4	0.16153
lbph	0.21	0.1	2.06	0.039399
svi	0.31	0.12	2.47	0.013511
lcp	-0.29	0.15	-1.87	0.061484
gleason	-0.02	0.15	-0.15	0.880765
pgg45	0.27	0.15	1.74	0.081859

Are all the predictors useful? If yes, explain why. If not, explain what you would try to do next with the data.

*lbph, lcp, and gleason seem to be the most relevant predictors, because they have relatively high p-values. I would next try to determine the*

5. (5 pts) We are trying to predict whether a patient will have a heart attack within one year. We have tried four different methods on a training dataset and have the following ROC curves. Which curve has the best performance on this training set? Justify your answer.



*The blue curve has the best performance because it has the highest amount of true positives, we should also be more comfortable with false positives than false negatives because we don't want someone to die when we said that they would be fine.*

*$R^2$  value for each combination of the predictors as well as model degree*



6. (15 Points) I get way too much email, so I decide to build a logistic regression model to predict whether a new incoming message is spam or not. I decide to just use a few variables:

$$X_1 = \text{Number\_of\_sentences}$$

$$X_2 = \text{Number\_of\_references\_to\_a\_Nigerian\_Prince}$$

and am training a logistic model to predict

$$Pr(Y = \text{spam} \mid X_1, X_2)$$

- (a) Write down the equation for the model you would train, using our standard notation with  $\beta_i$ 's.

$$\frac{e^{\beta_0 + \beta_1 x + \beta_2 x^2}}{1 + e^{\beta_0 + \beta_1 x + \beta_2 x^2}}$$

- (b) If my trained model used  $\beta_0 = -13.1$ ,  $\beta_1 = 1.9$ , and  $\beta_2 = 6.1$ , what is the probability that a 5 sentence email with one reference to a Nigerian prince is spam? .

$$\frac{e^{-13.1 + 1.9x + 6.1x^2}}{1 + e^{-13.1 + 1.9x + 6.1x^2}}$$

- (c) We know that if we miss an important email, it may cost a lot, How can you modify your model to avoid this?

In this case we would want to increase the number of false negatives so we should increase our model flexibility



7. (15 Points) Sparty generated 100 pair  $x$  and  $y$  via the following relationship:  $Y = 0.3 + 2x + x^2 + \epsilon$  but didn't tell you. You will try to find a good model to fit these data and more importantly to predict future response  $y$  when Sparty gives you another  $x$ . You will start with two models 1)  $Y = \beta_0 + \beta_1 x$ ; 2)  $Y = \alpha_0 + \alpha_1 x + \alpha_2 x^2$ ; and 3)  $Y = \gamma_0 + \gamma_1 x + \gamma_2 x^2 + \gamma_3 x^3$ .
- (a) If you use all 100 data to fit these three model and use the same 100 data to calculate MSE, which model will have the smallest MSE? Justify your answer

The model with the smallest mse will be model #2 because the shape of the data will match the shape of the model curve

- (b) Since we focus on the ability to predict unseen data, you decide use validation set to evaluate the performance of the three models. There are two way to split the data: 1) 80 testing points and 20 training points or 2) 20 testing points and 80 training points. Which one will you choose? Justify your choice

I would use 80 training points and 20 testing points because it is more important to have a well constructed model that is moderately well tested than a bad model that is well tested.

- (c) From the in-class lab, we know validation set is not a good choice to choose model. Explain the drawback of validation set and what procedure will you use to choose the best model?

Validation set is not a good choice because it is random and may give us a different model based on chance,  $k$ -fold would give us a more standardized idea for testing and if we had the computational freedom LOO might do an even better





8. (10 Points) Suppose we have a data set with five predictors,  $X_1 = \text{GPA}$ ,  $X_2 = \text{IQ}$ ,  $X_3 = \text{Level}$  (1 for College and 0 for High School),  $X_4 = \text{Interaction between GPA and IQ}$ , and  $X_5 = \text{Interaction between GPA and Level}$ . The response is starting salary after graduation (in thousands of dollars). Suppose we use least squares to fit the model, and get  $\beta_0 = 50$ ,  $\beta_1 = 20$ ,  $\beta_2 = 0.07$ ,  $\beta_3 = 35$ ,  $\beta_4 = 0.01$ ,  $\beta_5 = -10$ .

- (a) Predict the salary of a college graduate with IQ of 110 and a GPA of 4.0.

$$\hat{Y} = 50 + 0.07(110) + 4.0(0.01)^2$$

- (b) True or false: Since the coefficient for the GPA/IQ interaction term is very small, there is very little evidence of an interaction effect. Justify your answer.

This is not true based on the coefficient, because there are other predictors present in the model, the small coefficient may just be in order to make it so that the scales of each predictor line up. We would want to look at  $R^2$  to determine the relevance with or without.

12F007A7-9964-4A8F-AE77-354985AF52F2

midterm1-b9ace

#1 Page 9 of 10



**Scrap Paper**



5C38AE32-92B5-498C-B965-69097BC42CC1

midterm1-b9ace

#1 Page 10 of 10



Michigan State University  
CMSE381 - Data Science

Feb 10, 2023  
Dr. Xie

## CMSE381 - Midterm #1

First name (please write as legibly as possible within the boxes)

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

Last name

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

NetID

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

*I will adhere to the Spartan Code of Honor in completing this assignment.*

Signed: \_\_\_\_\_

1. Do not open this test booklet until you are directed to do so.
2. You will have class time (2:40-4:00pm) to complete the exam.
3. This exam is closed book, but you can use the cheatsheet provided by the instructor.
4. Throughout the test, show your work so that your reasoning is clear. Otherwise no credit will be given. **BOX** your answers. Partial credit will be given where warranted.
5. Do not spend too much time on any one problem. Read them all through first and attack them in the order that allows you to make the most progress. Good luck :P



5308F5CD-7112-4E3B-9286-800F2CE4D6DF

midterm1-b9ace

#2 Page 2 of 10

1. (15 points)

(a) Logistic regression is used for regression.

TRUE      FALSE

(b) Any model will never have training error below the irreducible error.

TRUE      FALSE

(c) Increasing your model flexibility always results in a better model.

TRUE      FALSE

(d) A logistic regression model is set up so that the odds are linear.

TRUE      FALSE

(e) Circle all of the following that would represent a qualitative variable.

Age      Year      Dog\_breed      Country\_of\_origin

Student\_(True/False)      Weight      Speed      MPG

(f) What equation would you use to evaluate the result of a regression model?

(g) What equation would you use to evaluate the result of a classification model?



2. (15 Points) I'm building a model to predict amount of a given brand of dog food eaten by a collection of dogs. I have 100 dogs eat this dog food, and I collect information on their height, weight, breed, and whether they live with another dog in the house.
- (a) List all input variables and specify whether they are quantitative or qualitative.
- (b) Say our dog breeds sampled are Huskies, Terriers, and Spaniels. Write down your prediction model.
- (c) We found that weight is a key predictor, and its relationship with the response is beyond linear. What extension can you come up with? Write down your updated model.



844335CC-6D0C-499B-9345-2C46CAE9AC60

midterm1-b9ace

#2 Page 4 of 10

3. (15 points)

(a) What is bias-variance tradeoff? Explain the meaning of each term in the formula.

(b) Provide a sketch of typical (squared) bias, variance, training error, test error, and Bayes (or irreducible) error curves, on a single plot, as we go from less flexible statistical learning methods towards more flexible approaches. The  $x$ -axis should represent the amount of flexibility in the method, and the  $y$ -axis should represent the values for each curve. There should be five curves. Make sure to label each one.



(c) Explain why the (i) training error and (ii) testing error lines in your drawing have the shape displayed.



4. (10 points) The data for this example come from a study by Stamey et al. (1989). The data comes from a number of clinical measures in men who were about to receive a radical prostatectomy. The goal is to predict the log of PSA (**lpsa**) (a prostate-specific antigen measured in nanograms of PSA per milliliter of blood (ng/mL)) from a number of measurements. The variables are log cancer volume (**lcavol**), log prostate weight (**lweight**), age, log of the amount of benign prostatic hyperplasia (**lbph**), seminal vesicle invasion (**svi**), log of capsular penetration (**lcp**), Gleason score (**gleason**), and percent of Gleason scores 4 or 5 (**pgg45**).

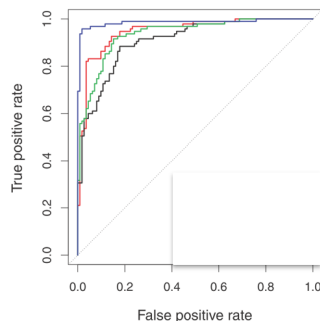
(a) Is this a case of regression or classification? Why?

(b) A linear model is fit to the data set, and the following table was returned.

Term	Coefficient	Std. Error	t-Score	p-value
Intercept	2.46	0.09	27.6	<0.00001
lcavol	0.68	0.13	5.37	<0.00001
lweight	0.26	0.1	2.75	0.00596
age	-0.14	0.1	-1.4	0.16153
lbph	0.21	0.1	2.06	0.039399
svi	0.31	0.12	2.47	0.013511
lcp	-0.29	0.15	-1.87	0.061484
gleason	-0.02	0.15	-0.15	0.880765
pgg45	0.27	0.15	1.74	0.081859

Are all the predictors useful? If yes, explain why. If not, explain what you would try to do next with the data.

5. (5 pts) We are trying to predict whether a patient will have a heart attack within one year. We have tried four different methods on a training dataset and have the following ROC curves. Which curve has the best performance on this training set? Justify your answer.







6. (15 Points) I get way too much email, so I decide to build a logistic regression model to predict whether a new incoming message is spam or not. I decide to just use a few variables:

$$\begin{aligned}X_1 &= \text{Number\_of\_sentences} \\X_2 &= \text{Number\_of\_references\_to\_a\_Nigerian\_Prince}\end{aligned}$$

and am training a logistic model to predict

$$Pr(Y = \text{spam} \mid X_1, X_2)$$

- (a) Write down the equation for the model you would train, using our standard notation with  $\beta_i$ 's.
- (b) If my trained model used  $\beta_0 = -13.1$ ,  $\beta_1 = 1.9$ , and  $\beta_2 = 6.1$ , what is the probability that a 5 sentence email with one reference to a Nigerian prince is spam? .
- (c) We know that if we miss an important email, it may cost a lot, How can you modify your model to avoid this?



7. (15 Points) Sparty generated 100 pair  $x$  and  $y$  via the following relationship:  $Y = 0.3 + 2x + x^2 + \epsilon$  but didn't tell you. You will try to find a good model to fit these data and more importantly to predict future response  $y$  when Sparty gives you another  $x$ . You will start with two models 1)  $Y = \beta_0 + \beta_1x$ ; 2)  $Y = \alpha_0 + \alpha_1x + \alpha_2x^2$ ; and 3)  $Y = \gamma_0 + \gamma_1x + \gamma_2x^2 + \gamma_3x^3$ .
- (a) If you use all 100 data to fit these three model and use the same 100 data to calculate MSE, which model will have the smallest MSE? Justify your answer
- (b) Since we focus on the ability to predict unseen data, you decide use validation set to evaluate the performance of the three models. There are two way to split the data: 1) 80 testing points and 20 training points or 2) 20 testing points and 80 training points. Which one will you choose? Justify your choice
- (c) From the in-class lab, we know validation set is not a good choice to choose model. Explain the drawback of validation set and what procedure will you use to choose the best model?



58258677-9FBE-478A-833B-0A5761AD73A7

midterm1-b9ace

#2 Page 8 of 10

8. (10 Points) Suppose we have a data set with five predictors,  $X_1 = \text{GPA}$ ,  $X_2 = \text{IQ}$ ,  $X_3 = \text{Level}$  (1 for College and 0 for High School),  $X_4 = \text{Interaction between GPA and IQ}$ , and  $X_5 = \text{Interaction between GPA and Level}$ . The response is starting salary after graduation (in thousands of dollars). Suppose we use least squares to fit the model, and get  $\beta_0 = 50$ ,  $\beta_1 = 20$ ,  $\beta_2 = 0.07$ ,  $\beta_3 = 35$ ,  $\beta_4 = 0.01$ ,  $\beta_5 = -10$ .

(a) Predict the salary of a college graduate with IQ of 110 and a GPA of 4.0.

- (b) True or false: Since the coefficient for the GPA/IQ interaction term is very small, there is very little evidence of an interaction effect. Justify your answer.

4B219E89-8B5B-4B8A-AB7A-DD1C25E22800

midterm1-b9ace

#2 Page 9 of 10



**Scrap Paper**



FA4F305E-2F45-40B5-97E4-48D1C5222AE6

midterm1-b9ace

#2 Page 10 of 10



Michigan State University  
CMSE381 - Data Science

Feb 10, 2023  
Dr. Xie

## CMSE381 - Midterm #1

First name (please write as legibly as possible within the boxes)

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

Last name

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

NetID

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

*I will adhere to the Spartan Code of Honor in completing this assignment.*

Signed: \_\_\_\_\_

1. Do not open this test booklet until you are directed to do so.
2. You will have class time (2:40-4:00pm) to complete the exam.
3. This exam is closed book, but you can use the cheatsheet provided by the instructor.
4. Throughout the test, show your work so that your reasoning is clear. Otherwise no credit will be given. **BOX** your answers. Partial credit will be given where warranted.
5. Do not spend too much time on any one problem. Read them all through first and attack them in the order that allows you to make the most progress. Good luck :P



CE5419D4-3FF0-4C65-B115-DD6711302568

midterm1-b9ace

#3 Page 2 of 10

1. (15 points)

(a) Logistic regression is used for regression.

TRUE      FALSE

(b) Any model will never have training error below the irreducible error.

TRUE      FALSE

(c) Increasing your model flexibility always results in a better model.

TRUE      FALSE

(d) A logistic regression model is set up so that the odds are linear.

TRUE      FALSE

(e) Circle all of the following that would represent a qualitative variable.

Age      Year      Dog\_breed      Country\_of\_origin

Student\_(True/False)      Weight      Speed      MPG

(f) What equation would you use to evaluate the result of a regression model?

(g) What equation would you use to evaluate the result of a classification model?



2. (15 Points) I'm building a model to predict amount of a given brand of dog food eaten by a collection of dogs. I have 100 dogs eat this dog food, and I collect information on their height, weight, breed, and whether they live with another dog in the house.
- (a) List all input variables and specify whether they are quantitative or qualitative.
- (b) Say our dog breeds sampled are Huskies, Terriers, and Spaniels. Write down your prediction model.
- (c) We found that weight is a key predictor, and its relationship with the response is beyond linear. What extension can you come up with? Write down your updated model.





B000D897-7BDF-4A90-A462-D4446D178A67

midterm1-b9ace

#3 Page 4 of 10

3. (15 points)

(a) What is bias-variance tradeoff? Explain the meaning of each term in the formula.

(b) Provide a sketch of typical (squared) bias, variance, training error, test error, and Bayes (or irreducible) error curves, on a single plot, as we go from less flexible statistical learning methods towards more flexible approaches. The  $x$ -axis should represent the amount of flexibility in the method, and the  $y$ -axis should represent the values for each curve. There should be five curves. Make sure to label each one.



(c) Explain why the (i) training error and (ii) testing error lines in your drawing have the shape displayed.



4. (10 points) The data for this example come from a study by Stamey et al. (1989). The data comes from a number of clinical measures in men who were about to receive a radical prostatectomy. The goal is to predict the log of PSA (**lpsa**) (a prostate-specific antigen measured in nanograms of PSA per milliliter of blood (ng/mL)) from a number of measurements. The variables are log cancer volume (**lcavol**), log prostate weight (**lweight**), age, log of the amount of benign prostatic hyperplasia (**lbph**), seminal vesicle invasion (**svi**), log of capsular penetration (**lcp**), Gleason score (**gleason**), and percent of Gleason scores 4 or 5 (**pgg45**).

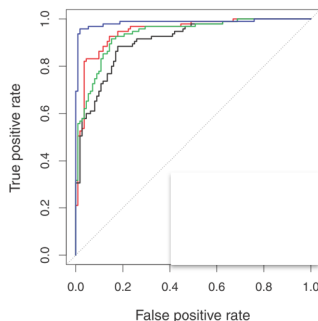
(a) Is this a case of regression or classification? Why?

(b) A linear model is fit to the data set, and the following table was returned.

Term	Coefficient	Std. Error	t-Score	p-value
Intercept	2.46	0.09	27.6	<0.00001
lcavol	0.68	0.13	5.37	<0.00001
lweight	0.26	0.1	2.75	0.00596
age	-0.14	0.1	-1.4	0.16153
lbph	0.21	0.1	2.06	0.039399
svi	0.31	0.12	2.47	0.013511
lcp	-0.29	0.15	-1.87	0.061484
gleason	-0.02	0.15	-0.15	0.880765
pgg45	0.27	0.15	1.74	0.081859

Are all the predictors useful? If yes, explain why. If not, explain what you would try to do next with the data.

5. (5 pts) We are trying to predict whether a patient will have a heart attack within one year. We have tried four different methods on a training dataset and have the following ROC curves. Which curve has the best performance on this training set? Justify your answer.





E1577DF9-EB58-4343-B2E4-233E65E5DB78

midterm1-b9ace

#3 Page 6 of 10

6. (15 Points) I get way too much email, so I decide to build a logistic regression model to predict whether a new incoming message is spam or not. I decide to just use a few variables:

$$\begin{aligned}X_1 &= \text{Number\_of\_sentences} \\X_2 &= \text{Number\_of\_references\_to\_a\_Nigerian\_Prince}\end{aligned}$$

and am training a logistic model to predict

$$Pr(Y = \text{spam} \mid X_1, X_2)$$

- (a) Write down the equation for the model you would train, using our standard notation with  $\beta_i$ 's.
- (b) If my trained model used  $\beta_0 = -13.1$ ,  $\beta_1 = 1.9$ , and  $\beta_2 = 6.1$ , what is the probability that a 5 sentence email with one reference to a Nigerian prince is spam? .
- (c) We know that if we miss an important email, it may cost a lot, How can you modify your model to avoid this?



7. (15 Points) Sparty generated 100 pair  $x$  and  $y$  via the following relationship:  $Y = 0.3 + 2x + x^2 + \epsilon$  but didn't tell you. You will try to find a good model to fit these data and more importantly to predict future response  $y$  when Sparty gives you another  $x$ . You will start with two models 1)  $Y = \beta_0 + \beta_1x$ ; 2)  $Y = \alpha_0 + \alpha_1x + \alpha_2x^2$ ; and 3)  $Y = \gamma_0 + \gamma_1x + \gamma_2x^2 + \gamma_3x^3$ .
- (a) If you use all 100 data to fit these three model and use the same 100 data to calculate MSE, which model will have the smallest MSE? Justify your answer
- (b) Since we focus on the ability to predict unseen data, you decide use validation set to evaluate the performance of the three models. There are two way to split the data: 1) 80 testing points and 20 training points or 2) 20 testing points and 80 training points. Which one will you choose? Justify your choice
- (c) From the in-class lab, we know validation set is not a good choice to choose model. Explain the drawback of validation set and what procedure will you use to choose the best model?



8. (10 Points) Suppose we have a data set with five predictors,  $X_1 = \text{GPA}$ ,  $X_2 = \text{IQ}$ ,  $X_3 = \text{Level}$  (1 for College and 0 for High School),  $X_4 = \text{Interaction between GPA and IQ}$ , and  $X_5 = \text{Interaction between GPA and Level}$ . The response is starting salary after graduation (in thousands of dollars). Suppose we use least squares to fit the model, and get  $\beta_0 = 50$ ,  $\beta_1 = 20$ ,  $\beta_2 = 0.07$ ,  $\beta_3 = 35$ ,  $\beta_4 = 0.01$ ,  $\beta_5 = -10$ .

(a) Predict the salary of a college graduate with IQ of 110 and a GPA of 4.0.

- (b) True or false: Since the coefficient for the GPA/IQ interaction term is very small, there is very little evidence of an interaction effect. Justify your answer.

A5ADA7D1-3799-4C31-9CA7-031CF581FDF4

midterm1-b9ace

#3 Page 9 of 10



**Scrap Paper**



0DFFDD06-6986-443E-9F8F-76B5827ECEDF

midterm1-b9ace

#3 Page 10 of 10



Michigan State University  
CMSE381 - Data Science

Feb 10, 2023  
Dr. Xie

## CMSE381 - Midterm #1

First name (please write as legibly as possible within the boxes)

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

Last name

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

NetID

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

*I will adhere to the Spartan Code of Honor in completing this assignment.*

Signed: \_\_\_\_\_

1. Do not open this test booklet until you are directed to do so.
2. You will have class time (2:40-4:00pm) to complete the exam.
3. This exam is closed book, but you can use the cheatsheet provided by the instructor.
4. Throughout the test, show your work so that your reasoning is clear. Otherwise no credit will be given. **BOX** your answers. Partial credit will be given where warranted.
5. Do not spend too much time on any one problem. Read them all through first and attack them in the order that allows you to make the most progress. Good luck :P





E81EBD48-FD87-4A27-820A-F2DC3338E240

midterm1-b9ace

#4 Page 2 of 10

1. (15 points)

(a) Logistic regression is used for regression.

TRUE      FALSE

(b) Any model will never have training error below the irreducible error.

TRUE      FALSE

(c) Increasing your model flexibility always results in a better model.

TRUE      FALSE

(d) A logistic regression model is set up so that the odds are linear.

TRUE      FALSE

(e) Circle all of the following that would represent a qualitative variable.

Age      Year      Dog\_breed      Country\_of\_origin

Student\_(True/False)      Weight      Speed      MPG

(f) What equation would you use to evaluate the result of a regression model?

(g) What equation would you use to evaluate the result of a classification model?



2. (15 Points) I'm building a model to predict amount of a given brand of dog food eaten by a collection of dogs. I have 100 dogs eat this dog food, and I collect information on their height, weight, breed, and whether they live with another dog in the house.
- (a) List all input variables and specify whether they are quantitative or qualitative.
- (b) Say our dog breeds sampled are Huskies, Terriers, and Spaniels. Write down your prediction model.
- (c) We found that weight is a key predictor, and its relationship with the response is beyond linear. What extension can you come up with? Write down your updated model.



D2DC1B36-F859-4337-9B3D-F6A63071DEA0

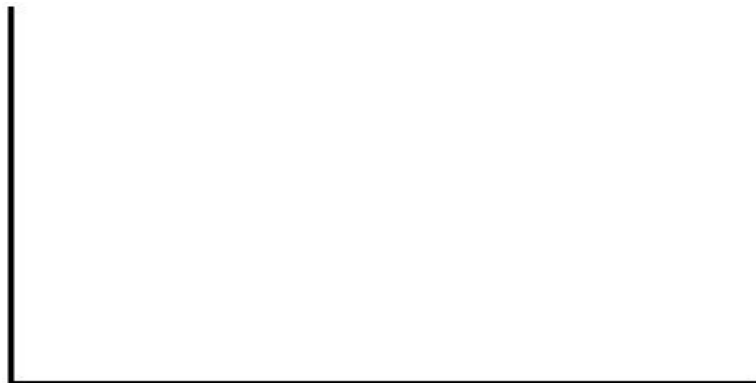
midterm1-b9ace

#4 Page 4 of 10

3. (15 points)

(a) What is bias-variance tradeoff? Explain the meaning of each term in the formula.

(b) Provide a sketch of typical (squared) bias, variance, training error, test error, and Bayes (or irreducible) error curves, on a single plot, as we go from less flexible statistical learning methods towards more flexible approaches. The  $x$ -axis should represent the amount of flexibility in the method, and the  $y$ -axis should represent the values for each curve. There should be five curves. Make sure to label each one.



(c) Explain why the (i) training error and (ii) testing error lines in your drawing have the shape displayed.



4. (10 points) The data for this example come from a study by Stamey et al. (1989). The data comes from a number of clinical measures in men who were about to receive a radical prostatectomy. The goal is to predict the log of PSA (**lpsa**) (a prostate-specific antigen measured in nanograms of PSA per milliliter of blood (ng/mL)) from a number of measurements. The variables are log cancer volume (**lcavol**), log prostate weight (**lweight**), age, log of the amount of benign prostatic hyperplasia (**lbph**), seminal vesicle invasion (**svi**), log of capsular penetration (**lcp**), Gleason score (**gleason**), and percent of Gleason scores 4 or 5 (**pgg45**).

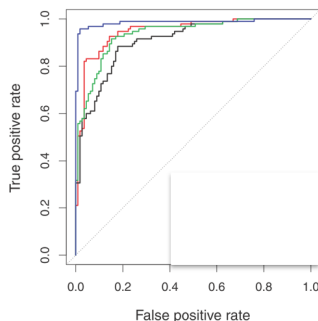
(a) Is this a case of regression or classification? Why?

(b) A linear model is fit to the data set, and the following table was returned.

Term	Coefficient	Std. Error	t-Score	p-value
Intercept	2.46	0.09	27.6	<0.00001
lcavol	0.68	0.13	5.37	<0.00001
lweight	0.26	0.1	2.75	0.00596
age	-0.14	0.1	-1.4	0.16153
lbph	0.21	0.1	2.06	0.039399
svi	0.31	0.12	2.47	0.013511
lcp	-0.29	0.15	-1.87	0.061484
gleason	-0.02	0.15	-0.15	0.880765
pgg45	0.27	0.15	1.74	0.081859

Are all the predictors useful? If yes, explain why. If not, explain what you would try to do next with the data.

5. (5 pts) We are trying to predict whether a patient will have a heart attack within one year. We have tried four different methods on a training dataset and have the following ROC curves. Which curve has the best performance on this training set? Justify your answer.





6. (15 Points) I get way too much email, so I decide to build a logistic regression model to predict whether a new incoming message is spam or not. I decide to just use a few variables:

$$\begin{aligned}X_1 &= \text{Number\_of\_sentences} \\X_2 &= \text{Number\_of\_references\_to\_a\_Nigerian\_Prince}\end{aligned}$$

and am training a logistic model to predict

$$Pr(Y = \text{spam} \mid X_1, X_2)$$

- (a) Write down the equation for the model you would train, using our standard notation with  $\beta_i$ 's.
- (b) If my trained model used  $\beta_0 = -13.1$ ,  $\beta_1 = 1.9$ , and  $\beta_2 = 6.1$ , what is the probability that a 5 sentence email with one reference to a Nigerian prince is spam? .
- (c) We know that if we miss an important email, it may cost a lot, How can you modify your model to avoid this?



7. (15 Points) Sparty generated 100 pair  $x$  and  $y$  via the following relationship:  $Y = 0.3 + 2x + x^2 + \epsilon$  but didn't tell you. You will try to find a good model to fit these data and more importantly to predict future response  $y$  when Sparty gives you another  $x$ . You will start with two models 1)  $Y = \beta_0 + \beta_1x$ ; 2)  $Y = \alpha_0 + \alpha_1x + \alpha_2x^2$ ; and 3)  $Y = \gamma_0 + \gamma_1x + \gamma_2x^2 + \gamma_3x^3$ .
- (a) If you use all 100 data to fit these three model and use the same 100 data to calculate MSE, which model will have the smallest MSE? Justify your answer
- (b) Since we focus on the ability to predict unseen data, you decide use validation set to evaluate the performance of the three models. There are two way to split the data: 1) 80 testing points and 20 training points or 2) 20 testing points and 80 training points. Which one will you choose? Justify your choice
- (c) From the in-class lab, we know validation set is not a good choice to choose model. Explain the drawback of validation set and what procedure will you use to choose the best model?



8. (10 Points) Suppose we have a data set with five predictors,  $X_1 = \text{GPA}$ ,  $X_2 = \text{IQ}$ ,  $X_3 = \text{Level}$  (1 for College and 0 for High School),  $X_4 = \text{Interaction between GPA and IQ}$ , and  $X_5 = \text{Interaction between GPA and Level}$ . The response is starting salary after graduation (in thousands of dollars). Suppose we use least squares to fit the model, and get  $\beta_0 = 50$ ,  $\beta_1 = 20$ ,  $\beta_2 = 0.07$ ,  $\beta_3 = 35$ ,  $\beta_4 = 0.01$ ,  $\beta_5 = -10$ .

(a) Predict the salary of a college graduate with IQ of 110 and a GPA of 4.0.

- (b) True or false: Since the coefficient for the GPA/IQ interaction term is very small, there is very little evidence of an interaction effect. Justify your answer.

8E6B7C2B-1985-4BFF-9997-4C221F97FFC4

midterm1-b9ace

#4 Page 9 of 10



## Scrap Paper





656DBCC5-287B-4AD0-B6E7-CEEF4C92BD9E

midterm1-b9ace

#4 Page 10 of 10



Michigan State University  
CMSE381 - Data Science

Feb 10, 2023  
Dr. Xie

## CMSE381 - Midterm #1

First name (please write as legibly as possible within the boxes)

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

Last name

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

NetID

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

*I will adhere to the Spartan Code of Honor in completing this assignment.*

Signed: \_\_\_\_\_

1. Do not open this test booklet until you are directed to do so.
2. You will have class time (2:40-4:00pm) to complete the exam.
3. This exam is closed book, but you can use the cheatsheet provided by the instructor.
4. Throughout the test, show your work so that your reasoning is clear. Otherwise no credit will be given. **BOX** your answers. Partial credit will be given where warranted.
5. Do not spend too much time on any one problem. Read them all through first and attack them in the order that allows you to make the most progress. Good luck :P



FD69CFB1-4F45-471A-AFC6-E147E123A32E

midterm1-b9ace

#5 Page 2 of 10

1. (15 points)

(a) Logistic regression is used for regression.

TRUE      FALSE

(b) Any model will never have training error below the irreducible error.

TRUE      FALSE

(c) Increasing your model flexibility always results in a better model.

TRUE      FALSE

(d) A logistic regression model is set up so that the odds are linear.

TRUE      FALSE

(e) Circle all of the following that would represent a qualitative variable.

Age      Year      Dog\_breed      Country\_of\_origin

Student\_(True/False)      Weight      Speed      MPG

(f) What equation would you use to evaluate the result of a regression model?

(g) What equation would you use to evaluate the result of a classification model?



2. (15 Points) I'm building a model to predict amount of a given brand of dog food eaten by a collection of dogs. I have 100 dogs eat this dog food, and I collect information on their height, weight, breed, and whether they live with another dog in the house.
- (a) List all input variables and specify whether they are quantitative or qualitative.
- (b) Say our dog breeds sampled are Huskies, Terriers, and Spaniels. Write down your prediction model.
- (c) We found that weight is a key predictor, and its relationship with the response is beyond linear. What extension can you come up with? Write down your updated model.



85C672F9-80FF-41B3-83B8-C80E5082FA46

midterm1-b9ace

#5 Page 4 of 10

3. (15 points)

(a) What is bias-variance tradeoff? Explain the meaning of each term in the formula.

(b) Provide a sketch of typical (squared) bias, variance, training error, test error, and Bayes (or irreducible) error curves, on a single plot, as we go from less flexible statistical learning methods towards more flexible approaches. The  $x$ -axis should represent the amount of flexibility in the method, and the  $y$ -axis should represent the values for each curve. There should be five curves. Make sure to label each one.



(c) Explain why the (i) training error and (ii) testing error lines in your drawing have the shape displayed.



4. (10 points) The data for this example come from a study by Stamey et al. (1989). The data comes from a number of clinical measures in men who were about to receive a radical prostatectomy. The goal is to predict the log of PSA (**lpsa**) (a prostate-specific antigen measured in nanograms of PSA per milliliter of blood (ng/mL)) from a number of measurements. The variables are log cancer volume (**lcavol**), log prostate weight (**lweight**), age, log of the amount of benign prostatic hyperplasia (**lbph**), seminal vesicle invasion (**svi**), log of capsular penetration (**lcp**), Gleason score (**gleason**), and percent of Gleason scores 4 or 5 (**pgg45**).

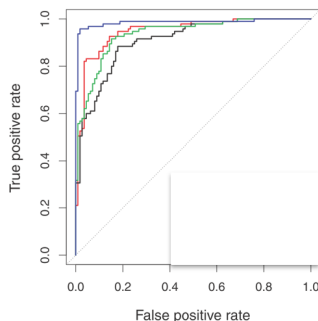
(a) Is this a case of regression or classification? Why?

(b) A linear model is fit to the data set, and the following table was returned.

Term	Coefficient	Std. Error	t-Score	p-value
Intercept	2.46	0.09	27.6	<0.00001
lcavol	0.68	0.13	5.37	<0.00001
lweight	0.26	0.1	2.75	0.00596
age	-0.14	0.1	-1.4	0.16153
lbph	0.21	0.1	2.06	0.039399
svi	0.31	0.12	2.47	0.013511
lcp	-0.29	0.15	-1.87	0.061484
gleason	-0.02	0.15	-0.15	0.880765
pgg45	0.27	0.15	1.74	0.081859

Are all the predictors useful? If yes, explain why. If not, explain what you would try to do next with the data.

5. (5 pts) We are trying to predict whether a patient will have a heart attack within one year. We have tried four different methods on a training dataset and have the following ROC curves. Which curve has the best performance on this training set? Justify your answer.





6. (15 Points) I get way too much email, so I decide to build a logistic regression model to predict whether a new incoming message is spam or not. I decide to just use a few variables:

$$\begin{aligned}X_1 &= \text{Number\_of\_sentences} \\X_2 &= \text{Number\_of\_references\_to\_a\_Nigerian\_Prince}\end{aligned}$$

and am training a logistic model to predict

$$Pr(Y = \text{spam} \mid X_1, X_2)$$

- (a) Write down the equation for the model you would train, using our standard notation with  $\beta_i$ 's.
- (b) If my trained model used  $\beta_0 = -13.1$ ,  $\beta_1 = 1.9$ , and  $\beta_2 = 6.1$ , what is the probability that a 5 sentence email with one reference to a Nigerian prince is spam? .
- (c) We know that if we miss an important email, it may cost a lot, How can you modify your model to avoid this?



7. (15 Points) Sparty generated 100 pair  $x$  and  $y$  via the following relationship:  $Y = 0.3 + 2x + x^2 + \epsilon$  but didn't tell you. You will try to find a good model to fit these data and more importantly to predict future response  $y$  when Sparty gives you another  $x$ . You will start with two models 1)  $Y = \beta_0 + \beta_1x$ ; 2)  $Y = \alpha_0 + \alpha_1x + \alpha_2x^2$ ; and 3)  $Y = \gamma_0 + \gamma_1x + \gamma_2x^2 + \gamma_3x^3$ .
- (a) If you use all 100 data to fit these three model and use the same 100 data to calculate MSE, which model will have the smallest MSE? Justify your answer
- (b) Since we focus on the ability to predict unseen data, you decide use validation set to evaluate the performance of the three models. There are two way to split the data: 1) 80 testing points and 20 training points or 2) 20 testing points and 80 training points. Which one will you choose? Justify your choice
- (c) From the in-class lab, we know validation set is not a good choice to choose model. Explain the drawback of validation set and what procedure will you use to choose the best model?





8. (10 Points) Suppose we have a data set with five predictors,  $X_1 = \text{GPA}$ ,  $X_2 = \text{IQ}$ ,  $X_3 = \text{Level}$  (1 for College and 0 for High School),  $X_4 = \text{Interaction between GPA and IQ}$ , and  $X_5 = \text{Interaction between GPA and Level}$ . The response is starting salary after graduation (in thousands of dollars). Suppose we use least squares to fit the model, and get  $\beta_0 = 50$ ,  $\beta_1 = 20$ ,  $\beta_2 = 0.07$ ,  $\beta_3 = 35$ ,  $\beta_4 = 0.01$ ,  $\beta_5 = -10$ .

(a) Predict the salary of a college graduate with IQ of 110 and a GPA of 4.0.

- (b) True or false: Since the coefficient for the GPA/IQ interaction term is very small, there is very little evidence of an interaction effect. Justify your answer.

647AB94E-75EB-49F3-9B87-D275BBA16A90

midterm1-b9ace

#5 Page 9 of 10



## Scrap Paper



874F568F-5F58-40C2-8A4D-065CE022C12B

midterm1-b9ace

#5 Page 10 of 10



Michigan State University  
CMSE381 - Data Science

Feb 10, 2023  
Dr. Xie

## CMSE381 - Midterm #1

First name (please write as legibly as possible within the boxes)

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

Last name

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

NetID

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

*I will adhere to the Spartan Code of Honor in completing this assignment.*

Signed: \_\_\_\_\_

1. Do not open this test booklet until you are directed to do so.
2. You will have class time (2:40-4:00pm) to complete the exam.
3. This exam is closed book, but you can use the cheatsheet provided by the instructor.
4. Throughout the test, show your work so that your reasoning is clear. Otherwise no credit will be given. **BOX** your answers. Partial credit will be given where warranted.
5. Do not spend too much time on any one problem. Read them all through first and attack them in the order that allows you to make the most progress. Good luck :P



BDF66EEE-FAB5-4EF5-901D-083A98EED0A7

midterm1-b9ace

#6 Page 2 of 10

1. (15 points)

(a) Logistic regression is used for regression.

TRUE      FALSE

(b) Any model will never have training error below the irreducible error.

TRUE      FALSE

(c) Increasing your model flexibility always results in a better model.

TRUE      FALSE

(d) A logistic regression model is set up so that the odds are linear.

TRUE      FALSE

(e) Circle all of the following that would represent a qualitative variable.

Age      Year      Dog\_breed      Country\_of\_origin

Student\_(True/False)      Weight      Speed      MPG

(f) What equation would you use to evaluate the result of a regression model?

(g) What equation would you use to evaluate the result of a classification model?



2. (15 Points) I'm building a model to predict amount of a given brand of dog food eaten by a collection of dogs. I have 100 dogs eat this dog food, and I collect information on their height, weight, breed, and whether they live with another dog in the house.
- (a) List all input variables and specify whether they are quantitative or qualitative.
  - (b) Say our dog breeds sampled are Huskies, Terriers, and Spaniels. Write down your prediction model.
  - (c) We found that weight is a key predictor, and its relationship with the response is beyond linear. What extension can you come up with? Write down your updated model.



B5770089-BC73-4EA0-B081-77310E187E31

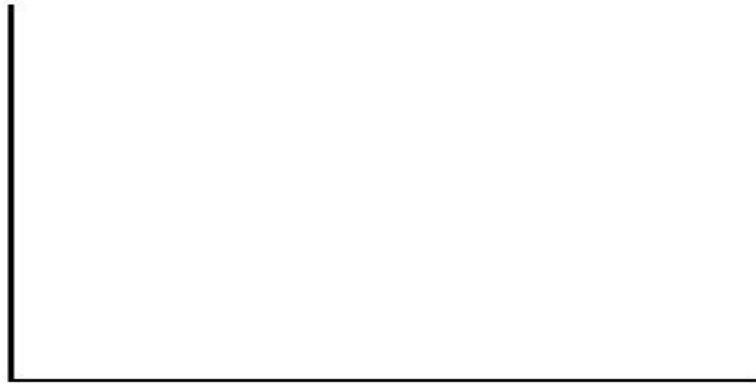
midterm1-b9ace

#6 Page 4 of 10

3. (15 points)

(a) What is bias-variance tradeoff? Explain the meaning of each term in the formula.

(b) Provide a sketch of typical (squared) bias, variance, training error, test error, and Bayes (or irreducible) error curves, on a single plot, as we go from less flexible statistical learning methods towards more flexible approaches. The  $x$ -axis should represent the amount of flexibility in the method, and the  $y$ -axis should represent the values for each curve. There should be five curves. Make sure to label each one.



(c) Explain why the (i) training error and (ii) testing error lines in your drawing have the shape displayed.



4. (10 points) The data for this example come from a study by Stamey et al. (1989). The data comes from a number of clinical measures in men who were about to receive a radical prostatectomy. The goal is to predict the log of PSA (**lpsa**) (a prostate-specific antigen measured in nanograms of PSA per milliliter of blood (ng/mL)) from a number of measurements. The variables are log cancer volume (**lcavol**), log prostate weight (**lweight**), age, log of the amount of benign prostatic hyperplasia (**lbph**), seminal vesicle invasion (**svi**), log of capsular penetration (**lcp**), Gleason score (**gleason**), and percent of Gleason scores 4 or 5 (**pgg45**).

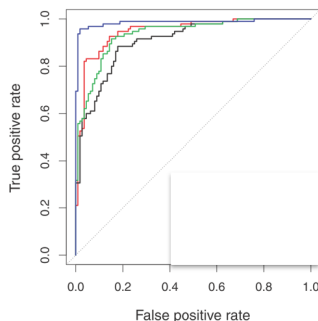
(a) Is this a case of regression or classification? Why?

(b) A linear model is fit to the data set, and the following table was returned.

Term	Coefficient	Std. Error	t-Score	p-value
Intercept	2.46	0.09	27.6	<0.00001
lcavol	0.68	0.13	5.37	<0.00001
lweight	0.26	0.1	2.75	0.00596
age	-0.14	0.1	-1.4	0.16153
lbph	0.21	0.1	2.06	0.039399
svi	0.31	0.12	2.47	0.013511
lcp	-0.29	0.15	-1.87	0.061484
gleason	-0.02	0.15	-0.15	0.880765
pgg45	0.27	0.15	1.74	0.081859

Are all the predictors useful? If yes, explain why. If not, explain what you would try to do next with the data.

5. (5 pts) We are trying to predict whether a patient will have a heart attack within one year. We have tried four different methods on a training dataset and have the following ROC curves. Which curve has the best performance on this training set? Justify your answer.







6. (15 Points) I get way too much email, so I decide to build a logistic regression model to predict whether a new incoming message is spam or not. I decide to just use a few variables:

$$\begin{aligned}X_1 &= \text{Number\_of\_sentences} \\X_2 &= \text{Number\_of\_references\_to\_a\_Nigerian\_Prince}\end{aligned}$$

and am training a logistic model to predict

$$Pr(Y = \text{spam} \mid X_1, X_2)$$

- (a) Write down the equation for the model you would train, using our standard notation with  $\beta_i$ 's.
- (b) If my trained model used  $\beta_0 = -13.1$ ,  $\beta_1 = 1.9$ , and  $\beta_2 = 6.1$ , what is the probability that a 5 sentence email with one reference to a Nigerian prince is spam? .
- (c) We know that if we miss an important email, it may cost a lot, How can you modify your model to avoid this?



7. (15 Points) Sparty generated 100 pair  $x$  and  $y$  via the following relationship:  $Y = 0.3 + 2x + x^2 + \epsilon$  but didn't tell you. You will try to find a good model to fit these data and more importantly to predict future response  $y$  when Sparty gives you another  $x$ . You will start with two models 1)  $Y = \beta_0 + \beta_1x$ ; 2)  $Y = \alpha_0 + \alpha_1x + \alpha_2x^2$ ; and 3)  $Y = \gamma_0 + \gamma_1x + \gamma_2x^2 + \gamma_3x^3$ .
- (a) If you use all 100 data to fit these three model and use the same 100 data to calculate MSE, which model will have the smallest MSE? Justify your answer
- (b) Since we focus on the ability to predict unseen data, you decide use validation set to evaluate the performance of the three models. There are two way to split the data: 1) 80 testing points and 20 training points or 2) 20 testing points and 80 training points. Which one will you choose? Justify your choice
- (c) From the in-class lab, we know validation set is not a good choice to choose model. Explain the drawback of validation set and what procedure will you use to choose the best model?



8. (10 Points) Suppose we have a data set with five predictors,  $X_1 = \text{GPA}$ ,  $X_2 = \text{IQ}$ ,  $X_3 = \text{Level}$  (1 for College and 0 for High School),  $X_4 = \text{Interaction between GPA and IQ}$ , and  $X_5 = \text{Interaction between GPA and Level}$ . The response is starting salary after graduation (in thousands of dollars). Suppose we use least squares to fit the model, and get  $\beta_0 = 50$ ,  $\beta_1 = 20$ ,  $\beta_2 = 0.07$ ,  $\beta_3 = 35$ ,  $\beta_4 = 0.01$ ,  $\beta_5 = -10$ .

(a) Predict the salary of a college graduate with IQ of 110 and a GPA of 4.0.

- (b) True or false: Since the coefficient for the GPA/IQ interaction term is very small, there is very little evidence of an interaction effect. Justify your answer.

00A74A0D-992B-4499-BF94-5A7BC41405CC

midterm1-b9ace

#6 Page 9 of 10



## Scrap Paper



C31D58B4-639D-4BEB-9B88-214BC04F05E3

midterm1-b9ace

#6 Page 10 of 10



Michigan State University  
CMSE381 - Data Science

Feb 10, 2023  
Dr. Xie

## CMSE381 - Midterm #1

First name (please write as legibly as possible within the boxes)

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

Last name

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

NetID

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

*I will adhere to the Spartan Code of Honor in completing this assignment.*

Signed: \_\_\_\_\_

1. Do not open this test booklet until you are directed to do so.
2. You will have class time (2:40-4:00pm) to complete the exam.
3. This exam is closed book, but you can use the cheatsheet provided by the instructor.
4. Throughout the test, show your work so that your reasoning is clear. Otherwise no credit will be given. **BOX** your answers. Partial credit will be given where warranted.
5. Do not spend too much time on any one problem. Read them all through first and attack them in the order that allows you to make the most progress. Good luck :P



14A4F015-F476-4949-AEF7-36A607EB318F

midterm1-b9ace

#7 Page 2 of 10

1. (15 points)

(a) Logistic regression is used for regression.

TRUE      FALSE

(b) Any model will never have training error below the irreducible error.

TRUE      FALSE

(c) Increasing your model flexibility always results in a better model.

TRUE      FALSE

(d) A logistic regression model is set up so that the odds are linear.

TRUE      FALSE

(e) Circle all of the following that would represent a qualitative variable.

Age      Year      Dog\_breed      Country\_of\_origin

Student\_(True/False)      Weight      Speed      MPG

(f) What equation would you use to evaluate the result of a regression model?

(g) What equation would you use to evaluate the result of a classification model?



2. (15 Points) I'm building a model to predict amount of a given brand of dog food eaten by a collection of dogs. I have 100 dogs eat this dog food, and I collect information on their height, weight, breed, and whether they live with another dog in the house.
- (a) List all input variables and specify whether they are quantitative or qualitative.
- (b) Say our dog breeds sampled are Huskies, Terriers, and Spaniels. Write down your prediction model.
- (c) We found that weight is a key predictor, and its relationship with the response is beyond linear. What extension can you come up with? Write down your updated model.





3. (15 points)

(a) What is bias-variance tradeoff? Explain the meaning of each term in the formula.

(b) Provide a sketch of typical (squared) bias, variance, training error, test error, and Bayes (or irreducible) error curves, on a single plot, as we go from less flexible statistical learning methods towards more flexible approaches. The  $x$ -axis should represent the amount of flexibility in the method, and the  $y$ -axis should represent the values for each curve. There should be five curves. Make sure to label each one.



(c) Explain why the (i) training error and (ii) testing error lines in your drawing have the shape displayed.



4. (10 points) The data for this example come from a study by Stamey et al. (1989). The data comes from a number of clinical measures in men who were about to receive a radical prostatectomy. The goal is to predict the log of PSA (**lpsa**) (a prostate-specific antigen measured in nanograms of PSA per milliliter of blood (ng/mL)) from a number of measurements. The variables are log cancer volume (**lcavol**), log prostate weight (**lweight**), age, log of the amount of benign prostatic hyperplasia (**lbph**), seminal vesicle invasion (**svi**), log of capsular penetration (**lcp**), Gleason score (**gleason**), and percent of Gleason scores 4 or 5 (**pgg45**).

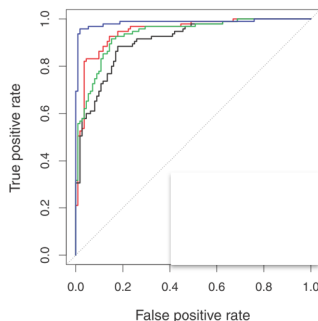
(a) Is this a case of regression or classification? Why?

(b) A linear model is fit to the data set, and the following table was returned.

Term	Coefficient	Std. Error	t-Score	p-value
Intercept	2.46	0.09	27.6	<0.00001
lcavol	0.68	0.13	5.37	<0.00001
lweight	0.26	0.1	2.75	0.00596
age	-0.14	0.1	-1.4	0.16153
lbph	0.21	0.1	2.06	0.039399
svi	0.31	0.12	2.47	0.013511
lcp	-0.29	0.15	-1.87	0.061484
gleason	-0.02	0.15	-0.15	0.880765
pgg45	0.27	0.15	1.74	0.081859

Are all the predictors useful? If yes, explain why. If not, explain what you would try to do next with the data.

5. (5 pts) We are trying to predict whether a patient will have a heart attack within one year. We have tried four different methods on a training dataset and have the following ROC curves. Which curve has the best performance on this training set? Justify your answer.





6. (15 Points) I get way too much email, so I decide to build a logistic regression model to predict whether a new incoming message is spam or not. I decide to just use a few variables:

$$\begin{aligned}X_1 &= \text{Number\_of\_sentences} \\X_2 &= \text{Number\_of\_references\_to\_a\_Nigerian\_Prince}\end{aligned}$$

and am training a logistic model to predict

$$Pr(Y = \text{spam} \mid X_1, X_2)$$

- (a) Write down the equation for the model you would train, using our standard notation with  $\beta_i$ 's.
- (b) If my trained model used  $\beta_0 = -13.1$ ,  $\beta_1 = 1.9$ , and  $\beta_2 = 6.1$ , what is the probability that a 5 sentence email with one reference to a Nigerian prince is spam? .
- (c) We know that if we miss an important email, it may cost a lot, How can you modify your model to avoid this?



7. (15 Points) Sparty generated 100 pair  $x$  and  $y$  via the following relationship:  $Y = 0.3 + 2x + x^2 + \epsilon$  but didn't tell you. You will try to find a good model to fit these data and more importantly to predict future response  $y$  when Sparty gives you another  $x$ . You will start with two models 1)  $Y = \beta_0 + \beta_1x$ ; 2)  $Y = \alpha_0 + \alpha_1x + \alpha_2x^2$ ; and 3)  $Y = \gamma_0 + \gamma_1x + \gamma_2x^2 + \gamma_3x^3$ .
- (a) If you use all 100 data to fit these three model and use the same 100 data to calculate MSE, which model will have the smallest MSE? Justify your answer
- (b) Since we focus on the ability to predict unseen data, you decide use validation set to evaluate the performance of the three models. There are two way to split the data: 1) 80 testing points and 20 training points or 2) 20 testing points and 80 training points. Which one will you choose? Justify your choice
- (c) From the in-class lab, we know validation set is not a good choice to choose model. Explain the drawback of validation set and what procedure will you use to choose the best model?



8. (10 Points) Suppose we have a data set with five predictors,  $X_1 = \text{GPA}$ ,  $X_2 = \text{IQ}$ ,  $X_3 = \text{Level}$  (1 for College and 0 for High School),  $X_4 = \text{Interaction between GPA and IQ}$ , and  $X_5 = \text{Interaction between GPA and Level}$ . The response is starting salary after graduation (in thousands of dollars). Suppose we use least squares to fit the model, and get  $\beta_0 = 50$ ,  $\beta_1 = 20$ ,  $\beta_2 = 0.07$ ,  $\beta_3 = 35$ ,  $\beta_4 = 0.01$ ,  $\beta_5 = -10$ .

(a) Predict the salary of a college graduate with IQ of 110 and a GPA of 4.0.

- (b) True or false: Since the coefficient for the GPA/IQ interaction term is very small, there is very little evidence of an interaction effect. Justify your answer.

6A7E1E75-D5B5-45F2-BB4E-13E92E718967

midterm1-b9ace

#7 Page 9 of 10



## Scrap Paper



6B2ED63E-D852-4DC7-8A39-6D7FF669E2F5

midterm1-b9ace

#7 Page 10 of 10



Michigan State University  
CMSE381 - Data Science

Feb 10, 2023  
Dr. Xie

## CMSE381 - Midterm #1

First name (please write as legibly as possible within the boxes)

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

Last name

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

NetID

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

*I will adhere to the Spartan Code of Honor in completing this assignment.*

Signed: \_\_\_\_\_

1. Do not open this test booklet until you are directed to do so.
2. You will have class time (2:40-4:00pm) to complete the exam.
3. This exam is closed book, but you can use the cheatsheet provided by the instructor.
4. Throughout the test, show your work so that your reasoning is clear. Otherwise no credit will be given. **BOX** your answers. Partial credit will be given where warranted.
5. Do not spend too much time on any one problem. Read them all through first and attack them in the order that allows you to make the most progress. Good luck :P





012EF59E-981A-4D6E-B7DA-FFF824BF9EA8

midterm1-b9ace

#8 Page 2 of 10

1. (15 points)

(a) Logistic regression is used for regression.

TRUE      FALSE

(b) Any model will never have training error below the irreducible error.

TRUE      FALSE

(c) Increasing your model flexibility always results in a better model.

TRUE      FALSE

(d) A logistic regression model is set up so that the odds are linear.

TRUE      FALSE

(e) Circle all of the following that would represent a qualitative variable.

Age      Year      Dog\_breed      Country\_of\_origin

Student\_(True/False)      Weight      Speed      MPG

(f) What equation would you use to evaluate the result of a regression model?

(g) What equation would you use to evaluate the result of a classification model?



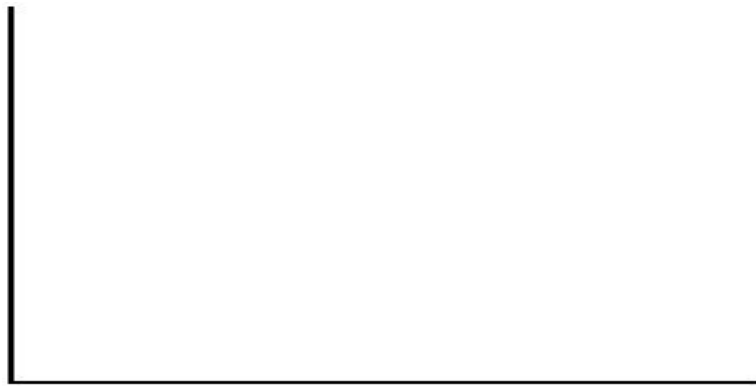
2. (15 Points) I'm building a model to predict amount of a given brand of dog food eaten by a collection of dogs. I have 100 dogs eat this dog food, and I collect information on their height, weight, breed, and whether they live with another dog in the house.
- (a) List all input variables and specify whether they are quantitative or qualitative.
- (b) Say our dog breeds sampled are Huskies, Terriers, and Spaniels. Write down your prediction model.
- (c) We found that weight is a key predictor, and its relationship with the response is beyond linear. What extension can you come up with? Write down your updated model.



3. (15 points)

(a) What is bias-variance tradeoff? Explain the meaning of each term in the formula.

(b) Provide a sketch of typical (squared) bias, variance, training error, test error, and Bayes (or irreducible) error curves, on a single plot, as we go from less flexible statistical learning methods towards more flexible approaches. The  $x$ -axis should represent the amount of flexibility in the method, and the  $y$ -axis should represent the values for each curve. There should be five curves. Make sure to label each one.



(c) Explain why the (i) training error and (ii) testing error lines in your drawing have the shape displayed.



4. (10 points) The data for this example come from a study by Stamey et al. (1989). The data comes from a number of clinical measures in men who were about to receive a radical prostatectomy. The goal is to predict the log of PSA (**lpsa**) (a prostate-specific antigen measured in nanograms of PSA per milliliter of blood (ng/mL)) from a number of measurements. The variables are log cancer volume (**lcavol**), log prostate weight (**lweight**), age, log of the amount of benign prostatic hyperplasia (**lbph**), seminal vesicle invasion (**svi**), log of capsular penetration (**lcp**), Gleason score (**gleason**), and percent of Gleason scores 4 or 5 (**pgg45**).

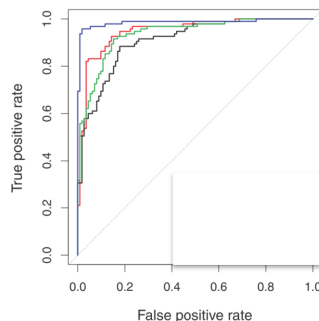
(a) Is this a case of regression or classification? Why?

(b) A linear model is fit to the data set, and the following table was returned.

Term	Coefficient	Std. Error	t-Score	p-value
Intercept	2.46	0.09	27.6	<0.00001
lcavol	0.68	0.13	5.37	<0.00001
lweight	0.26	0.1	2.75	0.00596
age	-0.14	0.1	-1.4	0.16153
lbph	0.21	0.1	2.06	0.039399
svi	0.31	0.12	2.47	0.013511
lcp	-0.29	0.15	-1.87	0.061484
gleason	-0.02	0.15	-0.15	0.880765
pgg45	0.27	0.15	1.74	0.081859

Are all the predictors useful? If yes, explain why. If not, explain what you would try to do next with the data.

5. (5 pts) We are trying to predict whether a patient will have a heart attack within one year. We have tried four different methods on a training dataset and have the following ROC curves. Which curve has the best performance on this training set? Justify your answer.





6. (15 Points) I get way too much email, so I decide to build a logistic regression model to predict whether a new incoming message is spam or not. I decide to just use a few variables:

$$\begin{aligned}X_1 &= \text{Number\_of\_sentences} \\X_2 &= \text{Number\_of\_references\_to\_a\_Nigerian\_Prince}\end{aligned}$$

and am training a logistic model to predict

$$Pr(Y = \text{spam} \mid X_1, X_2)$$

- (a) Write down the equation for the model you would train, using our standard notation with  $\beta_i$ 's.
- (b) If my trained model used  $\beta_0 = -13.1$ ,  $\beta_1 = 1.9$ , and  $\beta_2 = 6.1$ , what is the probability that a 5 sentence email with one reference to a Nigerian prince is spam? .
- (c) We know that if we miss an important email, it may cost a lot, How can you modify your model to avoid this?



7. (15 Points) Sparty generated 100 pair  $x$  and  $y$  via the following relationship:  $Y = 0.3 + 2x + x^2 + \epsilon$  but didn't tell you. You will try to find a good model to fit these data and more importantly to predict future response  $y$  when Sparty gives you another  $x$ . You will start with two models 1)  $Y = \beta_0 + \beta_1x$ ; 2)  $Y = \alpha_0 + \alpha_1x + \alpha_2x^2$ ; and 3)  $Y = \gamma_0 + \gamma_1x + \gamma_2x^2 + \gamma_3x^3$ .
- (a) If you use all 100 data to fit these three model and use the same 100 data to calculate MSE, which model will have the smallest MSE? Justify your answer
- (b) Since we focus on the ability to predict unseen data, you decide use validation set to evaluate the performance of the three models. There are two way to split the data: 1) 80 testing points and 20 training points or 2) 20 testing points and 80 training points. Which one will you choose? Justify your choice
- (c) From the in-class lab, we know validation set is not a good choice to choose model. Explain the drawback of validation set and what procedure will you use to choose the best model?



92BD9C1F-0BAD-4AE9-861C-541073D45D95

midterm1-b9ace

#8 Page 8 of 10

8. (10 Points) Suppose we have a data set with five predictors,  $X_1 = \text{GPA}$ ,  $X_2 = \text{IQ}$ ,  $X_3 = \text{Level}$  (1 for College and 0 for High School),  $X_4 = \text{Interaction between GPA and IQ}$ , and  $X_5 = \text{Interaction between GPA and Level}$ . The response is starting salary after graduation (in thousands of dollars). Suppose we use least squares to fit the model, and get  $\beta_0 = 50$ ,  $\beta_1 = 20$ ,  $\beta_2 = 0.07$ ,  $\beta_3 = 35$ ,  $\beta_4 = 0.01$ ,  $\beta_5 = -10$ .

(a) Predict the salary of a college graduate with IQ of 110 and a GPA of 4.0.

(b) True or false: Since the coefficient for the GPA/IQ interaction term is very small, there is very little evidence of an interaction effect. Justify your answer.

BA7C9303-6A69-4539-83F3-E6DFCF3440DB

midterm1-b9ace

#8 Page 9 of 10



**Scrap Paper**





DF7A8F31-E191-4307-9D85-BBDB43F5F8F8

midterm1-b9ace

#8 Page 10 of 10



Michigan State University  
CMSE381 - Data Science

Feb 10, 2023  
Dr. Xie

## CMSE381 - Midterm #1

First name (please write as legibly as possible within the boxes)

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

Last name

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

NetID

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

*I will adhere to the Spartan Code of Honor in completing this assignment.*

Signed: \_\_\_\_\_

1. Do not open this test booklet until you are directed to do so.
2. You will have class time (2:40-4:00pm) to complete the exam.
3. This exam is closed book, but you can use the cheatsheet provided by the instructor.
4. Throughout the test, show your work so that your reasoning is clear. Otherwise no credit will be given. **BOX** your answers. Partial credit will be given where warranted.
5. Do not spend too much time on any one problem. Read them all through first and attack them in the order that allows you to make the most progress. Good luck :P



1. (15 points)

(a) Logistic regression is used for regression.

TRUE      FALSE

(b) Any model will never have training error below the irreducible error.

TRUE      FALSE

(c) Increasing your model flexibility always results in a better model.

TRUE      FALSE

(d) A logistic regression model is set up so that the odds are linear.

TRUE      FALSE

(e) Circle all of the following that would represent a qualitative variable.

Age      Year      Dog\_breed      Country\_of\_origin

Student\_(True/False)      Weight      Speed      MPG

(f) What equation would you use to evaluate the result of a regression model?

(g) What equation would you use to evaluate the result of a classification model?



2. (15 Points) I'm building a model to predict amount of a given brand of dog food eaten by a collection of dogs. I have 100 dogs eat this dog food, and I collect information on their height, weight, breed, and whether they live with another dog in the house.
- (a) List all input variables and specify whether they are quantitative or qualitative.
- (b) Say our dog breeds sampled are Huskies, Terriers, and Spaniels. Write down your prediction model.
- (c) We found that weight is a key predictor, and its relationship with the response is beyond linear. What extension can you come up with? Write down your updated model.



8D8AE0A0-332A-49B6-B8C4-4C0CA921BABD

midterm1-b9ace

#9 Page 4 of 10

3. (15 points)

(a) What is bias-variance tradeoff? Explain the meaning of each term in the formula.

(b) Provide a sketch of typical (squared) bias, variance, training error, test error, and Bayes (or irreducible) error curves, on a single plot, as we go from less flexible statistical learning methods towards more flexible approaches. The  $x$ -axis should represent the amount of flexibility in the method, and the  $y$ -axis should represent the values for each curve. There should be five curves. Make sure to label each one.



(c) Explain why the (i) training error and (ii) testing error lines in your drawing have the shape displayed.



4. (10 points) The data for this example come from a study by Stamey et al. (1989). The data comes from a number of clinical measures in men who were about to receive a radical prostatectomy. The goal is to predict the log of PSA (**lpsa**) (a prostate-specific antigen measured in nanograms of PSA per milliliter of blood (ng/mL)) from a number of measurements. The variables are log cancer volume (**lcavol**), log prostate weight (**lweight**), age, log of the amount of benign prostatic hyperplasia (**lbph**), seminal vesicle invasion (**svi**), log of capsular penetration (**lcp**), Gleason score (**gleason**), and percent of Gleason scores 4 or 5 (**pgg45**).

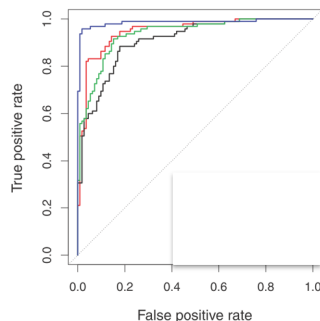
(a) Is this a case of regression or classification? Why?

(b) A linear model is fit to the data set, and the following table was returned.

Term	Coefficient	Std. Error	t-Score	p-value
Intercept	2.46	0.09	27.6	<0.00001
lcavol	0.68	0.13	5.37	<0.00001
lweight	0.26	0.1	2.75	0.00596
age	-0.14	0.1	-1.4	0.16153
lbph	0.21	0.1	2.06	0.039399
svi	0.31	0.12	2.47	0.013511
lcp	-0.29	0.15	-1.87	0.061484
gleason	-0.02	0.15	-0.15	0.880765
pgg45	0.27	0.15	1.74	0.081859

Are all the predictors useful? If yes, explain why. If not, explain what you would try to do next with the data.

5. (5 pts) We are trying to predict whether a patient will have a heart attack within one year. We have tried four different methods on a training dataset and have the following ROC curves. Which curve has the best performance on this training set? Justify your answer.





3B0C34DB-7F92-4F17-8B3B-D544034C402D

midterm1-b9ace

#9 Page 6 of 10

6. (15 Points) I get way too much email, so I decide to build a logistic regression model to predict whether a new incoming message is spam or not. I decide to just use a few variables:

$$\begin{aligned}X_1 &= \text{Number\_of\_sentences} \\X_2 &= \text{Number\_of\_references\_to\_a\_Nigerian\_Prince}\end{aligned}$$

and am training a logistic model to predict

$$Pr(Y = \text{spam} \mid X_1, X_2)$$

- (a) Write down the equation for the model you would train, using our standard notation with  $\beta_i$ 's.
- (b) If my trained model used  $\beta_0 = -13.1$ ,  $\beta_1 = 1.9$ , and  $\beta_2 = 6.1$ , what is the probability that a 5 sentence email with one reference to a Nigerian prince is spam? .
- (c) We know that if we miss an important email, it may cost a lot, How can you modify your model to avoid this?



7. (15 Points) Sparty generated 100 pair  $x$  and  $y$  via the following relationship:  $Y = 0.3 + 2x + x^2 + \epsilon$  but didn't tell you. You will try to find a good model to fit these data and more importantly to predict future response  $y$  when Sparty gives you another  $x$ . You will start with two models 1)  $Y = \beta_0 + \beta_1x$ ; 2)  $Y = \alpha_0 + \alpha_1x + \alpha_2x^2$ ; and 3)  $Y = \gamma_0 + \gamma_1x + \gamma_2x^2 + \gamma_3x^3$ .
- (a) If you use all 100 data to fit these three model and use the same 100 data to calculate MSE, which model will have the smallest MSE? Justify your answer
- (b) Since we focus on the ability to predict unseen data, you decide use validation set to evaluate the performance of the three models. There are two way to split the data: 1) 80 testing points and 20 training points or 2) 20 testing points and 80 training points. Which one will you choose? Justify your choice
- (c) From the in-class lab, we know validation set is not a good choice to choose model. Explain the drawback of validation set and what procedure will you use to choose the best model?





356CFB08-2A85-44CE-A487-1BB32A7D72C1

midterm1-b9ace

#9 Page 8 of 10

8. (10 Points) Suppose we have a data set with five predictors,  $X_1 = \text{GPA}$ ,  $X_2 = \text{IQ}$ ,  $X_3 = \text{Level}$  (1 for College and 0 for High School),  $X_4 = \text{Interaction between GPA and IQ}$ , and  $X_5 = \text{Interaction between GPA and Level}$ . The response is starting salary after graduation (in thousands of dollars). Suppose we use least squares to fit the model, and get  $\beta_0 = 50$ ,  $\beta_1 = 20$ ,  $\beta_2 = 0.07$ ,  $\beta_3 = 35$ ,  $\beta_4 = 0.01$ ,  $\beta_5 = -10$ .

(a) Predict the salary of a college graduate with IQ of 110 and a GPA of 4.0.

- (b) True or false: Since the coefficient for the GPA/IQ interaction term is very small, there is very little evidence of an interaction effect. Justify your answer.

DDFE69E9-3E30-4F68-98A8-365A71DE5009

midterm1-b9ace

#9 Page 9 of 10



**Scrap Paper**



834560D4-5E39-4F7B-AB9A-BDE048C04D62

midterm1-b9ace

#9 Page 10 of 10



Michigan State University  
CMSE381 - Data Science

Feb 10, 2023  
Dr. Xie

## CMSE381 - Midterm #1

First name (please write as legibly as possible within the boxes)

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

Last name

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

NetID

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

*I will adhere to the Spartan Code of Honor in completing this assignment.*

Signed: \_\_\_\_\_

1. Do not open this test booklet until you are directed to do so.
2. You will have class time (2:40-4:00pm) to complete the exam.
3. This exam is closed book, but you can use the cheatsheet provided by the instructor.
4. Throughout the test, show your work so that your reasoning is clear. Otherwise no credit will be given. **BOX** your answers. Partial credit will be given where warranted.
5. Do not spend too much time on any one problem. Read them all through first and attack them in the order that allows you to make the most progress. Good luck :P



6A550365-FE92-41CF-B889-D1DB0AC3B44D

midterm1-b9ace

#10 Page 2 of 10

1. (15 points)

(a) Logistic regression is used for regression.

TRUE      FALSE

(b) Any model will never have training error below the irreducible error.

TRUE      FALSE

(c) Increasing your model flexibility always results in a better model.

TRUE      FALSE

(d) A logistic regression model is set up so that the odds are linear.

TRUE      FALSE

(e) Circle all of the following that would represent a qualitative variable.

Age      Year      Dog\_breed      Country\_of\_origin

Student\_(True/False)      Weight      Speed      MPG

(f) What equation would you use to evaluate the result of a regression model?

(g) What equation would you use to evaluate the result of a classification model?



2. (15 Points) I'm building a model to predict amount of a given brand of dog food eaten by a collection of dogs. I have 100 dogs eat this dog food, and I collect information on their height, weight, breed, and whether they live with another dog in the house.
- (a) List all input variables and specify whether they are quantitative or qualitative.
- (b) Say our dog breeds sampled are Huskies, Terriers, and Spaniels. Write down your prediction model.
- (c) We found that weight is a key predictor, and its relationship with the response is beyond linear. What extension can you come up with? Write down your updated model.



76D65B6B-69F5-454D-B68F-B4F461217F6E

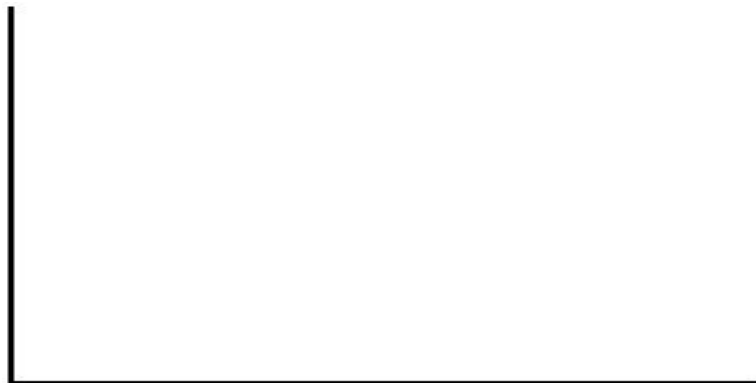
midterm1-b9ace

#10 Page 4 of 10

3. (15 points)

(a) What is bias-variance tradeoff? Explain the meaning of each term in the formula.

(b) Provide a sketch of typical (squared) bias, variance, training error, test error, and Bayes (or irreducible) error curves, on a single plot, as we go from less flexible statistical learning methods towards more flexible approaches. The  $x$ -axis should represent the amount of flexibility in the method, and the  $y$ -axis should represent the values for each curve. There should be five curves. Make sure to label each one.



(c) Explain why the (i) training error and (ii) testing error lines in your drawing have the shape displayed.



4. (10 points) The data for this example come from a study by Stamey et al. (1989). The data comes from a number of clinical measures in men who were about to receive a radical prostatectomy. The goal is to predict the log of PSA (**lpsa**) (a prostate-specific antigen measured in nanograms of PSA per milliliter of blood (ng/mL)) from a number of measurements. The variables are log cancer volume (**lcavol**), log prostate weight (**lweight**), age, log of the amount of benign prostatic hyperplasia (**lbph**), seminal vesicle invasion (**svi**), log of capsular penetration (**lcp**), Gleason score (**gleason**), and percent of Gleason scores 4 or 5 (**pgg45**).

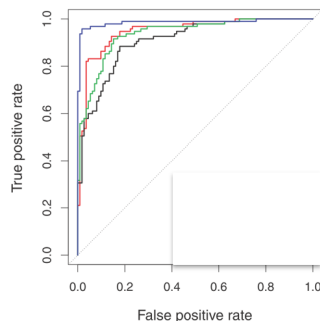
(a) Is this a case of regression or classification? Why?

(b) A linear model is fit to the data set, and the following table was returned.

Term	Coefficient	Std. Error	t-Score	p-value
Intercept	2.46	0.09	27.6	<0.00001
lcavol	0.68	0.13	5.37	<0.00001
lweight	0.26	0.1	2.75	0.00596
age	-0.14	0.1	-1.4	0.16153
lbph	0.21	0.1	2.06	0.039399
svi	0.31	0.12	2.47	0.013511
lcp	-0.29	0.15	-1.87	0.061484
gleason	-0.02	0.15	-0.15	0.880765
pgg45	0.27	0.15	1.74	0.081859

Are all the predictors useful? If yes, explain why. If not, explain what you would try to do next with the data.

5. (5 pts) We are trying to predict whether a patient will have a heart attack within one year. We have tried four different methods on a training dataset and have the following ROC curves. Which curve has the best performance on this training set? Justify your answer.







A9262E98-21A4-4879-BDFA-7FAC85962AC5

midterm1-b9ace

#10 Page 6 of 10

6. (15 Points) I get way too much email, so I decide to build a logistic regression model to predict whether a new incoming message is spam or not. I decide to just use a few variables:

$$\begin{aligned}X_1 &= \text{Number\_of\_sentences} \\X_2 &= \text{Number\_of\_references\_to\_a\_Nigerian\_Prince}\end{aligned}$$

and am training a logistic model to predict

$$Pr(Y = \text{spam} \mid X_1, X_2)$$

- (a) Write down the equation for the model you would train, using our standard notation with  $\beta_i$ 's.
- (b) If my trained model used  $\beta_0 = -13.1$ ,  $\beta_1 = 1.9$ , and  $\beta_2 = 6.1$ , what is the probability that a 5 sentence email with one reference to a Nigerian prince is spam? .
- (c) We know that if we miss an important email, it may cost a lot, How can you modify your model to avoid this?



7. (15 Points) Sparty generated 100 pair  $x$  and  $y$  via the following relationship:  $Y = 0.3 + 2x + x^2 + \epsilon$  but didn't tell you. You will try to find a good model to fit these data and more importantly to predict future response  $y$  when Sparty gives you another  $x$ . You will start with two models 1)  $Y = \beta_0 + \beta_1x$ ; 2)  $Y = \alpha_0 + \alpha_1x + \alpha_2x^2$ ; and 3)  $Y = \gamma_0 + \gamma_1x + \gamma_2x^2 + \gamma_3x^3$ .
- (a) If you use all 100 data to fit these three model and use the same 100 data to calculate MSE, which model will have the smallest MSE? Justify your answer
- (b) Since we focus on the ability to predict unseen data, you decide use validation set to evaluate the performance of the three models. There are two way to split the data: 1) 80 testing points and 20 training points or 2) 20 testing points and 80 training points. Which one will you choose? Justify your choice
- (c) From the in-class lab, we know validation set is not a good choice to choose model. Explain the drawback of validation set and what procedure will you use to choose the best model?



8545AE97-472F-4F24-A2FE-2C48A4C4492E

midterm1-b9ace

#10 Page 8 of 10

8. (10 Points) Suppose we have a data set with five predictors,  $X_1 = \text{GPA}$ ,  $X_2 = \text{IQ}$ ,  $X_3 = \text{Level}$  (1 for College and 0 for High School),  $X_4 = \text{Interaction between GPA and IQ}$ , and  $X_5 = \text{Interaction between GPA and Level}$ . The response is starting salary after graduation (in thousands of dollars). Suppose we use least squares to fit the model, and get  $\beta_0 = 50$ ,  $\beta_1 = 20$ ,  $\beta_2 = 0.07$ ,  $\beta_3 = 35$ ,  $\beta_4 = 0.01$ ,  $\beta_5 = -10$ .

(a) Predict the salary of a college graduate with IQ of 110 and a GPA of 4.0.

- (b) True or false: Since the coefficient for the GPA/IQ interaction term is very small, there is very little evidence of an interaction effect. Justify your answer.

F08F2532-E46B-44C5-B380-0C54F86AF247

midterm1-b9ace

#10 Page 9 of 10



**Scrap Paper**



D86475E5-CA82-4B79-BD31-BCED950D28F7

midterm1-b9ace

#10 Page 10 of 10



Michigan State University  
CMSE381 - Data Science

Feb 10, 2023  
Dr. Xie

## CMSE381 - Midterm #1

First name (please write as legibly as possible within the boxes)

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

Last name

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

NetID

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

*I will adhere to the Spartan Code of Honor in completing this assignment.*

Signed: \_\_\_\_\_

1. Do not open this test booklet until you are directed to do so.
2. You will have class time (2:40-4:00pm) to complete the exam.
3. This exam is closed book, but you can use the cheatsheet provided by the instructor.
4. Throughout the test, show your work so that your reasoning is clear. Otherwise no credit will be given. **BOX** your answers. Partial credit will be given where warranted.
5. Do not spend too much time on any one problem. Read them all through first and attack them in the order that allows you to make the most progress. Good luck :P



6F721904-C67B-42EF-998C-47FE0CB73841

midterm1-b9ace

#11 Page 2 of 10

1. (15 points)

(a) Logistic regression is used for regression.

TRUE      FALSE

(b) Any model will never have training error below the irreducible error.

TRUE      FALSE

(c) Increasing your model flexibility always results in a better model.

TRUE      FALSE

(d) A logistic regression model is set up so that the odds are linear.

TRUE      FALSE

(e) Circle all of the following that would represent a qualitative variable.

Age      Year      Dog\_breed      Country\_of\_origin

Student\_(True/False)      Weight      Speed      MPG

(f) What equation would you use to evaluate the result of a regression model?

(g) What equation would you use to evaluate the result of a classification model?



2. (15 Points) I'm building a model to predict amount of a given brand of dog food eaten by a collection of dogs. I have 100 dogs eat this dog food, and I collect information on their height, weight, breed, and whether they live with another dog in the house.
- (a) List all input variables and specify whether they are quantitative or qualitative.
  - (b) Say our dog breeds sampled are Huskies, Terriers, and Spaniels. Write down your prediction model.
  - (c) We found that weight is a key predictor, and its relationship with the response is beyond linear. What extension can you come up with? Write down your updated model.





D9D40294-438E-4F8C-A0DC-1BB442AB5818

midterm1-b9ace

#11 Page 4 of 10

3. (15 points)

(a) What is bias-variance tradeoff? Explain the meaning of each term in the formula.

(b) Provide a sketch of typical (squared) bias, variance, training error, test error, and Bayes (or irreducible) error curves, on a single plot, as we go from less flexible statistical learning methods towards more flexible approaches. The  $x$ -axis should represent the amount of flexibility in the method, and the  $y$ -axis should represent the values for each curve. There should be five curves. Make sure to label each one.



(c) Explain why the (i) training error and (ii) testing error lines in your drawing have the shape displayed.



4. (10 points) The data for this example come from a study by Stamey et al. (1989). The data comes from a number of clinical measures in men who were about to receive a radical prostatectomy. The goal is to predict the log of PSA (**lpsa**) (a prostate-specific antigen measured in nanograms of PSA per milliliter of blood (ng/mL)) from a number of measurements. The variables are log cancer volume (**lcavol**), log prostate weight (**lweight**), age, log of the amount of benign prostatic hyperplasia (**lbph**), seminal vesicle invasion (**svi**), log of capsular penetration (**lcp**), Gleason score (**gleason**), and percent of Gleason scores 4 or 5 (**pgg45**).

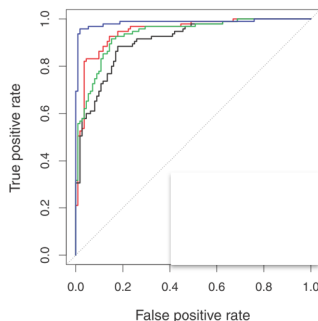
(a) Is this a case of regression or classification? Why?

(b) A linear model is fit to the data set, and the following table was returned.

Term	Coefficient	Std. Error	t-Score	p-value
Intercept	2.46	0.09	27.6	<0.00001
lcavol	0.68	0.13	5.37	<0.00001
lweight	0.26	0.1	2.75	0.00596
age	-0.14	0.1	-1.4	0.16153
lbph	0.21	0.1	2.06	0.039399
svi	0.31	0.12	2.47	0.013511
lcp	-0.29	0.15	-1.87	0.061484
gleason	-0.02	0.15	-0.15	0.880765
pgg45	0.27	0.15	1.74	0.081859

Are all the predictors useful? If yes, explain why. If not, explain what you would try to do next with the data.

5. (5 pts) We are trying to predict whether a patient will have a heart attach within one year. We have tried four different methods on a training dataset and have the following ROC curves Which curve has the best performance on this training set? Justify your answer.





42F376F4-AF20-4463-B935-C90ADC688A24

midterm1-b9ace

#11 Page 6 of 10

6. (15 Points) I get way too much email, so I decide to build a logistic regression model to predict whether a new incoming message is spam or not. I decide to just use a few variables:

$$\begin{aligned} X_1 &= \text{Number\_of\_sentences} \\ X_2 &= \text{Number\_of\_references\_to\_a\_Nigerian\_Prince} \end{aligned}$$

and am training a logistic model to predict

$$Pr(Y = \text{spam} \mid X_1, X_2)$$

- (a) Write down the equation for the model you would train, using our standard notation with  $\beta_i$ 's.
- (b) If my trained model used  $\beta_0 = -13.1$ ,  $\beta_1 = 1.9$ , and  $\beta_2 = 6.1$ , what is the probability that a 5 sentence email with one reference to a Nigerian prince is spam? .
- (c) We know that if we miss an important email, it may cost a lot, How can you modify your model to avoid this?



7. (15 Points) Sparty generated 100 pair  $x$  and  $y$  via the following relationship:  $Y = 0.3 + 2x + x^2 + \epsilon$  but didn't tell you. You will try to find a good model to fit these data and more importantly to predict future response  $y$  when Sparty gives you another  $x$ . You will start with two models 1)  $Y = \beta_0 + \beta_1 x$ ; 2)  $Y = \alpha_0 + \alpha_1 x + \alpha_2 x^2$ ; and 3)  $Y = \gamma_0 + \gamma_1 x + \gamma_2 x^2 + \gamma_3 x^3$ .
- (a) If you use all 100 data to fit these three model and use the same 100 data to calculate MSE, which model will have the smallest MSE? Justify your answer
- (b) Since we focus on the ability to predict unseen data, you decide use validation set to evaluate the performance of the three models. There are two way to split the data: 1) 80 testing points and 20 training points or 2) 20 testing points and 80 training points. Which one will you choose? Justify your choice
- (c) From the in-class lab, we know validation set is not a good choice to choose model. Explain the drawback of validation set and what procedure will you use to choose the best model?



8. (10 Points) Suppose we have a data set with five predictors,  $X_1 = \text{GPA}$ ,  $X_2 = \text{IQ}$ ,  $X_3 = \text{Level}$  (1 for College and 0 for High School),  $X_4 = \text{Interaction between GPA and IQ}$ , and  $X_5 = \text{Interaction between GPA and Level}$ . The response is starting salary after graduation (in thousands of dollars). Suppose we use least squares to fit the model, and get  $\beta_0 = 50$ ,  $\beta_1 = 20$ ,  $\beta_2 = 0.07$ ,  $\beta_3 = 35$ ,  $\beta_4 = 0.01$ ,  $\beta_5 = -10$ .

(a) Predict the salary of a college graduate with IQ of 110 and a GPA of 4.0.

- (b) True or false: Since the coefficient for the GPA/IQ interaction term is very small, there is very little evidence of an interaction effect. Justify your answer.

F63DD22D-3A57-46E3-BD5C-C44BF2966155

midterm1-b9ace

#11 Page 9 of 10



**Scrap Paper**



4B49B036-4029-4F7B-BA5F-6F2AFFC9C3FC

midterm1-b9ace

#11 Page 10 of 10



Michigan State University  
CMSE381 - Data Science

Feb 10, 2023  
Dr. Xie

## CMSE381 - Midterm #1

First name (please write as legibly as possible within the boxes)

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

Last name

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

NetID

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

*I will adhere to the Spartan Code of Honor in completing this assignment.*

Signed: \_\_\_\_\_

1. Do not open this test booklet until you are directed to do so.
2. You will have class time (2:40-4:00pm) to complete the exam.
3. This exam is closed book, but you can use the cheatsheet provided by the instructor.
4. Throughout the test, show your work so that your reasoning is clear. Otherwise no credit will be given. **BOX** your answers. Partial credit will be given where warranted.
5. Do not spend too much time on any one problem. Read them all through first and attack them in the order that allows you to make the most progress. Good luck :P





AA1014F6-D531-4B07-B497-D1B6D8E11494

midterm1-b9ace

#12 Page 2 of 10

1. (15 points)

(a) Logistic regression is used for regression.

TRUE      FALSE

(b) Any model will never have training error below the irreducible error.

TRUE      FALSE

(c) Increasing your model flexibility always results in a better model.

TRUE      FALSE

(d) A logistic regression model is set up so that the odds are linear.

TRUE      FALSE

(e) Circle all of the following that would represent a qualitative variable.

Age      Year      Dog\_breed      Country\_of\_origin

Student\_(True/False)      Weight      Speed      MPG

(f) What equation would you use to evaluate the result of a regression model?

(g) What equation would you use to evaluate the result of a classification model?



2. (15 Points) I'm building a model to predict amount of a given brand of dog food eaten by a collection of dogs. I have 100 dogs eat this dog food, and I collect information on their height, weight, breed, and whether they live with another dog in the house.
- (a) List all input variables and specify whether they are quantitative or qualitative.
- (b) Say our dog breeds sampled are Huskies, Terriers, and Spaniels. Write down your prediction model.
- (c) We found that weight is a key predictor, and its relationship with the response is beyond linear. What extension can you come up with? Write down your updated model.



8171BC5A-850B-49BD-8C84-7DECD7A760F0

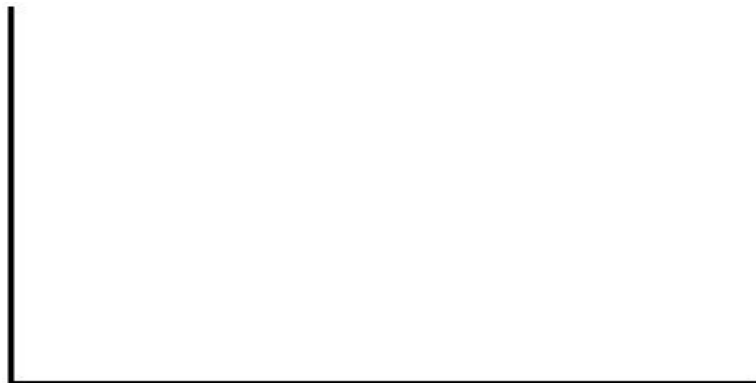
midterm1-b9ace

#12 Page 4 of 10

3. (15 points)

(a) What is bias-variance tradeoff? Explain the meaning of each term in the formula.

(b) Provide a sketch of typical (squared) bias, variance, training error, test error, and Bayes (or irreducible) error curves, on a single plot, as we go from less flexible statistical learning methods towards more flexible approaches. The  $x$ -axis should represent the amount of flexibility in the method, and the  $y$ -axis should represent the values for each curve. There should be five curves. Make sure to label each one.



(c) Explain why the (i) training error and (ii) testing error lines in your drawing have the shape displayed.



4. (10 points) The data for this example come from a study by Stamey et al. (1989). The data comes from a number of clinical measures in men who were about to receive a radical prostatectomy. The goal is to predict the log of PSA (**lpsa**) (a prostate-specific antigen measured in nanograms of PSA per milliliter of blood (ng/mL)) from a number of measurements. The variables are log cancer volume (**lcavol**), log prostate weight (**lweight**), age, log of the amount of benign prostatic hyperplasia (**lbph**), seminal vesicle invasion (**svi**), log of capsular penetration (**lcp**), Gleason score (**gleason**), and percent of Gleason scores 4 or 5 (**pgg45**).

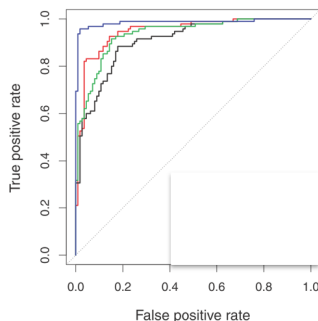
(a) Is this a case of regression or classification? Why?

(b) A linear model is fit to the data set, and the following table was returned.

Term	Coefficient	Std. Error	t-Score	p-value
Intercept	2.46	0.09	27.6	<0.00001
lcavol	0.68	0.13	5.37	<0.00001
lweight	0.26	0.1	2.75	0.00596
age	-0.14	0.1	-1.4	0.16153
lbph	0.21	0.1	2.06	0.039399
svi	0.31	0.12	2.47	0.013511
lcp	-0.29	0.15	-1.87	0.061484
gleason	-0.02	0.15	-0.15	0.880765
pgg45	0.27	0.15	1.74	0.081859

Are all the predictors useful? If yes, explain why. If not, explain what you would try to do next with the data.

5. (5 pts) We are trying to predict whether a patient will have a heart attack within one year. We have tried four different methods on a training dataset and have the following ROC curves. Which curve has the best performance on this training set? Justify your answer.





603B194F-A0BB-4322-8274-5C5B35E0FC50

midterm1-b9ace

#12 Page 6 of 10

6. (15 Points) I get way too much email, so I decide to build a logistic regression model to predict whether a new incoming message is spam or not. I decide to just use a few variables:

$$\begin{aligned}X_1 &= \text{Number\_of\_sentences} \\X_2 &= \text{Number\_of\_references\_to\_a\_Nigerian\_Prince}\end{aligned}$$

and am training a logistic model to predict

$$Pr(Y = \text{spam} \mid X_1, X_2)$$

- (a) Write down the equation for the model you would train, using our standard notation with  $\beta_i$ 's.
- (b) If my trained model used  $\beta_0 = -13.1$ ,  $\beta_1 = 1.9$ , and  $\beta_2 = 6.1$ , what is the probability that a 5 sentence email with one reference to a Nigerian prince is spam? .
- (c) We know that if we miss an important email, it may cost a lot, How can you modify your model to avoid this?



7. (15 Points) Sparty generated 100 pair  $x$  and  $y$  via the following relationship:  $Y = 0.3 + 2x + x^2 + \epsilon$  but didn't tell you. You will try to find a good model to fit these data and more importantly to predict future response  $y$  when Sparty gives you another  $x$ . You will start with two models 1)  $Y = \beta_0 + \beta_1x$ ; 2)  $Y = \alpha_0 + \alpha_1x + \alpha_2x^2$ ; and 3)  $Y = \gamma_0 + \gamma_1x + \gamma_2x^2 + \gamma_3x^3$ .
- (a) If you use all 100 data to fit these three model and use the same 100 data to calculate MSE, which model will have the smallest MSE? Justify your answer
- (b) Since we focus on the ability to predict unseen data, you decide use validation set to evaluate the performance of the three models. There are two way to split the data: 1) 80 testing points and 20 training points or 2) 20 testing points and 80 training points. Which one will you choose? Justify your choice
- (c) From the in-class lab, we know validation set is not a good choice to choose model. Explain the drawback of validation set and what procedure will you use to choose the best model?



8. (10 Points) Suppose we have a data set with five predictors,  $X_1 = \text{GPA}$ ,  $X_2 = \text{IQ}$ ,  $X_3 = \text{Level}$  (1 for College and 0 for High School),  $X_4 = \text{Interaction between GPA and IQ}$ , and  $X_5 = \text{Interaction between GPA and Level}$ . The response is starting salary after graduation (in thousands of dollars). Suppose we use least squares to fit the model, and get  $\beta_0 = 50$ ,  $\beta_1 = 20$ ,  $\beta_2 = 0.07$ ,  $\beta_3 = 35$ ,  $\beta_4 = 0.01$ ,  $\beta_5 = -10$ .

(a) Predict the salary of a college graduate with IQ of 110 and a GPA of 4.0.

(b) True or false: Since the coefficient for the GPA/IQ interaction term is very small, there is very little evidence of an interaction effect. Justify your answer.

F6E5108E-DBA2-4C03-A9EF-671734AE6E75

midterm1-b9ace

#12 Page 9 of 10



**Scrap Paper**





45E20D42-3C52-4F6D-AF32-25603E990570

midterm1-b9ace

#12 Page 10 of 10



Michigan State University  
CMSE381 - Data Science

Feb 10, 2023  
Dr. Xie

## CMSE381 - Midterm #1

First name (please write as legibly as possible within the boxes)

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

Last name

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

NetID

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

*I will adhere to the Spartan Code of Honor in completing this assignment.*

Signed: \_\_\_\_\_

1. Do not open this test booklet until you are directed to do so.
2. You will have class time (2:40-4:00pm) to complete the exam.
3. This exam is closed book, but you can use the cheatsheet provided by the instructor.
4. Throughout the test, show your work so that your reasoning is clear. Otherwise no credit will be given. **BOX** your answers. Partial credit will be given where warranted.
5. Do not spend too much time on any one problem. Read them all through first and attack them in the order that allows you to make the most progress. Good luck :P



71FF4F38-DE5B-449D-B74B-ACAB60B33EE9

midterm1-b9ace

#13 Page 2 of 10

1. (15 points)

(a) Logistic regression is used for regression.

TRUE      FALSE

(b) Any model will never have training error below the irreducible error.

TRUE      FALSE

(c) Increasing your model flexibility always results in a better model.

TRUE      FALSE

(d) A logistic regression model is set up so that the odds are linear.

TRUE      FALSE

(e) Circle all of the following that would represent a qualitative variable.

Age      Year      Dog\_breed      Country\_of\_origin

Student\_(True/False)      Weight      Speed      MPG

(f) What equation would you use to evaluate the result of a regression model?

(g) What equation would you use to evaluate the result of a classification model?



2. (15 Points) I'm building a model to predict amount of a given brand of dog food eaten by a collection of dogs. I have 100 dogs eat this dog food, and I collect information on their height, weight, breed, and whether they live with another dog in the house.
- (a) List all input variables and specify whether they are quantitative or qualitative.
  - (b) Say our dog breeds sampled are Huskies, Terriers, and Spaniels. Write down your prediction model.
  - (c) We found that weight is a key predictor, and its relationship with the response is beyond linear. What extension can you come up with? Write down your updated model.



452B3EB7-F278-44F5-B09E-F25008C56B46

midterm1-b9ace

#13 Page 4 of 10

3. (15 points)

(a) What is bias-variance tradeoff? Explain the meaning of each term in the formula.

(b) Provide a sketch of typical (squared) bias, variance, training error, test error, and Bayes (or irreducible) error curves, on a single plot, as we go from less flexible statistical learning methods towards more flexible approaches. The  $x$ -axis should represent the amount of flexibility in the method, and the  $y$ -axis should represent the values for each curve. There should be five curves. Make sure to label each one.



(c) Explain why the (i) training error and (ii) testing error lines in your drawing have the shape displayed.



4. (10 points) The data for this example come from a study by Stamey et al. (1989). The data comes from a number of clinical measures in men who were about to receive a radical prostatectomy. The goal is to predict the log of PSA (**lpsa**) (a prostate-specific antigen measured in nanograms of PSA per milliliter of blood (ng/mL)) from a number of measurements. The variables are log cancer volume (**lcavol**), log prostate weight (**lweight**), age, log of the amount of benign prostatic hyperplasia (**lbph**), seminal vesicle invasion (**svi**), log of capsular penetration (**lcp**), Gleason score (**gleason**), and percent of Gleason scores 4 or 5 (**pgg45**).

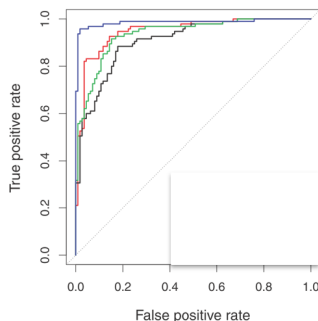
(a) Is this a case of regression or classification? Why?

(b) A linear model is fit to the data set, and the following table was returned.

Term	Coefficient	Std. Error	t-Score	p-value
Intercept	2.46	0.09	27.6	<0.00001
lcavol	0.68	0.13	5.37	<0.00001
lweight	0.26	0.1	2.75	0.00596
age	-0.14	0.1	-1.4	0.16153
lbph	0.21	0.1	2.06	0.039399
svi	0.31	0.12	2.47	0.013511
lcp	-0.29	0.15	-1.87	0.061484
gleason	-0.02	0.15	-0.15	0.880765
pgg45	0.27	0.15	1.74	0.081859

Are all the predictors useful? If yes, explain why. If not, explain what you would try to do next with the data.

5. (5 pts) We are trying to predict whether a patient will have a heart attack within one year. We have tried four different methods on a training dataset and have the following ROC curves. Which curve has the best performance on this training set? Justify your answer.





55B46C1E-E4B9-4F8E-BB35-2338CC294C43

midterm1-b9ace

#13 Page 6 of 10

6. (15 Points) I get way too much email, so I decide to build a logistic regression model to predict whether a new incoming message is spam or not. I decide to just use a few variables:

$$\begin{aligned}X_1 &= \text{Number\_of\_sentences} \\X_2 &= \text{Number\_of\_references\_to\_a\_Nigerian\_Prince}\end{aligned}$$

and am training a logistic model to predict

$$Pr(Y = \text{spam} \mid X_1, X_2)$$

- (a) Write down the equation for the model you would train, using our standard notation with  $\beta_i$ 's.
- (b) If my trained model used  $\beta_0 = -13.1$ ,  $\beta_1 = 1.9$ , and  $\beta_2 = 6.1$ , what is the probability that a 5 sentence email with one reference to a Nigerian prince is spam? .
- (c) We know that if we miss an important email, it may cost a lot, How can you modify your model to avoid this?



7. (15 Points) Sparty generated 100 pair  $x$  and  $y$  via the following relationship:  $Y = 0.3 + 2x + x^2 + \epsilon$  but didn't tell you. You will try to find a good model to fit these data and more importantly to predict future response  $y$  when Sparty gives you another  $x$ . You will start with two models 1)  $Y = \beta_0 + \beta_1x$ ; 2)  $Y = \alpha_0 + \alpha_1x + \alpha_2x^2$ ; and 3)  $Y = \gamma_0 + \gamma_1x + \gamma_2x^2 + \gamma_3x^3$ .
- (a) If you use all 100 data to fit these three model and use the same 100 data to calculate MSE, which model will have the smallest MSE? Justify your answer
- (b) Since we focus on the ability to predict unseen data, you decide use validation set to evaluate the performance of the three models. There are two way to split the data: 1) 80 testing points and 20 training points or 2) 20 testing points and 80 training points. Which one will you choose? Justify your choice
- (c) From the in-class lab, we know validation set is not a good choice to choose model. Explain the drawback of validation set and what procedure will you use to choose the best model?





F3ADB6E8-E415-4EFD-9C70-EC85D51A8D9C

midterm1-b9ace

#13 Page 8 of 10

8. (10 Points) Suppose we have a data set with five predictors,  $X_1 = \text{GPA}$ ,  $X_2 = \text{IQ}$ ,  $X_3 = \text{Level}$  (1 for College and 0 for High School),  $X_4 = \text{Interaction between GPA and IQ}$ , and  $X_5 = \text{Interaction between GPA and Level}$ . The response is starting salary after graduation (in thousands of dollars). Suppose we use least squares to fit the model, and get  $\beta_0 = 50$ ,  $\beta_1 = 20$ ,  $\beta_2 = 0.07$ ,  $\beta_3 = 35$ ,  $\beta_4 = 0.01$ ,  $\beta_5 = -10$ .

(a) Predict the salary of a college graduate with IQ of 110 and a GPA of 4.0.

(b) True or false: Since the coefficient for the GPA/IQ interaction term is very small, there is very little evidence of an interaction effect. Justify your answer.

A39CC251-AA79-4E79-B203-B9BA5138D678

midterm1-b9ace

#13 Page 9 of 10



**Scrap Paper**



4B95B56A-658F-4D14-B10A-7500BD70B4BB

midterm1-b9ace

#13 Page 10 of 10



Michigan State University  
CMSE381 - Data Science

Feb 10, 2023  
Dr. Xie

## CMSE381 - Midterm #1

First name (please write as legibly as possible within the boxes)

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

Last name

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

NetID

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

*I will adhere to the Spartan Code of Honor in completing this assignment.*

Signed: \_\_\_\_\_

1. Do not open this test booklet until you are directed to do so.
2. You will have class time (2:40-4:00pm) to complete the exam.
3. This exam is closed book, but you can use the cheatsheet provided by the instructor.
4. Throughout the test, show your work so that your reasoning is clear. Otherwise no credit will be given. **BOX** your answers. Partial credit will be given where warranted.
5. Do not spend too much time on any one problem. Read them all through first and attack them in the order that allows you to make the most progress. Good luck :P



D4FD4603-8F25-4BB1-8546-FEEAE8E0B2AD

midterm1-b9ace

#14 Page 2 of 10

1. (15 points)

(a) Logistic regression is used for regression.

TRUE      FALSE

(b) Any model will never have training error below the irreducible error.

TRUE      FALSE

(c) Increasing your model flexibility always results in a better model.

TRUE      FALSE

(d) A logistic regression model is set up so that the odds are linear.

TRUE      FALSE

(e) Circle all of the following that would represent a qualitative variable.

Age      Year      Dog\_breed      Country\_of\_origin

Student\_(True/False)      Weight      Speed      MPG

(f) What equation would you use to evaluate the result of a regression model?

(g) What equation would you use to evaluate the result of a classification model?



2. (15 Points) I'm building a model to predict amount of a given brand of dog food eaten by a collection of dogs. I have 100 dogs eat this dog food, and I collect information on their height, weight, breed, and whether they live with another dog in the house.
- (a) List all input variables and specify whether they are quantitative or qualitative.
- (b) Say our dog breeds sampled are Huskies, Terriers, and Spaniels. Write down your prediction model.
- (c) We found that weight is a key predictor, and its relationship with the response is beyond linear. What extension can you come up with? Write down your updated model.



3. (15 points)

(a) What is bias-variance tradeoff? Explain the meaning of each term in the formula.

(b) Provide a sketch of typical (squared) bias, variance, training error, test error, and Bayes (or irreducible) error curves, on a single plot, as we go from less flexible statistical learning methods towards more flexible approaches. The  $x$ -axis should represent the amount of flexibility in the method, and the  $y$ -axis should represent the values for each curve. There should be five curves. Make sure to label each one.



(c) Explain why the (i) training error and (ii) testing error lines in your drawing have the shape displayed.



4. (10 points) The data for this example come from a study by Stamey et al. (1989). The data comes from a number of clinical measures in men who were about to receive a radical prostatectomy. The goal is to predict the log of PSA (**lpsa**) (a prostate-specific antigen measured in nanograms of PSA per milliliter of blood (ng/mL)) from a number of measurements. The variables are log cancer volume (**lcavol**), log prostate weight (**lweight**), age, log of the amount of benign prostatic hyperplasia (**lbph**), seminal vesicle invasion (**svi**), log of capsular penetration (**lcp**), Gleason score (**gleason**), and percent of Gleason scores 4 or 5 (**pgg45**).

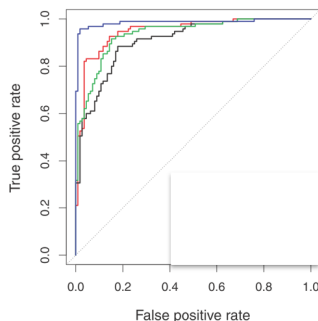
(a) Is this a case of regression or classification? Why?

(b) A linear model is fit to the data set, and the following table was returned.

Term	Coefficient	Std. Error	t-Score	p-value
Intercept	2.46	0.09	27.6	<0.00001
lcavol	0.68	0.13	5.37	<0.00001
lweight	0.26	0.1	2.75	0.00596
age	-0.14	0.1	-1.4	0.16153
lbph	0.21	0.1	2.06	0.039399
svi	0.31	0.12	2.47	0.013511
lcp	-0.29	0.15	-1.87	0.061484
gleason	-0.02	0.15	-0.15	0.880765
pgg45	0.27	0.15	1.74	0.081859

Are all the predictors useful? If yes, explain why. If not, explain what you would try to do next with the data.

5. (5 pts) We are trying to predict whether a patient will have a heart attack within one year. We have tried four different methods on a training dataset and have the following ROC curves. Which curve has the best performance on this training set? Justify your answer.







ABE38257-E940-47EF-9858-C3E85FBF43BC

midterm1-b9ace

#14 Page 6 of 10

6. (15 Points) I get way too much email, so I decide to build a logistic regression model to predict whether a new incoming message is spam or not. I decide to just use a few variables:

$$\begin{aligned}X_1 &= \text{Number\_of\_sentences} \\X_2 &= \text{Number\_of\_references\_to\_a\_Nigerian\_Prince}\end{aligned}$$

and am training a logistic model to predict

$$Pr(Y = \text{spam} \mid X_1, X_2)$$

- (a) Write down the equation for the model you would train, using our standard notation with  $\beta_i$ 's.
- (b) If my trained model used  $\beta_0 = -13.1$ ,  $\beta_1 = 1.9$ , and  $\beta_2 = 6.1$ , what is the probability that a 5 sentence email with one reference to a Nigerian prince is spam? .
- (c) We know that if we miss an important email, it may cost a lot, How can you modify your model to avoid this?



7. (15 Points) Sparty generated 100 pair  $x$  and  $y$  via the following relationship:  $Y = 0.3 + 2x + x^2 + \epsilon$  but didn't tell you. You will try to find a good model to fit these data and more importantly to predict future response  $y$  when Sparty gives you another  $x$ . You will start with two models 1)  $Y = \beta_0 + \beta_1x$ ; 2)  $Y = \alpha_0 + \alpha_1x + \alpha_2x^2$ ; and 3)  $Y = \gamma_0 + \gamma_1x + \gamma_2x^2 + \gamma_3x^3$ .
- (a) If you use all 100 data to fit these three model and use the same 100 data to calculate MSE, which model will have the smallest MSE? Justify your answer
- (b) Since we focus on the ability to predict unseen data, you decide use validation set to evaluate the performance of the three models. There are two way to split the data: 1) 80 testing points and 20 training points or 2) 20 testing points and 80 training points. Which one will you choose? Justify your choice
- (c) From the in-class lab, we know validation set is not a good choice to choose model. Explain the drawback of validation set and what procedure will you use to choose the best model?



F49F1BB6-001E-48EC-9D9E-A0E42CBE8AD5

midterm1-b9ace

#14 Page 8 of 10

8. (10 Points) Suppose we have a data set with five predictors,  $X_1 = \text{GPA}$ ,  $X_2 = \text{IQ}$ ,  $X_3 = \text{Level}$  (1 for College and 0 for High School),  $X_4 = \text{Interaction between GPA and IQ}$ , and  $X_5 = \text{Interaction between GPA and Level}$ . The response is starting salary after graduation (in thousands of dollars). Suppose we use least squares to fit the model, and get  $\beta_0 = 50$ ,  $\beta_1 = 20$ ,  $\beta_2 = 0.07$ ,  $\beta_3 = 35$ ,  $\beta_4 = 0.01$ ,  $\beta_5 = -10$ .

(a) Predict the salary of a college graduate with IQ of 110 and a GPA of 4.0.

(b) True or false: Since the coefficient for the GPA/IQ interaction term is very small, there is very little evidence of an interaction effect. Justify your answer.

61214AD1-4823-49D1-8118-2BC8162302A8

midterm1-b9ace

#14 Page 9 of 10



## Scrap Paper



2C419CE4-8A83-4CC0-B72B-970858DE5B3E

midterm1-b9ace

#14 Page 10 of 10



Michigan State University  
CMSE381 - Data Science

Feb 10, 2023  
Dr. Xie

## CMSE381 - Midterm #1

First name (please write as legibly as possible within the boxes)

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

Last name

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

NetID

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

*I will adhere to the Spartan Code of Honor in completing this assignment.*

Signed: \_\_\_\_\_

1. Do not open this test booklet until you are directed to do so.
2. You will have class time (2:40-4:00pm) to complete the exam.
3. This exam is closed book, but you can use the cheatsheet provided by the instructor.
4. Throughout the test, show your work so that your reasoning is clear. Otherwise no credit will be given. **BOX** your answers. Partial credit will be given where warranted.
5. Do not spend too much time on any one problem. Read them all through first and attack them in the order that allows you to make the most progress. Good luck :P



BD5105EE-F408-43F8-ADD0-A9D9A9E35A6F

midterm1-b9ace

#15 Page 2 of 10

1. (15 points)

(a) Logistic regression is used for regression.

TRUE      FALSE

(b) Any model will never have training error below the irreducible error.

TRUE      FALSE

(c) Increasing your model flexibility always results in a better model.

TRUE      FALSE

(d) A logistic regression model is set up so that the odds are linear.

TRUE      FALSE

(e) Circle all of the following that would represent a qualitative variable.

Age      Year      Dog\_breed      Country\_of\_origin

Student\_(True/False)      Weight      Speed      MPG

(f) What equation would you use to evaluate the result of a regression model?

(g) What equation would you use to evaluate the result of a classification model?



2. (15 Points) I'm building a model to predict amount of a given brand of dog food eaten by a collection of dogs. I have 100 dogs eat this dog food, and I collect information on their height, weight, breed, and whether they live with another dog in the house.
- (a) List all input variables and specify whether they are quantitative or qualitative.
  - (b) Say our dog breeds sampled are Huskies, Terriers, and Spaniels. Write down your prediction model.
  - (c) We found that weight is a key predictor, and its relationship with the response is beyond linear. What extension can you come up with? Write down your updated model.





1C862F12-9E89-45B9-9576-F53890D64EE5

midterm1-b9ace

#15 Page 4 of 10

3. (15 points)

(a) What is bias-variance tradeoff? Explain the meaning of each term in the formula.

(b) Provide a sketch of typical (squared) bias, variance, training error, test error, and Bayes (or irreducible) error curves, on a single plot, as we go from less flexible statistical learning methods towards more flexible approaches. The  $x$ -axis should represent the amount of flexibility in the method, and the  $y$ -axis should represent the values for each curve. There should be five curves. Make sure to label each one.



(c) Explain why the (i) training error and (ii) testing error lines in your drawing have the shape displayed.



4. (10 points) The data for this example come from a study by Stamey et al. (1989). The data comes from a number of clinical measures in men who were about to receive a radical prostatectomy. The goal is to predict the log of PSA (**lpsa**) (a prostate-specific antigen measured in nanograms of PSA per milliliter of blood (ng/mL)) from a number of measurements. The variables are log cancer volume (**lcavol**), log prostate weight (**lweight**), age, log of the amount of benign prostatic hyperplasia (**lbph**), seminal vesicle invasion (**svi**), log of capsular penetration (**lcp**), Gleason score (**gleason**), and percent of Gleason scores 4 or 5 (**pgg45**).

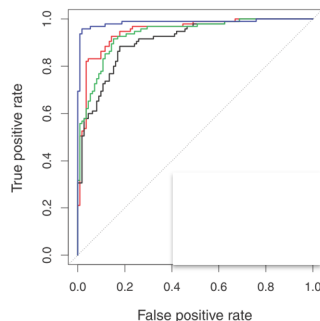
(a) Is this a case of regression or classification? Why?

(b) A linear model is fit to the data set, and the following table was returned.

Term	Coefficient	Std. Error	t-Score	p-value
Intercept	2.46	0.09	27.6	<0.00001
lcavol	0.68	0.13	5.37	<0.00001
lweight	0.26	0.1	2.75	0.00596
age	-0.14	0.1	-1.4	0.16153
lbph	0.21	0.1	2.06	0.039399
svi	0.31	0.12	2.47	0.013511
lcp	-0.29	0.15	-1.87	0.061484
gleason	-0.02	0.15	-0.15	0.880765
pgg45	0.27	0.15	1.74	0.081859

Are all the predictors useful? If yes, explain why. If not, explain what you would try to do next with the data.

5. (5 pts) We are trying to predict whether a patient will have a heart attack within one year. We have tried four different methods on a training dataset and have the following ROC curves. Which curve has the best performance on this training set? Justify your answer.





7F798E88-7A40-4F90-B68C-E5B902DE47BB

midterm1-b9ace

#15 Page 6 of 10

6. (15 Points) I get way too much email, so I decide to build a logistic regression model to predict whether a new incoming message is spam or not. I decide to just use a few variables:

$$\begin{aligned}X_1 &= \text{Number\_of\_sentences} \\X_2 &= \text{Number\_of\_references\_to\_a\_Nigerian\_Prince}\end{aligned}$$

and am training a logistic model to predict

$$Pr(Y = \text{spam} \mid X_1, X_2)$$

- (a) Write down the equation for the model you would train, using our standard notation with  $\beta_i$ 's.
- (b) If my trained model used  $\beta_0 = -13.1$ ,  $\beta_1 = 1.9$ , and  $\beta_2 = 6.1$ , what is the probability that a 5 sentence email with one reference to a Nigerian prince is spam? .
- (c) We know that if we miss an important email, it may cost a lot, How can you modify your model to avoid this?



7. (15 Points) Sparty generated 100 pair  $x$  and  $y$  via the following relationship:  $Y = 0.3 + 2x + x^2 + \epsilon$  but didn't tell you. You will try to find a good model to fit these data and more importantly to predict future response  $y$  when Sparty gives you another  $x$ . You will start with two models 1)  $Y = \beta_0 + \beta_1x$ ; 2)  $Y = \alpha_0 + \alpha_1x + \alpha_2x^2$ ; and 3)  $Y = \gamma_0 + \gamma_1x + \gamma_2x^2 + \gamma_3x^3$ .
- (a) If you use all 100 data to fit these three model and use the same 100 data to calculate MSE, which model will have the smallest MSE? Justify your answer
- (b) Since we focus on the ability to predict unseen data, you decide use validation set to evaluate the performance of the three models. There are two way to split the data: 1) 80 testing points and 20 training points or 2) 20 testing points and 80 training points. Which one will you choose? Justify your choice
- (c) From the in-class lab, we know validation set is not a good choice to choose model. Explain the drawback of validation set and what procedure will you use to choose the best model?



2CEF0D75-CAF9-4E45-98DB-F9072607F391

midterm1-b9ace

#15 Page 8 of 10

8. (10 Points) Suppose we have a data set with five predictors,  $X_1 = \text{GPA}$ ,  $X_2 = \text{IQ}$ ,  $X_3 = \text{Level}$  (1 for College and 0 for High School),  $X_4 = \text{Interaction between GPA and IQ}$ , and  $X_5 = \text{Interaction between GPA and Level}$ . The response is starting salary after graduation (in thousands of dollars). Suppose we use least squares to fit the model, and get  $\beta_0 = 50$ ,  $\beta_1 = 20$ ,  $\beta_2 = 0.07$ ,  $\beta_3 = 35$ ,  $\beta_4 = 0.01$ ,  $\beta_5 = -10$ .

(a) Predict the salary of a college graduate with IQ of 110 and a GPA of 4.0.

- (b) True or false: Since the coefficient for the GPA/IQ interaction term is very small, there is very little evidence of an interaction effect. Justify your answer.

F8563084-9500-44E1-B5ED-2FF2954356F8

midterm1-b9ace

#15 Page 9 of 10



**Scrap Paper**



D17FCD54-1D8B-4DE4-9810-82AF0F4A4ED6

midterm1-b9ace

#15 Page 10 of 10



Michigan State University  
CMSE381 - Data Science

Feb 10, 2023  
Dr. Xie

## CMSE381 - Midterm #1

First name (please write as legibly as possible within the boxes)

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

Last name

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

NetID

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

*I will adhere to the Spartan Code of Honor in completing this assignment.*

Signed: \_\_\_\_\_

1. Do not open this test booklet until you are directed to do so.
2. You will have class time (2:40-4:00pm) to complete the exam.
3. This exam is closed book, but you can use the cheatsheet provided by the instructor.
4. Throughout the test, show your work so that your reasoning is clear. Otherwise no credit will be given. **BOX** your answers. Partial credit will be given where warranted.
5. Do not spend too much time on any one problem. Read them all through first and attack them in the order that allows you to make the most progress. Good luck :P





AA1CADE6-AF45-47BB-BFB5-774D24304EF8

midterm1-b9ace

#16 Page 2 of 10

1. (15 points)

(a) Logistic regression is used for regression.

TRUE      FALSE

(b) Any model will never have training error below the irreducible error.

TRUE      FALSE

(c) Increasing your model flexibility always results in a better model.

TRUE      FALSE

(d) A logistic regression model is set up so that the odds are linear.

TRUE      FALSE

(e) Circle all of the following that would represent a qualitative variable.

Age      Year      Dog\_breed      Country\_of\_origin

Student\_(True/False)      Weight      Speed      MPG

(f) What equation would you use to evaluate the result of a regression model?

(g) What equation would you use to evaluate the result of a classification model?



2. (15 Points) I'm building a model to predict amount of a given brand of dog food eaten by a collection of dogs. I have 100 dogs eat this dog food, and I collect information on their height, weight, breed, and whether they live with another dog in the house.
- (a) List all input variables and specify whether they are quantitative or qualitative.
- (b) Say our dog breeds sampled are Huskies, Terriers, and Spaniels. Write down your prediction model.
- (c) We found that weight is a key predictor, and its relationship with the response is beyond linear. What extension can you come up with? Write down your updated model.



15664234-EF15-4A14-BB1E-B945EB0084DA

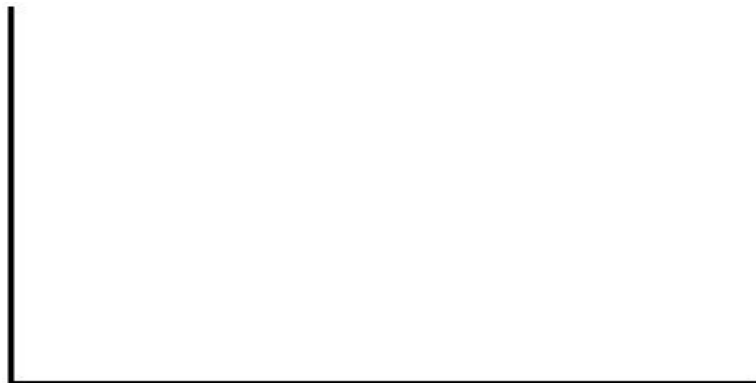
midterm1-b9ace

#16 Page 4 of 10

3. (15 points)

(a) What is bias-variance tradeoff? Explain the meaning of each term in the formula.

(b) Provide a sketch of typical (squared) bias, variance, training error, test error, and Bayes (or irreducible) error curves, on a single plot, as we go from less flexible statistical learning methods towards more flexible approaches. The  $x$ -axis should represent the amount of flexibility in the method, and the  $y$ -axis should represent the values for each curve. There should be five curves. Make sure to label each one.



(c) Explain why the (i) training error and (ii) testing error lines in your drawing have the shape displayed.



4. (10 points) The data for this example come from a study by Stamey et al. (1989). The data comes from a number of clinical measures in men who were about to receive a radical prostatectomy. The goal is to predict the log of PSA (**lpsa**) (a prostate-specific antigen measured in nanograms of PSA per milliliter of blood (ng/mL)) from a number of measurements. The variables are log cancer volume (**lcavol**), log prostate weight (**lweight**), age, log of the amount of benign prostatic hyperplasia (**lbph**), seminal vesicle invasion (**svi**), log of capsular penetration (**lcp**), Gleason score (**gleason**), and percent of Gleason scores 4 or 5 (**pgg45**).

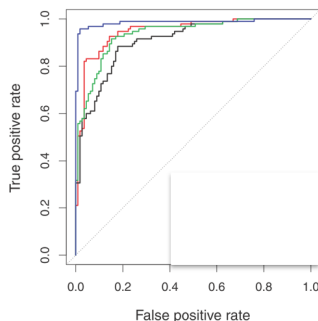
(a) Is this a case of regression or classification? Why?

(b) A linear model is fit to the data set, and the following table was returned.

Term	Coefficient	Std. Error	t-Score	p-value
Intercept	2.46	0.09	27.6	<0.00001
lcavol	0.68	0.13	5.37	<0.00001
lweight	0.26	0.1	2.75	0.00596
age	-0.14	0.1	-1.4	0.16153
lbph	0.21	0.1	2.06	0.039399
svi	0.31	0.12	2.47	0.013511
lcp	-0.29	0.15	-1.87	0.061484
gleason	-0.02	0.15	-0.15	0.880765
pgg45	0.27	0.15	1.74	0.081859

Are all the predictors useful? If yes, explain why. If not, explain what you would try to do next with the data.

5. (5 pts) We are trying to predict whether a patient will have a heart attack within one year. We have tried four different methods on a training dataset and have the following ROC curves. Which curve has the best performance on this training set? Justify your answer.





0DF963E4-C5BD-4B2C-89F4-282C73AF0957

midterm1-b9ace

#16 Page 6 of 10

6. (15 Points) I get way too much email, so I decide to build a logistic regression model to predict whether a new incoming message is spam or not. I decide to just use a few variables:

$$\begin{aligned} X_1 &= \text{Number\_of\_sentences} \\ X_2 &= \text{Number\_of\_references\_to\_a\_Nigerian\_Prince} \end{aligned}$$

and am training a logistic model to predict

$$Pr(Y = \text{spam} \mid X_1, X_2)$$

- (a) Write down the equation for the model you would train, using our standard notation with  $\beta_i$ 's.
- (b) If my trained model used  $\beta_0 = -13.1$ ,  $\beta_1 = 1.9$ , and  $\beta_2 = 6.1$ , what is the probability that a 5 sentence email with one reference to a Nigerian prince is spam? .
- (c) We know that if we miss an important email, it may cost a lot, How can you modify your model to avoid this?



7. (15 Points) Sparty generated 100 pair  $x$  and  $y$  via the following relationship:  $Y = 0.3 + 2x + x^2 + \epsilon$  but didn't tell you. You will try to find a good model to fit these data and more importantly to predict future response  $y$  when Sparty gives you another  $x$ . You will start with two models 1)  $Y = \beta_0 + \beta_1x$ ; 2)  $Y = \alpha_0 + \alpha_1x + \alpha_2x^2$ ; and 3)  $Y = \gamma_0 + \gamma_1x + \gamma_2x^2 + \gamma_3x^3$ .
- (a) If you use all 100 data to fit these three model and use the same 100 data to calculate MSE, which model will have the smallest MSE? Justify your answer
- (b) Since we focus on the ability to predict unseen data, you decide use validation set to evaluate the performance of the three models. There are two way to split the data: 1) 80 testing points and 20 training points or 2) 20 testing points and 80 training points. Which one will you choose? Justify your choice
- (c) From the in-class lab, we know validation set is not a good choice to choose model. Explain the drawback of validation set and what procedure will you use to choose the best model?



8. (10 Points) Suppose we have a data set with five predictors,  $X_1 = \text{GPA}$ ,  $X_2 = \text{IQ}$ ,  $X_3 = \text{Level}$  (1 for College and 0 for High School),  $X_4 = \text{Interaction between GPA and IQ}$ , and  $X_5 = \text{Interaction between GPA and Level}$ . The response is starting salary after graduation (in thousands of dollars). Suppose we use least squares to fit the model, and get  $\beta_0 = 50$ ,  $\beta_1 = 20$ ,  $\beta_2 = 0.07$ ,  $\beta_3 = 35$ ,  $\beta_4 = 0.01$ ,  $\beta_5 = -10$ .

(a) Predict the salary of a college graduate with IQ of 110 and a GPA of 4.0.

- (b) True or false: Since the coefficient for the GPA/IQ interaction term is very small, there is very little evidence of an interaction effect. Justify your answer.

41DD8A5A-37FF-41F7-83C4-946DB23F8A51

midterm1-b9ace

#16 Page 9 of 10



**Scrap Paper**





ACAC650F-7717-4E12-9014-F31DD541F6B7

midterm1-b9ace

#16 Page 10 of 10



Michigan State University  
CMSE381 - Data Science

Feb 10, 2023  
Dr. Xie

## CMSE381 - Midterm #1

First name (please write as legibly as possible within the boxes)

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

Last name

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

NetID

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

*I will adhere to the Spartan Code of Honor in completing this assignment.*

Signed: \_\_\_\_\_

1. Do not open this test booklet until you are directed to do so.
2. You will have class time (2:40-4:00pm) to complete the exam.
3. This exam is closed book, but you can use the cheatsheet provided by the instructor.
4. Throughout the test, show your work so that your reasoning is clear. Otherwise no credit will be given. **BOX** your answers. Partial credit will be given where warranted.
5. Do not spend too much time on any one problem. Read them all through first and attack them in the order that allows you to make the most progress. Good luck :P



1. (15 points)

(a) Logistic regression is used for regression.

TRUE      FALSE

(b) Any model will never have training error below the irreducible error.

TRUE      FALSE

(c) Increasing your model flexibility always results in a better model.

TRUE      FALSE

(d) A logistic regression model is set up so that the odds are linear.

TRUE      FALSE

(e) Circle all of the following that would represent a qualitative variable.

Age      Year      Dog\_breed      Country\_of\_origin

Student\_(True/False)      Weight      Speed      MPG

(f) What equation would you use to evaluate the result of a regression model?

(g) What equation would you use to evaluate the result of a classification model?



2. (15 Points) I'm building a model to predict amount of a given brand of dog food eaten by a collection of dogs. I have 100 dogs eat this dog food, and I collect information on their height, weight, breed, and whether they live with another dog in the house.
- (a) List all input variables and specify whether they are quantitative or qualitative.
- (b) Say our dog breeds sampled are Huskies, Terriers, and Spaniels. Write down your prediction model.
- (c) We found that weight is a key predictor, and its relationship with the response is beyond linear. What extension can you come up with? Write down your updated model.



3EBA03BE-34B9-4AC9-BB71-231A3386E326

midterm1-b9ace

#17 Page 4 of 10

3. (15 points)

(a) What is bias-variance tradeoff? Explain the meaning of each term in the formula.

(b) Provide a sketch of typical (squared) bias, variance, training error, test error, and Bayes (or irreducible) error curves, on a single plot, as we go from less flexible statistical learning methods towards more flexible approaches. The  $x$ -axis should represent the amount of flexibility in the method, and the  $y$ -axis should represent the values for each curve. There should be five curves. Make sure to label each one.



(c) Explain why the (i) training error and (ii) testing error lines in your drawing have the shape displayed.



4. (10 points) The data for this example come from a study by Stamey et al. (1989). The data comes from a number of clinical measures in men who were about to receive a radical prostatectomy. The goal is to predict the log of PSA (**lpsa**) (a prostate-specific antigen measured in nanograms of PSA per milliliter of blood (ng/mL)) from a number of measurements. The variables are log cancer volume (**lcavol**), log prostate weight (**lweight**), age, log of the amount of benign prostatic hyperplasia (**lbph**), seminal vesicle invasion (**svi**), log of capsular penetration (**lcp**), Gleason score (**gleason**), and percent of Gleason scores 4 or 5 (**pgg45**).

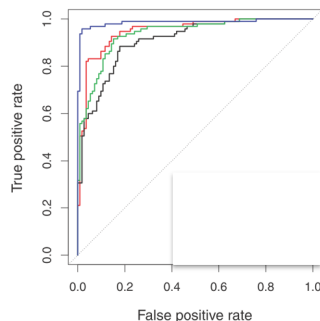
(a) Is this a case of regression or classification? Why?

(b) A linear model is fit to the data set, and the following table was returned.

Term	Coefficient	Std. Error	t-Score	p-value
Intercept	2.46	0.09	27.6	<0.00001
lcavol	0.68	0.13	5.37	<0.00001
lweight	0.26	0.1	2.75	0.00596
age	-0.14	0.1	-1.4	0.16153
lbph	0.21	0.1	2.06	0.039399
svi	0.31	0.12	2.47	0.013511
lcp	-0.29	0.15	-1.87	0.061484
gleason	-0.02	0.15	-0.15	0.880765
pgg45	0.27	0.15	1.74	0.081859

Are all the predictors useful? If yes, explain why. If not, explain what you would try to do next with the data.

5. (5 pts) We are trying to predict whether a patient will have a heart attack within one year. We have tried four different methods on a training dataset and have the following ROC curves. Which curve has the best performance on this training set? Justify your answer.





6. (15 Points) I get way too much email, so I decide to build a logistic regression model to predict whether a new incoming message is spam or not. I decide to just use a few variables:

$$\begin{aligned}X_1 &= \text{Number\_of\_sentences} \\X_2 &= \text{Number\_of\_references\_to\_a\_Nigerian\_Prince}\end{aligned}$$

and am training a logistic model to predict

$$Pr(Y = \text{spam} \mid X_1, X_2)$$

- (a) Write down the equation for the model you would train, using our standard notation with  $\beta_i$ 's.
- (b) If my trained model used  $\beta_0 = -13.1$ ,  $\beta_1 = 1.9$ , and  $\beta_2 = 6.1$ , what is the probability that a 5 sentence email with one reference to a Nigerian prince is spam? .
- (c) We know that if we miss an important email, it may cost a lot, How can you modify your model to avoid this?



7. (15 Points) Sparty generated 100 pair  $x$  and  $y$  via the following relationship:  $Y = 0.3 + 2x + x^2 + \epsilon$  but didn't tell you. You will try to find a good model to fit these data and more importantly to predict future response  $y$  when Sparty gives you another  $x$ . You will start with two models 1)  $Y = \beta_0 + \beta_1x$ ; 2)  $Y = \alpha_0 + \alpha_1x + \alpha_2x^2$ ; and 3)  $Y = \gamma_0 + \gamma_1x + \gamma_2x^2 + \gamma_3x^3$ .
- (a) If you use all 100 data to fit these three model and use the same 100 data to calculate MSE, which model will have the smallest MSE? Justify your answer
- (b) Since we focus on the ability to predict unseen data, you decide use validation set to evaluate the performance of the three models. There are two way to split the data: 1) 80 testing points and 20 training points or 2) 20 testing points and 80 training points. Which one will you choose? Justify your choice
- (c) From the in-class lab, we know validation set is not a good choice to choose model. Explain the drawback of validation set and what procedure will you use to choose the best model?





8. (10 Points) Suppose we have a data set with five predictors,  $X_1 = \text{GPA}$ ,  $X_2 = \text{IQ}$ ,  $X_3 = \text{Level}$  (1 for College and 0 for High School),  $X_4 = \text{Interaction between GPA and IQ}$ , and  $X_5 = \text{Interaction between GPA and Level}$ . The response is starting salary after graduation (in thousands of dollars). Suppose we use least squares to fit the model, and get  $\beta_0 = 50$ ,  $\beta_1 = 20$ ,  $\beta_2 = 0.07$ ,  $\beta_3 = 35$ ,  $\beta_4 = 0.01$ ,  $\beta_5 = -10$ .

(a) Predict the salary of a college graduate with IQ of 110 and a GPA of 4.0.

- (b) True or false: Since the coefficient for the GPA/IQ interaction term is very small, there is very little evidence of an interaction effect. Justify your answer.

EB1D741A-4E87-4A35-93B1-FD23ED1CDBE5

midterm1-b9ace

#17 Page 9 of 10



## Scrap Paper



659AF04F-4218-4071-9373-228CBA4A4A38

midterm1-b9ace

#17 Page 10 of 10



Michigan State University  
CMSE381 - Data Science

Feb 10, 2023  
Dr. Xie

## CMSE381 - Midterm #1

First name (please write as legibly as possible within the boxes)

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

Last name

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

NetID

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

*I will adhere to the Spartan Code of Honor in completing this assignment.*

Signed: \_\_\_\_\_

1. Do not open this test booklet until you are directed to do so.
2. You will have class time (2:40-4:00pm) to complete the exam.
3. This exam is closed book, but you can use the cheatsheet provided by the instructor.
4. Throughout the test, show your work so that your reasoning is clear. Otherwise no credit will be given. **BOX** your answers. Partial credit will be given where warranted.
5. Do not spend too much time on any one problem. Read them all through first and attack them in the order that allows you to make the most progress. Good luck :P



A3BA673C-2B29-43B7-A2BD-4222FDC59479

midterm1-b9ace

#18 Page 2 of 10

1. (15 points)

(a) Logistic regression is used for regression.

TRUE      FALSE

(b) Any model will never have training error below the irreducible error.

TRUE      FALSE

(c) Increasing your model flexibility always results in a better model.

TRUE      FALSE

(d) A logistic regression model is set up so that the odds are linear.

TRUE      FALSE

(e) Circle all of the following that would represent a qualitative variable.

Age      Year      Dog\_breed      Country\_of\_origin

Student\_(True/False)      Weight      Speed      MPG

(f) What equation would you use to evaluate the result of a regression model?

(g) What equation would you use to evaluate the result of a classification model?



2. (15 Points) I'm building a model to predict amount of a given brand of dog food eaten by a collection of dogs. I have 100 dogs eat this dog food, and I collect information on their height, weight, breed, and whether they live with another dog in the house.
- (a) List all input variables and specify whether they are quantitative or qualitative.
- (b) Say our dog breeds sampled are Huskies, Terriers, and Spaniels. Write down your prediction model.
- (c) We found that weight is a key predictor, and its relationship with the response is beyond linear. What extension can you come up with? Write down your updated model.



E416883B-CA79-4D84-A0F9-C7DE4C5B1DEE

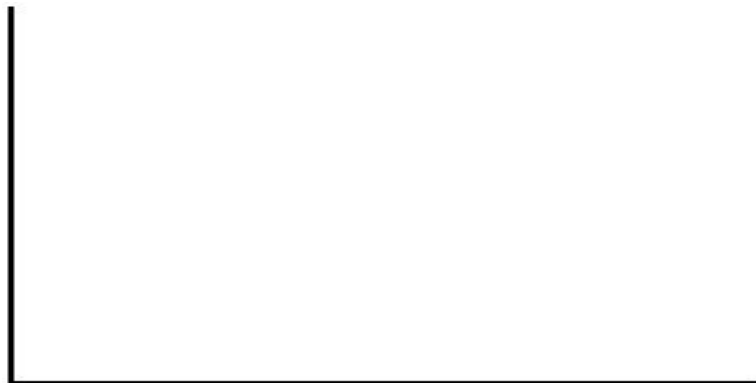
midterm1-b9ace

#18 Page 4 of 10

3. (15 points)

(a) What is bias-variance tradeoff? Explain the meaning of each term in the formula.

(b) Provide a sketch of typical (squared) bias, variance, training error, test error, and Bayes (or irreducible) error curves, on a single plot, as we go from less flexible statistical learning methods towards more flexible approaches. The  $x$ -axis should represent the amount of flexibility in the method, and the  $y$ -axis should represent the values for each curve. There should be five curves. Make sure to label each one.



(c) Explain why the (i) training error and (ii) testing error lines in your drawing have the shape displayed.



4. (10 points) The data for this example come from a study by Stamey et al. (1989). The data comes from a number of clinical measures in men who were about to receive a radical prostatectomy. The goal is to predict the log of PSA (**lpsa**) (a prostate-specific antigen measured in nanograms of PSA per milliliter of blood (ng/mL)) from a number of measurements. The variables are log cancer volume (**lcavol**), log prostate weight (**lweight**), age, log of the amount of benign prostatic hyperplasia (**lbph**), seminal vesicle invasion (**svi**), log of capsular penetration (**lcp**), Gleason score (**gleason**), and percent of Gleason scores 4 or 5 (**pgg45**).

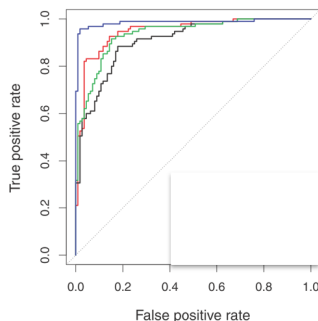
(a) Is this a case of regression or classification? Why?

(b) A linear model is fit to the data set, and the following table was returned.

Term	Coefficient	Std. Error	t-Score	p-value
Intercept	2.46	0.09	27.6	<0.00001
lcavol	0.68	0.13	5.37	<0.00001
lweight	0.26	0.1	2.75	0.00596
age	-0.14	0.1	-1.4	0.16153
lbph	0.21	0.1	2.06	0.039399
svi	0.31	0.12	2.47	0.013511
lcp	-0.29	0.15	-1.87	0.061484
gleason	-0.02	0.15	-0.15	0.880765
pgg45	0.27	0.15	1.74	0.081859

Are all the predictors useful? If yes, explain why. If not, explain what you would try to do next with the data.

5. (5 pts) We are trying to predict whether a patient will have a heart attack within one year. We have tried four different methods on a training dataset and have the following ROC curves. Which curve has the best performance on this training set? Justify your answer.







6FEE1452-0C07-449B-9635-9C6271512A8D

midterm1-b9ace

#18 Page 6 of 10

6. (15 Points) I get way too much email, so I decide to build a logistic regression model to predict whether a new incoming message is spam or not. I decide to just use a few variables:

$$\begin{aligned}X_1 &= \text{Number\_of\_sentences} \\X_2 &= \text{Number\_of\_references\_to\_a\_Nigerian\_Prince}\end{aligned}$$

and am training a logistic model to predict

$$Pr(Y = \text{spam} \mid X_1, X_2)$$

- (a) Write down the equation for the model you would train, using our standard notation with  $\beta_i$ 's.
- (b) If my trained model used  $\beta_0 = -13.1$ ,  $\beta_1 = 1.9$ , and  $\beta_2 = 6.1$ , what is the probability that a 5 sentence email with one reference to a Nigerian prince is spam? .
- (c) We know that if we miss an important email, it may cost a lot, How can you modify your model to avoid this?



7. (15 Points) Sparty generated 100 pair  $x$  and  $y$  via the following relationship:  $Y = 0.3 + 2x + x^2 + \epsilon$  but didn't tell you. You will try to find a good model to fit these data and more importantly to predict future response  $y$  when Sparty gives you another  $x$ . You will start with two models 1)  $Y = \beta_0 + \beta_1 x$ ; 2)  $Y = \alpha_0 + \alpha_1 x + \alpha_2 x^2$ ; and 3)  $Y = \gamma_0 + \gamma_1 x + \gamma_2 x^2 + \gamma_3 x^3$ .
- (a) If you use all 100 data to fit these three model and use the same 100 data to calculate MSE, which model will have the smallest MSE? Justify your answer
- (b) Since we focus on the ability to predict unseen data, you decide use validation set to evaluate the performance of the three models. There are two way to split the data: 1) 80 testing points and 20 training points or 2) 20 testing points and 80 training points. Which one will you choose? Justify your choice
- (c) From the in-class lab, we know validation set is not a good choice to choose model. Explain the drawback of validation set and what procedure will you use to choose the best model?



4EF3F22E-4AFC-4D4F-8F8D-464DB6C63EB8

midterm1-b9ace

#18 Page 8 of 10

8. (10 Points) Suppose we have a data set with five predictors,  $X_1 = \text{GPA}$ ,  $X_2 = \text{IQ}$ ,  $X_3 = \text{Level}$  (1 for College and 0 for High School),  $X_4 = \text{Interaction between GPA and IQ}$ , and  $X_5 = \text{Interaction between GPA and Level}$ . The response is starting salary after graduation (in thousands of dollars). Suppose we use least squares to fit the model, and get  $\beta_0 = 50$ ,  $\beta_1 = 20$ ,  $\beta_2 = 0.07$ ,  $\beta_3 = 35$ ,  $\beta_4 = 0.01$ ,  $\beta_5 = -10$ .

(a) Predict the salary of a college graduate with IQ of 110 and a GPA of 4.0.

- (b) True or false: Since the coefficient for the GPA/IQ interaction term is very small, there is very little evidence of an interaction effect. Justify your answer.

8D04B1D7-FD3F-4F9E-B38F-963B26632F8C

midterm1-b9ace

#18 Page 9 of 10



**Scrap Paper**



13EB73AD-32A1-4B2F-AF92-BAD7D07ABB02

midterm1-b9ace

#18 Page 10 of 10



Michigan State University  
CMSE381 - Data Science

Feb 10, 2023  
Dr. Xie

## CMSE381 - Midterm #1

First name (please write as legibly as possible within the boxes)

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

Last name

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

NetID

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

*I will adhere to the Spartan Code of Honor in completing this assignment.*

Signed: \_\_\_\_\_

1. Do not open this test booklet until you are directed to do so.
2. You will have class time (2:40-4:00pm) to complete the exam.
3. This exam is closed book, but you can use the cheatsheet provided by the instructor.
4. Throughout the test, show your work so that your reasoning is clear. Otherwise no credit will be given. **BOX** your answers. Partial credit will be given where warranted.
5. Do not spend too much time on any one problem. Read them all through first and attack them in the order that allows you to make the most progress. Good luck :P



89A330FD-25EB-466B-8D0C-D9F5DF388296

midterm1-b9ace

#19 Page 2 of 10

1. (15 points)

(a) Logistic regression is used for regression.

TRUE      FALSE

(b) Any model will never have training error below the irreducible error.

TRUE      FALSE

(c) Increasing your model flexibility always results in a better model.

TRUE      FALSE

(d) A logistic regression model is set up so that the odds are linear.

TRUE      FALSE

(e) Circle all of the following that would represent a qualitative variable.

Age      Year      Dog\_breed      Country\_of\_origin

Student\_(True/False)      Weight      Speed      MPG

(f) What equation would you use to evaluate the result of a regression model?

(g) What equation would you use to evaluate the result of a classification model?



2. (15 Points) I'm building a model to predict amount of a given brand of dog food eaten by a collection of dogs. I have 100 dogs eat this dog food, and I collect information on their height, weight, breed, and whether they live with another dog in the house.
- (a) List all input variables and specify whether they are quantitative or qualitative.
- (b) Say our dog breeds sampled are Huskies, Terriers, and Spaniels. Write down your prediction model.
- (c) We found that weight is a key predictor, and its relationship with the response is beyond linear. What extension can you come up with? Write down your updated model.





5216A884-6AA9-4A9C-9FD9-30FC2CA9B80A

midterm1-b9ace

#19 Page 4 of 10

3. (15 points)

(a) What is bias-variance tradeoff? Explain the meaning of each term in the formula.

(b) Provide a sketch of typical (squared) bias, variance, training error, test error, and Bayes (or irreducible) error curves, on a single plot, as we go from less flexible statistical learning methods towards more flexible approaches. The  $x$ -axis should represent the amount of flexibility in the method, and the  $y$ -axis should represent the values for each curve. There should be five curves. Make sure to label each one.



(c) Explain why the (i) training error and (ii) testing error lines in your drawing have the shape displayed.



4. (10 points) The data for this example come from a study by Stamey et al. (1989). The data comes from a number of clinical measures in men who were about to receive a radical prostatectomy. The goal is to predict the log of PSA (**lpsa**) (a prostate-specific antigen measured in nanograms of PSA per milliliter of blood (ng/mL)) from a number of measurements. The variables are log cancer volume (**lcavol**), log prostate weight (**lweight**), age, log of the amount of benign prostatic hyperplasia (**lbph**), seminal vesicle invasion (**svi**), log of capsular penetration (**lcp**), Gleason score (**gleason**), and percent of Gleason scores 4 or 5 (**pgg45**).

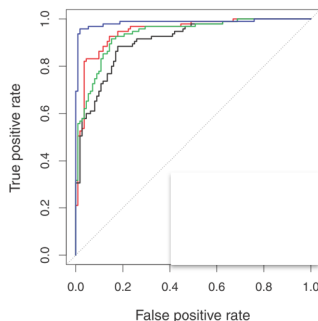
(a) Is this a case of regression or classification? Why?

(b) A linear model is fit to the data set, and the following table was returned.

Term	Coefficient	Std. Error	t-Score	p-value
Intercept	2.46	0.09	27.6	<0.00001
lcavol	0.68	0.13	5.37	<0.00001
lweight	0.26	0.1	2.75	0.00596
age	-0.14	0.1	-1.4	0.16153
lbph	0.21	0.1	2.06	0.039399
svi	0.31	0.12	2.47	0.013511
lcp	-0.29	0.15	-1.87	0.061484
gleason	-0.02	0.15	-0.15	0.880765
pgg45	0.27	0.15	1.74	0.081859

Are all the predictors useful? If yes, explain why. If not, explain what you would try to do next with the data.

5. (5 pts) We are trying to predict whether a patient will have a heart attack within one year. We have tried four different methods on a training dataset and have the following ROC curves. Which curve has the best performance on this training set? Justify your answer.





A5B0DB85-F388-4FF0-8B26-8E2CF0C2734A

midterm1-b9ace

#19 Page 6 of 10

6. (15 Points) I get way too much email, so I decide to build a logistic regression model to predict whether a new incoming message is spam or not. I decide to just use a few variables:

$$\begin{aligned}X_1 &= \text{Number\_of\_sentences} \\X_2 &= \text{Number\_of\_references\_to\_a\_Nigerian\_Prince}\end{aligned}$$

and am training a logistic model to predict

$$Pr(Y = \text{spam} \mid X_1, X_2)$$

- (a) Write down the equation for the model you would train, using our standard notation with  $\beta_i$ 's.
- (b) If my trained model used  $\beta_0 = -13.1$ ,  $\beta_1 = 1.9$ , and  $\beta_2 = 6.1$ , what is the probability that a 5 sentence email with one reference to a Nigerian prince is spam? .
- (c) We know that if we miss an important email, it may cost a lot, How can you modify your model to avoid this?



7. (15 Points) Sparty generated 100 pair  $x$  and  $y$  via the following relationship:  $Y = 0.3 + 2x + x^2 + \epsilon$  but didn't tell you. You will try to find a good model to fit these data and more importantly to predict future response  $y$  when Sparty gives you another  $x$ . You will start with two models 1)  $Y = \beta_0 + \beta_1x$ ; 2)  $Y = \alpha_0 + \alpha_1x + \alpha_2x^2$ ; and 3)  $Y = \gamma_0 + \gamma_1x + \gamma_2x^2 + \gamma_3x^3$ .
- (a) If you use all 100 data to fit these three model and use the same 100 data to calculate MSE, which model will have the smallest MSE? Justify your answer
- (b) Since we focus on the ability to predict unseen data, you decide use validation set to evaluate the performance of the three models. There are two way to split the data: 1) 80 testing points and 20 training points or 2) 20 testing points and 80 training points. Which one will you choose? Justify your choice
- (c) From the in-class lab, we know validation set is not a good choice to choose model. Explain the drawback of validation set and what procedure will you use to choose the best model?



8. (10 Points) Suppose we have a data set with five predictors,  $X_1 = \text{GPA}$ ,  $X_2 = \text{IQ}$ ,  $X_3 = \text{Level}$  (1 for College and 0 for High School),  $X_4 = \text{Interaction between GPA and IQ}$ , and  $X_5 = \text{Interaction between GPA and Level}$ . The response is starting salary after graduation (in thousands of dollars). Suppose we use least squares to fit the model, and get  $\beta_0 = 50$ ,  $\beta_1 = 20$ ,  $\beta_2 = 0.07$ ,  $\beta_3 = 35$ ,  $\beta_4 = 0.01$ ,  $\beta_5 = -10$ .

(a) Predict the salary of a college graduate with IQ of 110 and a GPA of 4.0.

- (b) True or false: Since the coefficient for the GPA/IQ interaction term is very small, there is very little evidence of an interaction effect. Justify your answer.

EBEF78F2-F2CE-4C94-9700-6F6BEE0B7956

midterm1-b9ace

#19 Page 9 of 10



## Scrap Paper



4FBAFD7E-2C14-4890-B7CE-AE7D9774A11D

midterm1-b9ace

#19 Page 10 of 10



Michigan State University  
CMSE381 - Data Science

Feb 10, 2023  
Dr. Xie

## CMSE381 - Midterm #1

First name (please write as legibly as possible within the boxes)

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

Last name

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

NetID

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

*I will adhere to the Spartan Code of Honor in completing this assignment.*

Signed: \_\_\_\_\_

1. Do not open this test booklet until you are directed to do so.
2. You will have class time (2:40-4:00pm) to complete the exam.
3. This exam is closed book, but you can use the cheatsheet provided by the instructor.
4. Throughout the test, show your work so that your reasoning is clear. Otherwise no credit will be given. **BOX** your answers. Partial credit will be given where warranted.
5. Do not spend too much time on any one problem. Read them all through first and attack them in the order that allows you to make the most progress. Good luck :P





1. (15 points)

(a) Logistic regression is used for regression.

TRUE      FALSE

(b) Any model will never have training error below the irreducible error.

TRUE      FALSE

(c) Increasing your model flexibility always results in a better model.

TRUE      FALSE

(d) A logistic regression model is set up so that the odds are linear.

TRUE      FALSE

(e) Circle all of the following that would represent a qualitative variable.

Age      Year      Dog\_breed      Country\_of\_origin

Student\_(True/False)      Weight      Speed      MPG

(f) What equation would you use to evaluate the result of a regression model?

(g) What equation would you use to evaluate the result of a classification model?



2. (15 Points) I'm building a model to predict amount of a given brand of dog food eaten by a collection of dogs. I have 100 dogs eat this dog food, and I collect information on their height, weight, breed, and whether they live with another dog in the house.
- (a) List all input variables and specify whether they are quantitative or qualitative.
  - (b) Say our dog breeds sampled are Huskies, Terriers, and Spaniels. Write down your prediction model.
  - (c) We found that weight is a key predictor, and its relationship with the response is beyond linear. What extension can you come up with? Write down your updated model.



1EF400A6-A328-43E8-B65A-58CE445F93F4

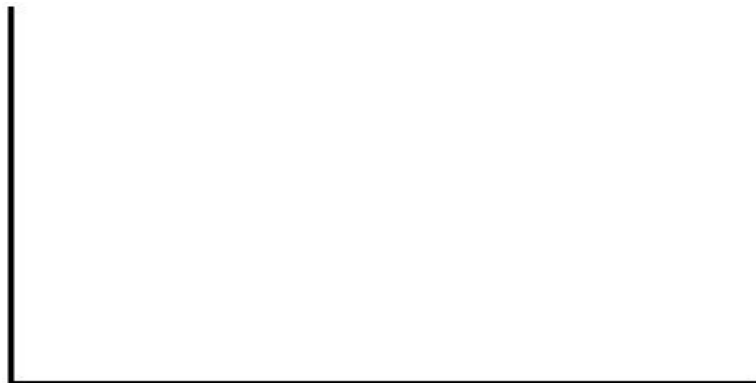
midterm1-b9ace

#20 Page 4 of 10

3. (15 points)

(a) What is bias-variance tradeoff? Explain the meaning of each term in the formula.

(b) Provide a sketch of typical (squared) bias, variance, training error, test error, and Bayes (or irreducible) error curves, on a single plot, as we go from less flexible statistical learning methods towards more flexible approaches. The  $x$ -axis should represent the amount of flexibility in the method, and the  $y$ -axis should represent the values for each curve. There should be five curves. Make sure to label each one.



(c) Explain why the (i) training error and (ii) testing error lines in your drawing have the shape displayed.



4. (10 points) The data for this example come from a study by Stamey et al. (1989). The data comes from a number of clinical measures in men who were about to receive a radical prostatectomy. The goal is to predict the log of PSA (**lpsa**) (a prostate-specific antigen measured in nanograms of PSA per milliliter of blood (ng/mL)) from a number of measurements. The variables are log cancer volume (**lcavol**), log prostate weight (**lweight**), age, log of the amount of benign prostatic hyperplasia (**lbph**), seminal vesicle invasion (**svi**), log of capsular penetration (**lcp**), Gleason score (**gleason**), and percent of Gleason scores 4 or 5 (**pgg45**).

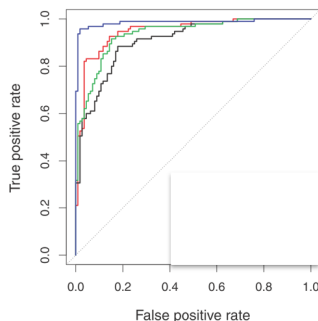
(a) Is this a case of regression or classification? Why?

(b) A linear model is fit to the data set, and the following table was returned.

Term	Coefficient	Std. Error	t-Score	p-value
Intercept	2.46	0.09	27.6	<0.00001
lcavol	0.68	0.13	5.37	<0.00001
lweight	0.26	0.1	2.75	0.00596
age	-0.14	0.1	-1.4	0.16153
lbph	0.21	0.1	2.06	0.039399
svi	0.31	0.12	2.47	0.013511
lcp	-0.29	0.15	-1.87	0.061484
gleason	-0.02	0.15	-0.15	0.880765
pgg45	0.27	0.15	1.74	0.081859

Are all the predictors useful? If yes, explain why. If not, explain what you would try to do next with the data.

5. (5 pts) We are trying to predict whether a patient will have a heart attack within one year. We have tried four different methods on a training dataset and have the following ROC curves. Which curve has the best performance on this training set? Justify your answer.





8B5D5AA5-9618-4926-87E6-2CB22233E328

midterm1-b9ace

#20 Page 6 of 10

6. (15 Points) I get way too much email, so I decide to build a logistic regression model to predict whether a new incoming message is spam or not. I decide to just use a few variables:

$$\begin{aligned}X_1 &= \text{Number\_of\_sentences} \\X_2 &= \text{Number\_of\_references\_to\_a\_Nigerian\_Prince}\end{aligned}$$

and am training a logistic model to predict

$$Pr(Y = \text{spam} \mid X_1, X_2)$$

- (a) Write down the equation for the model you would train, using our standard notation with  $\beta_i$ 's.
- (b) If my trained model used  $\beta_0 = -13.1$ ,  $\beta_1 = 1.9$ , and  $\beta_2 = 6.1$ , what is the probability that a 5 sentence email with one reference to a Nigerian prince is spam? .
- (c) We know that if we miss an important email, it may cost a lot, How can you modify your model to avoid this?



7. (15 Points) Sparty generated 100 pair  $x$  and  $y$  via the following relationship:  $Y = 0.3 + 2x + x^2 + \epsilon$  but didn't tell you. You will try to find a good model to fit these data and more importantly to predict future response  $y$  when Sparty gives you another  $x$ . You will start with two models 1)  $Y = \beta_0 + \beta_1x$ ; 2)  $Y = \alpha_0 + \alpha_1x + \alpha_2x^2$ ; and 3)  $Y = \gamma_0 + \gamma_1x + \gamma_2x^2 + \gamma_3x^3$ .
- (a) If you use all 100 data to fit these three model and use the same 100 data to calculate MSE, which model will have the smallest MSE? Justify your answer
- (b) Since we focus on the ability to predict unseen data, you decide use validation set to evaluate the performance of the three models. There are two way to split the data: 1) 80 testing points and 20 training points or 2) 20 testing points and 80 training points. Which one will you choose? Justify your choice
- (c) From the in-class lab, we know validation set is not a good choice to choose model. Explain the drawback of validation set and what procedure will you use to choose the best model?



D84DEDE0-84C7-48D0-AE54-CAE20CD4F8C2

midterm1-b9ace

#20 Page 8 of 10

8. (10 Points) Suppose we have a data set with five predictors,  $X_1 = \text{GPA}$ ,  $X_2 = \text{IQ}$ ,  $X_3 = \text{Level}$  (1 for College and 0 for High School),  $X_4 = \text{Interaction between GPA and IQ}$ , and  $X_5 = \text{Interaction between GPA and Level}$ . The response is starting salary after graduation (in thousands of dollars). Suppose we use least squares to fit the model, and get  $\beta_0 = 50$ ,  $\beta_1 = 20$ ,  $\beta_2 = 0.07$ ,  $\beta_3 = 35$ ,  $\beta_4 = 0.01$ ,  $\beta_5 = -10$ .

(a) Predict the salary of a college graduate with IQ of 110 and a GPA of 4.0.

- (b) True or false: Since the coefficient for the GPA/IQ interaction term is very small, there is very little evidence of an interaction effect. Justify your answer.

D91401DA-87C7-4218-B71F-FC49CD0F5FB9

midterm1-b9ace

#20 Page 9 of 10



**Scrap Paper**





758F6C79-A29C-4A0B-A9CE-0FDD6772E440

midterm1-b9ace

#20 Page 10 of 10



Michigan State University  
CMSE381 - Data Science

Feb 10, 2023  
Dr. Xie

## CMSE381 - Midterm #1

First name (please write as legibly as possible within the boxes)

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

Last name

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

NetID

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

*I will adhere to the Spartan Code of Honor in completing this assignment.*

Signed: \_\_\_\_\_

1. Do not open this test booklet until you are directed to do so.
2. You will have class time (2:40-4:00pm) to complete the exam.
3. This exam is closed book, but you can use the cheatsheet provided by the instructor.
4. Throughout the test, show your work so that your reasoning is clear. Otherwise no credit will be given. **BOX** your answers. Partial credit will be given where warranted.
5. Do not spend too much time on any one problem. Read them all through first and attack them in the order that allows you to make the most progress. Good luck :P



7F49F2B5-6164-4594-8403-14FEE8EFD2F5

midterm1-b9ace

#21 Page 2 of 10

1. (15 points)

(a) Logistic regression is used for regression.

TRUE      FALSE

(b) Any model will never have training error below the irreducible error.

TRUE      FALSE

(c) Increasing your model flexibility always results in a better model.

TRUE      FALSE

(d) A logistic regression model is set up so that the odds are linear.

TRUE      FALSE

(e) Circle all of the following that would represent a qualitative variable.

Age      Year      Dog\_breed      Country\_of\_origin

Student\_(True/False)      Weight      Speed      MPG

(f) What equation would you use to evaluate the result of a regression model?

(g) What equation would you use to evaluate the result of a classification model?



2. (15 Points) I'm building a model to predict amount of a given brand of dog food eaten by a collection of dogs. I have 100 dogs eat this dog food, and I collect information on their height, weight, breed, and whether they live with another dog in the house.
- (a) List all input variables and specify whether they are quantitative or qualitative.
- (b) Say our dog breeds sampled are Huskies, Terriers, and Spaniels. Write down your prediction model.
- (c) We found that weight is a key predictor, and its relationship with the response is beyond linear. What extension can you come up with? Write down your updated model.



7C74AF98-9FF2-4C1E-A212-8EA8FF1735AF

midterm1-b9ace

#21 Page 4 of 10

3. (15 points)

(a) What is bias-variance tradeoff? Explain the meaning of each term in the formula.

(b) Provide a sketch of typical (squared) bias, variance, training error, test error, and Bayes (or irreducible) error curves, on a single plot, as we go from less flexible statistical learning methods towards more flexible approaches. The  $x$ -axis should represent the amount of flexibility in the method, and the  $y$ -axis should represent the values for each curve. There should be five curves. Make sure to label each one.



(c) Explain why the (i) training error and (ii) testing error lines in your drawing have the shape displayed.



4. (10 points) The data for this example come from a study by Stamey et al. (1989). The data comes from a number of clinical measures in men who were about to receive a radical prostatectomy. The goal is to predict the log of PSA (**lpsa**) (a prostate-specific antigen measured in nanograms of PSA per milliliter of blood (ng/mL)) from a number of measurements. The variables are log cancer volume (**lcavol**), log prostate weight (**lweight**), age, log of the amount of benign prostatic hyperplasia (**lbph**), seminal vesicle invasion (**svi**), log of capsular penetration (**lcp**), Gleason score (**gleason**), and percent of Gleason scores 4 or 5 (**pgg45**).

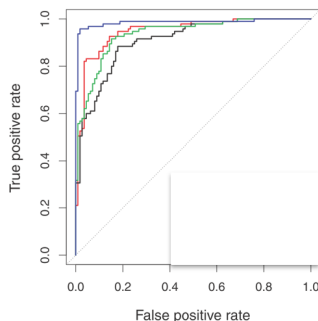
(a) Is this a case of regression or classification? Why?

(b) A linear model is fit to the data set, and the following table was returned.

Term	Coefficient	Std. Error	t-Score	p-value
Intercept	2.46	0.09	27.6	<0.00001
lcavol	0.68	0.13	5.37	<0.00001
lweight	0.26	0.1	2.75	0.00596
age	-0.14	0.1	-1.4	0.16153
lbph	0.21	0.1	2.06	0.039399
svi	0.31	0.12	2.47	0.013511
lcp	-0.29	0.15	-1.87	0.061484
gleason	-0.02	0.15	-0.15	0.880765
pgg45	0.27	0.15	1.74	0.081859

Are all the predictors useful? If yes, explain why. If not, explain what you would try to do next with the data.

5. (5 pts) We are trying to predict whether a patient will have a heart attack within one year. We have tried four different methods on a training dataset and have the following ROC curves. Which curve has the best performance on this training set? Justify your answer.





2404C9AD-C3A1-4248-9D0C-7A0770F8983C

midterm1-b9ace

#21 Page 6 of 10

6. (15 Points) I get way too much email, so I decide to build a logistic regression model to predict whether a new incoming message is spam or not. I decide to just use a few variables:

$$\begin{aligned}X_1 &= \text{Number\_of\_sentences} \\X_2 &= \text{Number\_of\_references\_to\_a\_Nigerian\_Prince}\end{aligned}$$

and am training a logistic model to predict

$$Pr(Y = \text{spam} \mid X_1, X_2)$$

- (a) Write down the equation for the model you would train, using our standard notation with  $\beta_i$ 's.
- (b) If my trained model used  $\beta_0 = -13.1$ ,  $\beta_1 = 1.9$ , and  $\beta_2 = 6.1$ , what is the probability that a 5 sentence email with one reference to a Nigerian prince is spam? .
- (c) We know that if we miss an important email, it may cost a lot, How can you modify your model to avoid this?



7. (15 Points) Sparty generated 100 pair  $x$  and  $y$  via the following relationship:  $Y = 0.3 + 2x + x^2 + \epsilon$  but didn't tell you. You will try to find a good model to fit these data and more importantly to predict future response  $y$  when Sparty gives you another  $x$ . You will start with two models 1)  $Y = \beta_0 + \beta_1x$ ; 2)  $Y = \alpha_0 + \alpha_1x + \alpha_2x^2$ ; and 3)  $Y = \gamma_0 + \gamma_1x + \gamma_2x^2 + \gamma_3x^3$ .
- (a) If you use all 100 data to fit these three model and use the same 100 data to calculate MSE, which model will have the smallest MSE? Justify your answer
- (b) Since we focus on the ability to predict unseen data, you decide use validation set to evaluate the performance of the three models. There are two way to split the data: 1) 80 testing points and 20 training points or 2) 20 testing points and 80 training points. Which one will you choose? Justify your choice
- (c) From the in-class lab, we know validation set is not a good choice to choose model. Explain the drawback of validation set and what procedure will you use to choose the best model?





8. (10 Points) Suppose we have a data set with five predictors,  $X_1 = \text{GPA}$ ,  $X_2 = \text{IQ}$ ,  $X_3 = \text{Level}$  (1 for College and 0 for High School),  $X_4 = \text{Interaction between GPA and IQ}$ , and  $X_5 = \text{Interaction between GPA and Level}$ . The response is starting salary after graduation (in thousands of dollars). Suppose we use least squares to fit the model, and get  $\beta_0 = 50$ ,  $\beta_1 = 20$ ,  $\beta_2 = 0.07$ ,  $\beta_3 = 35$ ,  $\beta_4 = 0.01$ ,  $\beta_5 = -10$ .

(a) Predict the salary of a college graduate with IQ of 110 and a GPA of 4.0.

(b) True or false: Since the coefficient for the GPA/IQ interaction term is very small, there is very little evidence of an interaction effect. Justify your answer.

78EDA807-F104-45F8-B9C5-DB9AD4CF3549

midterm1-b9ace

#21 Page 9 of 10



**Scrap Paper**



EBB66F22-21EC-4F74-B308-F0EA30C3DF6E

midterm1-b9ace

#21 Page 10 of 10



Michigan State University  
CMSE381 - Data Science

Feb 10, 2023  
Dr. Xie

## CMSE381 - Midterm #1

First name (please write as legibly as possible within the boxes)

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

Last name

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

NetID

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

*I will adhere to the Spartan Code of Honor in completing this assignment.*

Signed: \_\_\_\_\_

1. Do not open this test booklet until you are directed to do so.
2. You will have class time (2:40-4:00pm) to complete the exam.
3. This exam is closed book, but you can use the cheatsheet provided by the instructor.
4. Throughout the test, show your work so that your reasoning is clear. Otherwise no credit will be given. **BOX** your answers. Partial credit will be given where warranted.
5. Do not spend too much time on any one problem. Read them all through first and attack them in the order that allows you to make the most progress. Good luck :P



4716E2D0-362F-46E9-8989-6B5F5FDAE055

midterm1-b9ace

#22 Page 2 of 10

1. (15 points)

(a) Logistic regression is used for regression.

TRUE      FALSE

(b) Any model will never have training error below the irreducible error.

TRUE      FALSE

(c) Increasing your model flexibility always results in a better model.

TRUE      FALSE

(d) A logistic regression model is set up so that the odds are linear.

TRUE      FALSE

(e) Circle all of the following that would represent a qualitative variable.

Age      Year      Dog\_breed      Country\_of\_origin

Student\_(True/False)      Weight      Speed      MPG

(f) What equation would you use to evaluate the result of a regression model?

(g) What equation would you use to evaluate the result of a classification model?



2. (15 Points) I'm building a model to predict amount of a given brand of dog food eaten by a collection of dogs. I have 100 dogs eat this dog food, and I collect information on their height, weight, breed, and whether they live with another dog in the house.
- (a) List all input variables and specify whether they are quantitative or qualitative.
- (b) Say our dog breeds sampled are Huskies, Terriers, and Spaniels. Write down your prediction model.
- (c) We found that weight is a key predictor, and its relationship with the response is beyond linear. What extension can you come up with? Write down your updated model.



36C30B02-490E-4624-97F3-2C0BBAC0DD89

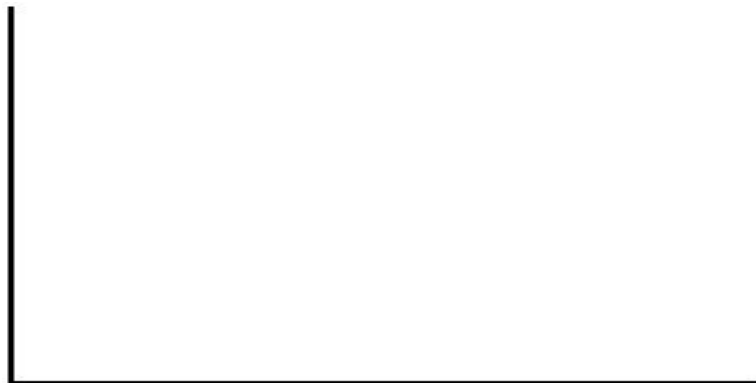
midterm1-b9ace

#22 Page 4 of 10

3. (15 points)

(a) What is bias-variance tradeoff? Explain the meaning of each term in the formula.

(b) Provide a sketch of typical (squared) bias, variance, training error, test error, and Bayes (or irreducible) error curves, on a single plot, as we go from less flexible statistical learning methods towards more flexible approaches. The  $x$ -axis should represent the amount of flexibility in the method, and the  $y$ -axis should represent the values for each curve. There should be five curves. Make sure to label each one.



(c) Explain why the (i) training error and (ii) testing error lines in your drawing have the shape displayed.



4. (10 points) The data for this example come from a study by Stamey et al. (1989). The data comes from a number of clinical measures in men who were about to receive a radical prostatectomy. The goal is to predict the log of PSA (**lpsa**) (a prostate-specific antigen measured in nanograms of PSA per milliliter of blood (ng/mL)) from a number of measurements. The variables are log cancer volume (**lcavol**), log prostate weight (**lweight**), age, log of the amount of benign prostatic hyperplasia (**lbph**), seminal vesicle invasion (**svi**), log of capsular penetration (**lcp**), Gleason score (**gleason**), and percent of Gleason scores 4 or 5 (**pgg45**).

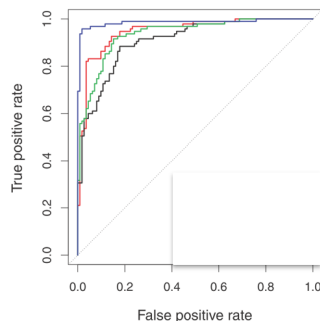
(a) Is this a case of regression or classification? Why?

(b) A linear model is fit to the data set, and the following table was returned.

Term	Coefficient	Std. Error	t-Score	p-value
Intercept	2.46	0.09	27.6	<0.00001
lcavol	0.68	0.13	5.37	<0.00001
lweight	0.26	0.1	2.75	0.00596
age	-0.14	0.1	-1.4	0.16153
lbph	0.21	0.1	2.06	0.039399
svi	0.31	0.12	2.47	0.013511
lcp	-0.29	0.15	-1.87	0.061484
gleason	-0.02	0.15	-0.15	0.880765
pgg45	0.27	0.15	1.74	0.081859

Are all the predictors useful? If yes, explain why. If not, explain what you would try to do next with the data.

5. (5 pts) We are trying to predict whether a patient will have a heart attack within one year. We have tried four different methods on a training dataset and have the following ROC curves. Which curve has the best performance on this training set? Justify your answer.







6. (15 Points) I get way too much email, so I decide to build a logistic regression model to predict whether a new incoming message is spam or not. I decide to just use a few variables:

$$\begin{aligned}X_1 &= \text{Number\_of\_sentences} \\X_2 &= \text{Number\_of\_references\_to\_a\_Nigerian\_Prince}\end{aligned}$$

and am training a logistic model to predict

$$Pr(Y = \text{spam} \mid X_1, X_2)$$

- (a) Write down the equation for the model you would train, using our standard notation with  $\beta_i$ 's.
- (b) If my trained model used  $\beta_0 = -13.1$ ,  $\beta_1 = 1.9$ , and  $\beta_2 = 6.1$ , what is the probability that a 5 sentence email with one reference to a Nigerian prince is spam? .
- (c) We know that if we miss an important email, it may cost a lot, How can you modify your model to avoid this?



7. (15 Points) Sparty generated 100 pair  $x$  and  $y$  via the following relationship:  $Y = 0.3 + 2x + x^2 + \epsilon$  but didn't tell you. You will try to find a good model to fit these data and more importantly to predict future response  $y$  when Sparty gives you another  $x$ . You will start with two models 1)  $Y = \beta_0 + \beta_1x$ ; 2)  $Y = \alpha_0 + \alpha_1x + \alpha_2x^2$ ; and 3)  $Y = \gamma_0 + \gamma_1x + \gamma_2x^2 + \gamma_3x^3$ .
- (a) If you use all 100 data to fit these three model and use the same 100 data to calculate MSE, which model will have the smallest MSE? Justify your answer
- (b) Since we focus on the ability to predict unseen data, you decide use validation set to evaluate the performance of the three models. There are two way to split the data: 1) 80 testing points and 20 training points or 2) 20 testing points and 80 training points. Which one will you choose? Justify your choice
- (c) From the in-class lab, we know validation set is not a good choice to choose model. Explain the drawback of validation set and what procedure will you use to choose the best model?



270B53BE-33FF-447E-AD49-0E88710961EB

midterm1-b9ace

#22 Page 8 of 10

8. (10 Points) Suppose we have a data set with five predictors,  $X_1 = \text{GPA}$ ,  $X_2 = \text{IQ}$ ,  $X_3 = \text{Level}$  (1 for College and 0 for High School),  $X_4 = \text{Interaction between GPA and IQ}$ , and  $X_5 = \text{Interaction between GPA and Level}$ . The response is starting salary after graduation (in thousands of dollars). Suppose we use least squares to fit the model, and get  $\beta_0 = 50$ ,  $\beta_1 = 20$ ,  $\beta_2 = 0.07$ ,  $\beta_3 = 35$ ,  $\beta_4 = 0.01$ ,  $\beta_5 = -10$ .

(a) Predict the salary of a college graduate with IQ of 110 and a GPA of 4.0.

(b) True or false: Since the coefficient for the GPA/IQ interaction term is very small, there is very little evidence of an interaction effect. Justify your answer.

0132EA9C-C672-4557-A44E-6D0B60C27AC5

midterm1-b9ace

#22 Page 9 of 10



## Scrap Paper



DC867504-C89C-47B5-9229-AF1E06B45B4E

midterm1-b9ace

#22 Page 10 of 10



Michigan State University  
CMSE381 - Data Science

Feb 10, 2023  
Dr. Xie

## CMSE381 - Midterm #1

First name (please write as legibly as possible within the boxes)

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

Last name

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

NetID

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

*I will adhere to the Spartan Code of Honor in completing this assignment.*

Signed: \_\_\_\_\_

1. Do not open this test booklet until you are directed to do so.
2. You will have class time (2:40-4:00pm) to complete the exam.
3. This exam is closed book, but you can use the cheatsheet provided by the instructor.
4. Throughout the test, show your work so that your reasoning is clear. Otherwise no credit will be given. **BOX** your answers. Partial credit will be given where warranted.
5. Do not spend too much time on any one problem. Read them all through first and attack them in the order that allows you to make the most progress. Good luck :P



80C8FA48-F637-4F54-BB06-E648216D6A81

midterm1-b9ace

#23 Page 2 of 10

1. (15 points)

(a) Logistic regression is used for regression.

TRUE      FALSE

(b) Any model will never have training error below the irreducible error.

TRUE      FALSE

(c) Increasing your model flexibility always results in a better model.

TRUE      FALSE

(d) A logistic regression model is set up so that the odds are linear.

TRUE      FALSE

(e) Circle all of the following that would represent a qualitative variable.

Age      Year      Dog\_breed      Country\_of\_origin

Student\_(True/False)      Weight      Speed      MPG

(f) What equation would you use to evaluate the result of a regression model?

(g) What equation would you use to evaluate the result of a classification model?



2. (15 Points) I'm building a model to predict amount of a given brand of dog food eaten by a collection of dogs. I have 100 dogs eat this dog food, and I collect information on their height, weight, breed, and whether they live with another dog in the house.
- (a) List all input variables and specify whether they are quantitative or qualitative.
  - (b) Say our dog breeds sampled are Huskies, Terriers, and Spaniels. Write down your prediction model.
  - (c) We found that weight is a key predictor, and its relationship with the response is beyond linear. What extension can you come up with? Write down your updated model.





4E8FCEAA-99D2-4BFC-844D-B3AFBA3FE41E

midterm1-b9ace

#23 Page 4 of 10

3. (15 points)

(a) What is bias-variance tradeoff? Explain the meaning of each term in the formula.

(b) Provide a sketch of typical (squared) bias, variance, training error, test error, and Bayes (or irreducible) error curves, on a single plot, as we go from less flexible statistical learning methods towards more flexible approaches. The  $x$ -axis should represent the amount of flexibility in the method, and the  $y$ -axis should represent the values for each curve. There should be five curves. Make sure to label each one.



(c) Explain why the (i) training error and (ii) testing error lines in your drawing have the shape displayed.



4. (10 points) The data for this example come from a study by Stamey et al. (1989). The data comes from a number of clinical measures in men who were about to receive a radical prostatectomy. The goal is to predict the log of PSA (**lpsa**) (a prostate-specific antigen measured in nanograms of PSA per milliliter of blood (ng/mL)) from a number of measurements. The variables are log cancer volume (**lcavol**), log prostate weight (**lweight**), age, log of the amount of benign prostatic hyperplasia (**lbph**), seminal vesicle invasion (**svi**), log of capsular penetration (**lcp**), Gleason score (**gleason**), and percent of Gleason scores 4 or 5 (**pgg45**).

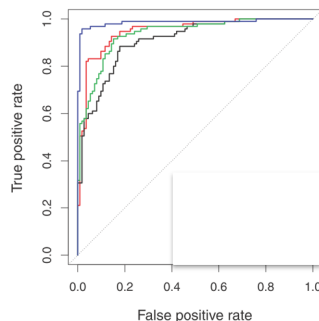
(a) Is this a case of regression or classification? Why?

(b) A linear model is fit to the data set, and the following table was returned.

Term	Coefficient	Std. Error	t-Score	p-value
Intercept	2.46	0.09	27.6	<0.00001
lcavol	0.68	0.13	5.37	<0.00001
lweight	0.26	0.1	2.75	0.00596
age	-0.14	0.1	-1.4	0.16153
lbph	0.21	0.1	2.06	0.039399
svi	0.31	0.12	2.47	0.013511
lcp	-0.29	0.15	-1.87	0.061484
gleason	-0.02	0.15	-0.15	0.880765
pgg45	0.27	0.15	1.74	0.081859

Are all the predictors useful? If yes, explain why. If not, explain what you would try to do next with the data.

5. (5 pts) We are trying to predict whether a patient will have a heart attack within one year. We have tried four different methods on a training dataset and have the following ROC curves. Which curve has the best performance on this training set? Justify your answer.





F6F5AB98-8548-4BD0-9782-3E4954D6F090

midterm1-b9ace

#23 Page 6 of 10

6. (15 Points) I get way too much email, so I decide to build a logistic regression model to predict whether a new incoming message is spam or not. I decide to just use a few variables:

$$X_1 = \text{Number\_of\_sentences}$$

$$X_2 = \text{Number\_of\_references\_to\_a\_Nigerian\_Prince}$$

and am training a logistic model to predict

$$Pr(Y = \text{spam} \mid X_1, X_2)$$

- (a) Write down the equation for the model you would train, using our standard notation with  $\beta_i$ 's.
- (b) If my trained model used  $\beta_0 = -13.1$ ,  $\beta_1 = 1.9$ , and  $\beta_2 = 6.1$ , what is the probability that a 5 sentence email with one reference to a Nigerian prince is spam? .
- (c) We know that if we miss an important email, it may cost a lot, How can you modify your model to avoid this?



7. (15 Points) Sparty generated 100 pair  $x$  and  $y$  via the following relationship:  $Y = 0.3 + 2x + x^2 + \epsilon$  but didn't tell you. You will try to find a good model to fit these data and more importantly to predict future response  $y$  when Sparty gives you another  $x$ . You will start with two models 1)  $Y = \beta_0 + \beta_1x$ ; 2)  $Y = \alpha_0 + \alpha_1x + \alpha_2x^2$ ; and 3)  $Y = \gamma_0 + \gamma_1x + \gamma_2x^2 + \gamma_3x^3$ .
- (a) If you use all 100 data to fit these three model and use the same 100 data to calculate MSE, which model will have the smallest MSE? Justify your answer
- (b) Since we focus on the ability to predict unseen data, you decide use validation set to evaluate the performance of the three models. There are two way to split the data: 1) 80 testing points and 20 training points or 2) 20 testing points and 80 training points. Which one will you choose? Justify your choice
- (c) From the in-class lab, we know validation set is not a good choice to choose model. Explain the drawback of validation set and what procedure will you use to choose the best model?



8. (10 Points) Suppose we have a data set with five predictors,  $X_1 = \text{GPA}$ ,  $X_2 = \text{IQ}$ ,  $X_3 = \text{Level}$  (1 for College and 0 for High School),  $X_4 = \text{Interaction between GPA and IQ}$ , and  $X_5 = \text{Interaction between GPA and Level}$ . The response is starting salary after graduation (in thousands of dollars). Suppose we use least squares to fit the model, and get  $\beta_0 = 50$ ,  $\beta_1 = 20$ ,  $\beta_2 = 0.07$ ,  $\beta_3 = 35$ ,  $\beta_4 = 0.01$ ,  $\beta_5 = -10$ .

(a) Predict the salary of a college graduate with IQ of 110 and a GPA of 4.0.

- (b) True or false: Since the coefficient for the GPA/IQ interaction term is very small, there is very little evidence of an interaction effect. Justify your answer.

D6885AB3-ED5E-4F8D-B2AB-B4B997EC38BD

midterm1-b9ace

#23 Page 9 of 10



## Scrap Paper



222E9477-7BF6-448D-A414-45E66DEF8C71

midterm1-b9ace

#23 Page 10 of 10



Michigan State University  
CMSE381 - Data Science

Feb 10, 2023  
Dr. Xie

## CMSE381 - Midterm #1

First name (please write as legibly as possible within the boxes)

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

Last name

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

NetID

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

*I will adhere to the Spartan Code of Honor in completing this assignment.*

Signed: \_\_\_\_\_

1. Do not open this test booklet until you are directed to do so.
2. You will have class time (2:40-4:00pm) to complete the exam.
3. This exam is closed book, but you can use the cheatsheet provided by the instructor.
4. Throughout the test, show your work so that your reasoning is clear. Otherwise no credit will be given. **BOX** your answers. Partial credit will be given where warranted.
5. Do not spend too much time on any one problem. Read them all through first and attack them in the order that allows you to make the most progress. Good luck :P





1C356DAC-1115-485D-BBBB-2BF4D7F3B921

midterm1-b9ace

#24 Page 2 of 10

1. (15 points)

(a) Logistic regression is used for regression.

TRUE      FALSE

(b) Any model will never have training error below the irreducible error.

TRUE      FALSE

(c) Increasing your model flexibility always results in a better model.

TRUE      FALSE

(d) A logistic regression model is set up so that the odds are linear.

TRUE      FALSE

(e) Circle all of the following that would represent a qualitative variable.

Age      Year      Dog\_breed      Country\_of\_origin

Student\_(True/False)      Weight      Speed      MPG

(f) What equation would you use to evaluate the result of a regression model?

(g) What equation would you use to evaluate the result of a classification model?



2. (15 Points) I'm building a model to predict amount of a given brand of dog food eaten by a collection of dogs. I have 100 dogs eat this dog food, and I collect information on their height, weight, breed, and whether they live with another dog in the house.
- (a) List all input variables and specify whether they are quantitative or qualitative.
- (b) Say our dog breeds sampled are Huskies, Terriers, and Spaniels. Write down your prediction model.
- (c) We found that weight is a key predictor, and its relationship with the response is beyond linear. What extension can you come up with? Write down your updated model.



47DE1243-B0CE-48A4-A3A5-B99A7648ADA8

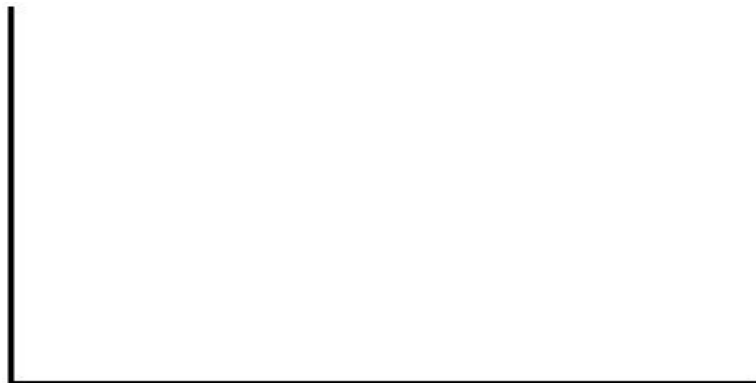
midterm1-b9ace

#24 Page 4 of 10

3. (15 points)

(a) What is bias-variance tradeoff? Explain the meaning of each term in the formula.

(b) Provide a sketch of typical (squared) bias, variance, training error, test error, and Bayes (or irreducible) error curves, on a single plot, as we go from less flexible statistical learning methods towards more flexible approaches. The  $x$ -axis should represent the amount of flexibility in the method, and the  $y$ -axis should represent the values for each curve. There should be five curves. Make sure to label each one.



(c) Explain why the (i) training error and (ii) testing error lines in your drawing have the shape displayed.



4. (10 points) The data for this example come from a study by Stamey et al. (1989). The data comes from a number of clinical measures in men who were about to receive a radical prostatectomy. The goal is to predict the log of PSA (**lpsa**) (a prostate-specific antigen measured in nanograms of PSA per milliliter of blood (ng/mL)) from a number of measurements. The variables are log cancer volume (**lcavol**), log prostate weight (**lweight**), age, log of the amount of benign prostatic hyperplasia (**lbph**), seminal vesicle invasion (**svi**), log of capsular penetration (**lcp**), Gleason score (**gleason**), and percent of Gleason scores 4 or 5 (**pgg45**).

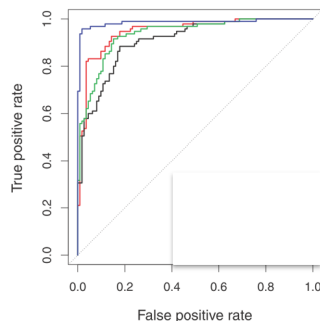
(a) Is this a case of regression or classification? Why?

(b) A linear model is fit to the data set, and the following table was returned.

Term	Coefficient	Std. Error	t-Score	p-value
Intercept	2.46	0.09	27.6	<0.00001
lcavol	0.68	0.13	5.37	<0.00001
lweight	0.26	0.1	2.75	0.00596
age	-0.14	0.1	-1.4	0.16153
lbph	0.21	0.1	2.06	0.039399
svi	0.31	0.12	2.47	0.013511
lcp	-0.29	0.15	-1.87	0.061484
gleason	-0.02	0.15	-0.15	0.880765
pgg45	0.27	0.15	1.74	0.081859

Are all the predictors useful? If yes, explain why. If not, explain what you would try to do next with the data.

5. (5 pts) We are trying to predict whether a patient will have a heart attack within one year. We have tried four different methods on a training dataset and have the following ROC curves. Which curve has the best performance on this training set? Justify your answer.





6. (15 Points) I get way too much email, so I decide to build a logistic regression model to predict whether a new incoming message is spam or not. I decide to just use a few variables:

$$\begin{aligned}X_1 &= \text{Number\_of\_sentences} \\X_2 &= \text{Number\_of\_references\_to\_a\_Nigerian\_Prince}\end{aligned}$$

and am training a logistic model to predict

$$Pr(Y = \text{spam} \mid X_1, X_2)$$

- (a) Write down the equation for the model you would train, using our standard notation with  $\beta_i$ 's.
- (b) If my trained model used  $\beta_0 = -13.1$ ,  $\beta_1 = 1.9$ , and  $\beta_2 = 6.1$ , what is the probability that a 5 sentence email with one reference to a Nigerian prince is spam? .
- (c) We know that if we miss an important email, it may cost a lot, How can you modify your model to avoid this?



7. (15 Points) Sparty generated 100 pair  $x$  and  $y$  via the following relationship:  $Y = 0.3 + 2x + x^2 + \epsilon$  but didn't tell you. You will try to find a good model to fit these data and more importantly to predict future response  $y$  when Sparty gives you another  $x$ . You will start with two models 1)  $Y = \beta_0 + \beta_1x$ ; 2)  $Y = \alpha_0 + \alpha_1x + \alpha_2x^2$ ; and 3)  $Y = \gamma_0 + \gamma_1x + \gamma_2x^2 + \gamma_3x^3$ .
- (a) If you use all 100 data to fit these three model and use the same 100 data to calculate MSE, which model will have the smallest MSE? Justify your answer
- (b) Since we focus on the ability to predict unseen data, you decide use validation set to evaluate the performance of the three models. There are two way to split the data: 1) 80 testing points and 20 training points or 2) 20 testing points and 80 training points. Which one will you choose? Justify your choice
- (c) From the in-class lab, we know validation set is not a good choice to choose model. Explain the drawback of validation set and what procedure will you use to choose the best model?



067048CD-CA91-402A-A644-A3E91985F90F

midterm1-b9ace

#24 Page 8 of 10

8. (10 Points) Suppose we have a data set with five predictors,  $X_1 = \text{GPA}$ ,  $X_2 = \text{IQ}$ ,  $X_3 = \text{Level}$  (1 for College and 0 for High School),  $X_4 = \text{Interaction between GPA and IQ}$ , and  $X_5 = \text{Interaction between GPA and Level}$ . The response is starting salary after graduation (in thousands of dollars). Suppose we use least squares to fit the model, and get  $\beta_0 = 50$ ,  $\beta_1 = 20$ ,  $\beta_2 = 0.07$ ,  $\beta_3 = 35$ ,  $\beta_4 = 0.01$ ,  $\beta_5 = -10$ .

(a) Predict the salary of a college graduate with IQ of 110 and a GPA of 4.0.

- (b) True or false: Since the coefficient for the GPA/IQ interaction term is very small, there is very little evidence of an interaction effect. Justify your answer.

689ECD9F-0684-443D-951B-C06AF4202E97

midterm1-b9ace

#24 Page 9 of 10



## Scrap Paper





E4BCCDF3-D4D2-4502-B2CF-6452CB96B06E

midterm1-b9ace

#24 Page 10 of 10



Michigan State University  
CMSE381 - Data Science

Feb 10, 2023  
Dr. Xie

## CMSE381 - Midterm #1

First name (please write as legibly as possible within the boxes)

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

Last name

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

NetID

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

*I will adhere to the Spartan Code of Honor in completing this assignment.*

Signed: \_\_\_\_\_

1. Do not open this test booklet until you are directed to do so.
2. You will have class time (2:40-4:00pm) to complete the exam.
3. This exam is closed book, but you can use the cheatsheet provided by the instructor.
4. Throughout the test, show your work so that your reasoning is clear. Otherwise no credit will be given. **BOX** your answers. Partial credit will be given where warranted.
5. Do not spend too much time on any one problem. Read them all through first and attack them in the order that allows you to make the most progress. Good luck :P



DDCCBA6A-72F6-47D2-B43F-E0B89CDFC5C8

midterm1-b9ace

#25 Page 2 of 10

1. (15 points)

(a) Logistic regression is used for regression.

TRUE      FALSE

(b) Any model will never have training error below the irreducible error.

TRUE      FALSE

(c) Increasing your model flexibility always results in a better model.

TRUE      FALSE

(d) A logistic regression model is set up so that the odds are linear.

TRUE      FALSE

(e) Circle all of the following that would represent a qualitative variable.

Age      Year      Dog\_breed      Country\_of\_origin

Student\_(True/False)      Weight      Speed      MPG

(f) What equation would you use to evaluate the result of a regression model?

(g) What equation would you use to evaluate the result of a classification model?



2. (15 Points) I'm building a model to predict amount of a given brand of dog food eaten by a collection of dogs. I have 100 dogs eat this dog food, and I collect information on their height, weight, breed, and whether they live with another dog in the house.
- (a) List all input variables and specify whether they are quantitative or qualitative.
- (b) Say our dog breeds sampled are Huskies, Terriers, and Spaniels. Write down your prediction model.
- (c) We found that weight is a key predictor, and its relationship with the response is beyond linear. What extension can you come up with? Write down your updated model.



517BA6AF-6BB0-45CA-B162-2B2F52B93DB6

midterm1-b9ace

#25 Page 4 of 10

3. (15 points)

(a) What is bias-variance tradeoff? Explain the meaning of each term in the formula.

(b) Provide a sketch of typical (squared) bias, variance, training error, test error, and Bayes (or irreducible) error curves, on a single plot, as we go from less flexible statistical learning methods towards more flexible approaches. The  $x$ -axis should represent the amount of flexibility in the method, and the  $y$ -axis should represent the values for each curve. There should be five curves. Make sure to label each one.



(c) Explain why the (i) training error and (ii) testing error lines in your drawing have the shape displayed.



4. (10 points) The data for this example come from a study by Stamey et al. (1989). The data comes from a number of clinical measures in men who were about to receive a radical prostatectomy. The goal is to predict the log of PSA (**lpsa**) (a prostate-specific antigen measured in nanograms of PSA per milliliter of blood (ng/mL)) from a number of measurements. The variables are log cancer volume (**lcavol**), log prostate weight (**lweight**), age, log of the amount of benign prostatic hyperplasia (**lbph**), seminal vesicle invasion (**svi**), log of capsular penetration (**lcp**), Gleason score (**gleason**), and percent of Gleason scores 4 or 5 (**pgg45**).

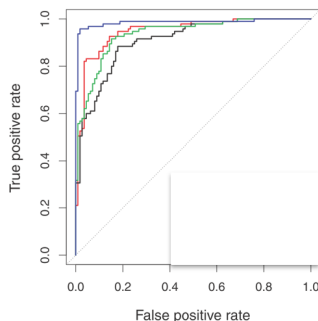
(a) Is this a case of regression or classification? Why?

(b) A linear model is fit to the data set, and the following table was returned.

Term	Coefficient	Std. Error	t-Score	p-value
Intercept	2.46	0.09	27.6	<0.00001
lcavol	0.68	0.13	5.37	<0.00001
lweight	0.26	0.1	2.75	0.00596
age	-0.14	0.1	-1.4	0.16153
lbph	0.21	0.1	2.06	0.039399
svi	0.31	0.12	2.47	0.013511
lcp	-0.29	0.15	-1.87	0.061484
gleason	-0.02	0.15	-0.15	0.880765
pgg45	0.27	0.15	1.74	0.081859

Are all the predictors useful? If yes, explain why. If not, explain what you would try to do next with the data.

5. (5 pts) We are trying to predict whether a patient will have a heart attack within one year. We have tried four different methods on a training dataset and have the following ROC curves. Which curve has the best performance on this training set? Justify your answer.





6. (15 Points) I get way too much email, so I decide to build a logistic regression model to predict whether a new incoming message is spam or not. I decide to just use a few variables:

$$\begin{aligned}X_1 &= \text{Number\_of\_sentences} \\X_2 &= \text{Number\_of\_references\_to\_a\_Nigerian\_Prince}\end{aligned}$$

and am training a logistic model to predict

$$Pr(Y = \text{spam} \mid X_1, X_2)$$

- (a) Write down the equation for the model you would train, using our standard notation with  $\beta_i$ 's.
- (b) If my trained model used  $\beta_0 = -13.1$ ,  $\beta_1 = 1.9$ , and  $\beta_2 = 6.1$ , what is the probability that a 5 sentence email with one reference to a Nigerian prince is spam? .
- (c) We know that if we miss an important email, it may cost a lot, How can you modify your model to avoid this?



7. (15 Points) Sparty generated 100 pair  $x$  and  $y$  via the following relationship:  $Y = 0.3 + 2x + x^2 + \epsilon$  but didn't tell you. You will try to find a good model to fit these data and more importantly to predict future response  $y$  when Sparty gives you another  $x$ . You will start with two models 1)  $Y = \beta_0 + \beta_1x$ ; 2)  $Y = \alpha_0 + \alpha_1x + \alpha_2x^2$ ; and 3)  $Y = \gamma_0 + \gamma_1x + \gamma_2x^2 + \gamma_3x^3$ .
- (a) If you use all 100 data to fit these three model and use the same 100 data to calculate MSE, which model will have the smallest MSE? Justify your answer
- (b) Since we focus on the ability to predict unseen data, you decide use validation set to evaluate the performance of the three models. There are two way to split the data: 1) 80 testing points and 20 training points or 2) 20 testing points and 80 training points. Which one will you choose? Justify your choice
- (c) From the in-class lab, we know validation set is not a good choice to choose model. Explain the drawback of validation set and what procedure will you use to choose the best model?





C96D8C08-675B-444E-B632-11C962A93B5C

midterm1-b9ace

#25 Page 8 of 10

8. (10 Points) Suppose we have a data set with five predictors,  $X_1 = \text{GPA}$ ,  $X_2 = \text{IQ}$ ,  $X_3 = \text{Level}$  (1 for College and 0 for High School),  $X_4 = \text{Interaction between GPA and IQ}$ , and  $X_5 = \text{Interaction between GPA and Level}$ . The response is starting salary after graduation (in thousands of dollars). Suppose we use least squares to fit the model, and get  $\beta_0 = 50$ ,  $\beta_1 = 20$ ,  $\beta_2 = 0.07$ ,  $\beta_3 = 35$ ,  $\beta_4 = 0.01$ ,  $\beta_5 = -10$ .

(a) Predict the salary of a college graduate with IQ of 110 and a GPA of 4.0.

- (b) True or false: Since the coefficient for the GPA/IQ interaction term is very small, there is very little evidence of an interaction effect. Justify your answer.

37A6CF06-A0B6-42A2-A0F7-00F31E648DAF

midterm1-b9ace

#25 Page 9 of 10



**Scrap Paper**



C55D32CE-A1C1-453F-A23B-AAB66E606B86

midterm1-b9ace

#25 Page 10 of 10



Michigan State University  
CMSE381 - Data Science

Feb 10, 2023  
Dr. Xie

## CMSE381 - Midterm #1

First name (please write as legibly as possible within the boxes)

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

Last name

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

NetID

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

*I will adhere to the Spartan Code of Honor in completing this assignment.*

Signed: \_\_\_\_\_

1. Do not open this test booklet until you are directed to do so.
2. You will have class time (2:40-4:00pm) to complete the exam.
3. This exam is closed book, but you can use the cheatsheet provided by the instructor.
4. Throughout the test, show your work so that your reasoning is clear. Otherwise no credit will be given. **BOX** your answers. Partial credit will be given where warranted.
5. Do not spend too much time on any one problem. Read them all through first and attack them in the order that allows you to make the most progress. Good luck :P



821BC322-7070-4652-A703-1A68436940BC

midterm1-b9ace

#26 Page 2 of 10

1. (15 points)

(a) Logistic regression is used for regression.

TRUE      FALSE

(b) Any model will never have training error below the irreducible error.

TRUE      FALSE

(c) Increasing your model flexibility always results in a better model.

TRUE      FALSE

(d) A logistic regression model is set up so that the odds are linear.

TRUE      FALSE

(e) Circle all of the following that would represent a qualitative variable.

Age      Year      Dog\_breed      Country\_of\_origin

Student\_(True/False)      Weight      Speed      MPG

(f) What equation would you use to evaluate the result of a regression model?

(g) What equation would you use to evaluate the result of a classification model?



2. (15 Points) I'm building a model to predict amount of a given brand of dog food eaten by a collection of dogs. I have 100 dogs eat this dog food, and I collect information on their height, weight, breed, and whether they live with another dog in the house.
- (a) List all input variables and specify whether they are quantitative or qualitative.
- (b) Say our dog breeds sampled are Huskies, Terriers, and Spaniels. Write down your prediction model.
- (c) We found that weight is a key predictor, and its relationship with the response is beyond linear. What extension can you come up with? Write down your updated model.



D9C3F542-E26F-4F33-974E-2A203FFDA010

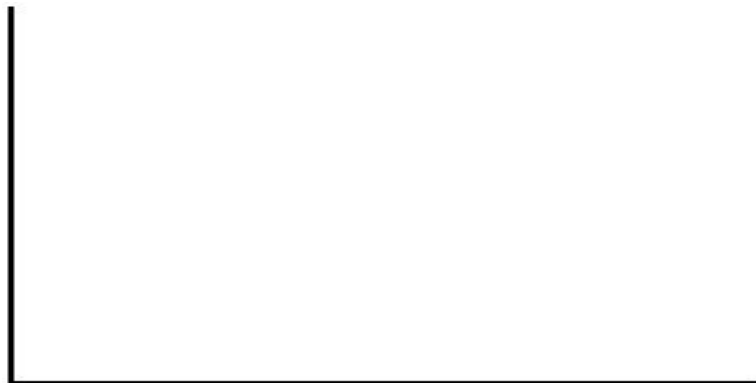
midterm1-b9ace

#26 Page 4 of 10

3. (15 points)

(a) What is bias-variance tradeoff? Explain the meaning of each term in the formula.

(b) Provide a sketch of typical (squared) bias, variance, training error, test error, and Bayes (or irreducible) error curves, on a single plot, as we go from less flexible statistical learning methods towards more flexible approaches. The  $x$ -axis should represent the amount of flexibility in the method, and the  $y$ -axis should represent the values for each curve. There should be five curves. Make sure to label each one.



(c) Explain why the (i) training error and (ii) testing error lines in your drawing have the shape displayed.



4. (10 points) The data for this example come from a study by Stamey et al. (1989). The data comes from a number of clinical measures in men who were about to receive a radical prostatectomy. The goal is to predict the log of PSA (**lpsa**) (a prostate-specific antigen measured in nanograms of PSA per milliliter of blood (ng/mL)) from a number of measurements. The variables are log cancer volume (**lcavol**), log prostate weight (**lweight**), age, log of the amount of benign prostatic hyperplasia (**lbph**), seminal vesicle invasion (**svi**), log of capsular penetration (**lcp**), Gleason score (**gleason**), and percent of Gleason scores 4 or 5 (**pgg45**).

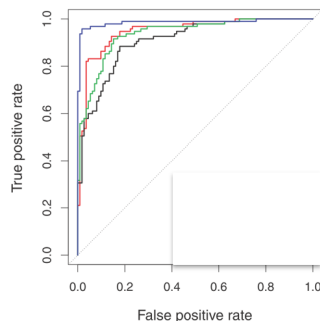
(a) Is this a case of regression or classification? Why?

(b) A linear model is fit to the data set, and the following table was returned.

Term	Coefficient	Std. Error	t-Score	p-value
Intercept	2.46	0.09	27.6	<0.00001
lcavol	0.68	0.13	5.37	<0.00001
lweight	0.26	0.1	2.75	0.00596
age	-0.14	0.1	-1.4	0.16153
lbph	0.21	0.1	2.06	0.039399
svi	0.31	0.12	2.47	0.013511
lcp	-0.29	0.15	-1.87	0.061484
gleason	-0.02	0.15	-0.15	0.880765
pgg45	0.27	0.15	1.74	0.081859

Are all the predictors useful? If yes, explain why. If not, explain what you would try to do next with the data.

5. (5 pts) We are trying to predict whether a patient will have a heart attack within one year. We have tried four different methods on a training dataset and have the following ROC curves. Which curve has the best performance on this training set? Justify your answer.







D4AF16FF-0FE7-4A06-876E-324A66CAFDDF

midterm1-b9ace

#26 Page 6 of 10

6. (15 Points) I get way too much email, so I decide to build a logistic regression model to predict whether a new incoming message is spam or not. I decide to just use a few variables:

$$\begin{aligned}X_1 &= \text{Number\_of\_sentences} \\X_2 &= \text{Number\_of\_references\_to\_a\_Nigerian\_Prince}\end{aligned}$$

and am training a logistic model to predict

$$Pr(Y = \text{spam} \mid X_1, X_2)$$

- (a) Write down the equation for the model you would train, using our standard notation with  $\beta_i$ 's.
- (b) If my trained model used  $\beta_0 = -13.1$ ,  $\beta_1 = 1.9$ , and  $\beta_2 = 6.1$ , what is the probability that a 5 sentence email with one reference to a Nigerian prince is spam? .
- (c) We know that if we miss an important email, it may cost a lot, How can you modify your model to avoid this?



7. (15 Points) Sparty generated 100 pair  $x$  and  $y$  via the following relationship:  $Y = 0.3 + 2x + x^2 + \epsilon$  but didn't tell you. You will try to find a good model to fit these data and more importantly to predict future response  $y$  when Sparty gives you another  $x$ . You will start with two models 1)  $Y = \beta_0 + \beta_1x$ ; 2)  $Y = \alpha_0 + \alpha_1x + \alpha_2x^2$ ; and 3)  $Y = \gamma_0 + \gamma_1x + \gamma_2x^2 + \gamma_3x^3$ .
- (a) If you use all 100 data to fit these three model and use the same 100 data to calculate MSE, which model will have the smallest MSE? Justify your answer
- (b) Since we focus on the ability to predict unseen data, you decide use validation set to evaluate the performance of the three models. There are two way to split the data: 1) 80 testing points and 20 training points or 2) 20 testing points and 80 training points. Which one will you choose? Justify your choice
- (c) From the in-class lab, we know validation set is not a good choice to choose model. Explain the drawback of validation set and what procedure will you use to choose the best model?



367104A6-B7D9-4173-9AB2-DCF641A9999C

midterm1-b9ace

#26 Page 8 of 10

8. (10 Points) Suppose we have a data set with five predictors,  $X_1 = \text{GPA}$ ,  $X_2 = \text{IQ}$ ,  $X_3 = \text{Level}$  (1 for College and 0 for High School),  $X_4 = \text{Interaction between GPA and IQ}$ , and  $X_5 = \text{Interaction between GPA and Level}$ . The response is starting salary after graduation (in thousands of dollars). Suppose we use least squares to fit the model, and get  $\beta_0 = 50$ ,  $\beta_1 = 20$ ,  $\beta_2 = 0.07$ ,  $\beta_3 = 35$ ,  $\beta_4 = 0.01$ ,  $\beta_5 = -10$ .

(a) Predict the salary of a college graduate with IQ of 110 and a GPA of 4.0.

(b) True or false: Since the coefficient for the GPA/IQ interaction term is very small, there is very little evidence of an interaction effect. Justify your answer.

BFD11AD2-DB8C-463D-89E7-CBC33FC85E81

midterm1-b9ace

#26 Page 9 of 10



**Scrap Paper**



A046BEA7-E8C9-40FF-A3C6-E7DE558FDE6B

midterm1-b9ace

#26 Page 10 of 10



Michigan State University  
CMSE381 - Data Science

Feb 10, 2023  
Dr. Xie

## CMSE381 - Midterm #1

First name (please write as legibly as possible within the boxes)

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

Last name

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

NetID

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

*I will adhere to the Spartan Code of Honor in completing this assignment.*

Signed: \_\_\_\_\_

1. Do not open this test booklet until you are directed to do so.
2. You will have class time (2:40-4:00pm) to complete the exam.
3. This exam is closed book, but you can use the cheatsheet provided by the instructor.
4. Throughout the test, show your work so that your reasoning is clear. Otherwise no credit will be given. **BOX** your answers. Partial credit will be given where warranted.
5. Do not spend too much time on any one problem. Read them all through first and attack them in the order that allows you to make the most progress. Good luck :P



628191E9-2E22-4DD7-AC08-168684AF224D

midterm1-b9ace

#27 Page 2 of 10

1. (15 points)

(a) Logistic regression is used for regression.

TRUE      FALSE

(b) Any model will never have training error below the irreducible error.

TRUE      FALSE

(c) Increasing your model flexibility always results in a better model.

TRUE      FALSE

(d) A logistic regression model is set up so that the odds are linear.

TRUE      FALSE

(e) Circle all of the following that would represent a qualitative variable.

Age      Year      Dog\_breed      Country\_of\_origin

Student\_(True/False)      Weight      Speed      MPG

(f) What equation would you use to evaluate the result of a regression model?

(g) What equation would you use to evaluate the result of a classification model?



2. (15 Points) I'm building a model to predict amount of a given brand of dog food eaten by a collection of dogs. I have 100 dogs eat this dog food, and I collect information on their height, weight, breed, and whether they live with another dog in the house.
- (a) List all input variables and specify whether they are quantitative or qualitative.
- (b) Say our dog breeds sampled are Huskies, Terriers, and Spaniels. Write down your prediction model.
- (c) We found that weight is a key predictor, and its relationship with the response is beyond linear. What extension can you come up with? Write down your updated model.





31E39FD9-0599-4A6D-9492-6820BD014245

midterm1-b9ace

#27 Page 4 of 10

3. (15 points)

(a) What is bias-variance tradeoff? Explain the meaning of each term in the formula.

(b) Provide a sketch of typical (squared) bias, variance, training error, test error, and Bayes (or irreducible) error curves, on a single plot, as we go from less flexible statistical learning methods towards more flexible approaches. The  $x$ -axis should represent the amount of flexibility in the method, and the  $y$ -axis should represent the values for each curve. There should be five curves. Make sure to label each one.



(c) Explain why the (i) training error and (ii) testing error lines in your drawing have the shape displayed.



4. (10 points) The data for this example come from a study by Stamey et al. (1989). The data comes from a number of clinical measures in men who were about to receive a radical prostatectomy. The goal is to predict the log of PSA (**lpsa**) (a prostate-specific antigen measured in nanograms of PSA per milliliter of blood (ng/mL)) from a number of measurements. The variables are log cancer volume (**lcavol**), log prostate weight (**lweight**), age, log of the amount of benign prostatic hyperplasia (**lbph**), seminal vesicle invasion (**svi**), log of capsular penetration (**lcp**), Gleason score (**gleason**), and percent of Gleason scores 4 or 5 (**pgg45**).

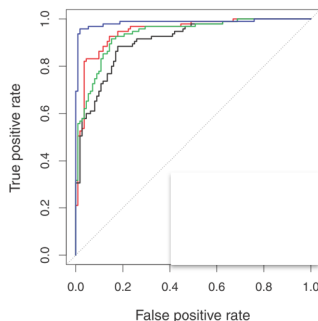
(a) Is this a case of regression or classification? Why?

(b) A linear model is fit to the data set, and the following table was returned.

Term	Coefficient	Std. Error	t-Score	p-value
Intercept	2.46	0.09	27.6	<0.00001
lcavol	0.68	0.13	5.37	<0.00001
lweight	0.26	0.1	2.75	0.00596
age	-0.14	0.1	-1.4	0.16153
lbph	0.21	0.1	2.06	0.039399
svi	0.31	0.12	2.47	0.013511
lcp	-0.29	0.15	-1.87	0.061484
gleason	-0.02	0.15	-0.15	0.880765
pgg45	0.27	0.15	1.74	0.081859

Are all the predictors useful? If yes, explain why. If not, explain what you would try to do next with the data.

5. (5 pts) We are trying to predict whether a patient will have a heart attack within one year. We have tried four different methods on a training dataset and have the following ROC curves. Which curve has the best performance on this training set? Justify your answer.





6. (15 Points) I get way too much email, so I decide to build a logistic regression model to predict whether a new incoming message is spam or not. I decide to just use a few variables:

$$\begin{aligned}X_1 &= \text{Number\_of\_sentences} \\X_2 &= \text{Number\_of\_references\_to\_a\_Nigerian\_Prince}\end{aligned}$$

and am training a logistic model to predict

$$Pr(Y = \text{spam} \mid X_1, X_2)$$

- (a) Write down the equation for the model you would train, using our standard notation with  $\beta_i$ 's.
- (b) If my trained model used  $\beta_0 = -13.1$ ,  $\beta_1 = 1.9$ , and  $\beta_2 = 6.1$ , what is the probability that a 5 sentence email with one reference to a Nigerian prince is spam? .
- (c) We know that if we miss an important email, it may cost a lot, How can you modify your model to avoid this?



7. (15 Points) Sparty generated 100 pair  $x$  and  $y$  via the following relationship:  $Y = 0.3 + 2x + x^2 + \epsilon$  but didn't tell you. You will try to find a good model to fit these data and more importantly to predict future response  $y$  when Sparty gives you another  $x$ . You will start with two models 1)  $Y = \beta_0 + \beta_1x$ ; 2)  $Y = \alpha_0 + \alpha_1x + \alpha_2x^2$ ; and 3)  $Y = \gamma_0 + \gamma_1x + \gamma_2x^2 + \gamma_3x^3$ .
- (a) If you use all 100 data to fit these three model and use the same 100 data to calculate MSE, which model will have the smallest MSE? Justify your answer
- (b) Since we focus on the ability to predict unseen data, you decide use validation set to evaluate the performance of the three models. There are two way to split the data: 1) 80 testing points and 20 training points or 2) 20 testing points and 80 training points. Which one will you choose? Justify your choice
- (c) From the in-class lab, we know validation set is not a good choice to choose model. Explain the drawback of validation set and what procedure will you use to choose the best model?



8. (10 Points) Suppose we have a data set with five predictors,  $X_1 = \text{GPA}$ ,  $X_2 = \text{IQ}$ ,  $X_3 = \text{Level}$  (1 for College and 0 for High School),  $X_4 = \text{Interaction between GPA and IQ}$ , and  $X_5 = \text{Interaction between GPA and Level}$ . The response is starting salary after graduation (in thousands of dollars). Suppose we use least squares to fit the model, and get  $\beta_0 = 50$ ,  $\beta_1 = 20$ ,  $\beta_2 = 0.07$ ,  $\beta_3 = 35$ ,  $\beta_4 = 0.01$ ,  $\beta_5 = -10$ .

(a) Predict the salary of a college graduate with IQ of 110 and a GPA of 4.0.

(b) True or false: Since the coefficient for the GPA/IQ interaction term is very small, there is very little evidence of an interaction effect. Justify your answer.

92E55DD4-680C-49B2-A12F-A0A7D1CAB5C6

midterm1-b9ace

#27 Page 9 of 10



## Scrap Paper



A86B1C78-0794-4514-96AD-B78A8C2D3DBF

midterm1-b9ace

#27 Page 10 of 10



Michigan State University  
CMSE381 - Data Science

Feb 10, 2023  
Dr. Xie

## CMSE381 - Midterm #1

First name (please write as legibly as possible within the boxes)

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

Last name

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

NetID

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

*I will adhere to the Spartan Code of Honor in completing this assignment.*

Signed: \_\_\_\_\_

1. Do not open this test booklet until you are directed to do so.
2. You will have class time (2:40-4:00pm) to complete the exam.
3. This exam is closed book, but you can use the cheatsheet provided by the instructor.
4. Throughout the test, show your work so that your reasoning is clear. Otherwise no credit will be given. **BOX** your answers. Partial credit will be given where warranted.
5. Do not spend too much time on any one problem. Read them all through first and attack them in the order that allows you to make the most progress. Good luck :P





11B13C42-8003-4896-BA9F-D197F08F8D29

midterm1-b9ace

#28 Page 2 of 10

1. (15 points)

(a) Logistic regression is used for regression.

TRUE      FALSE

(b) Any model will never have training error below the irreducible error.

TRUE      FALSE

(c) Increasing your model flexibility always results in a better model.

TRUE      FALSE

(d) A logistic regression model is set up so that the odds are linear.

TRUE      FALSE

(e) Circle all of the following that would represent a qualitative variable.

Age      Year      Dog\_breed      Country\_of\_origin

Student\_(True/False)      Weight      Speed      MPG

(f) What equation would you use to evaluate the result of a regression model?

(g) What equation would you use to evaluate the result of a classification model?



2. (15 Points) I'm building a model to predict amount of a given brand of dog food eaten by a collection of dogs. I have 100 dogs eat this dog food, and I collect information on their height, weight, breed, and whether they live with another dog in the house.
- (a) List all input variables and specify whether they are quantitative or qualitative.
- (b) Say our dog breeds sampled are Huskies, Terriers, and Spaniels. Write down your prediction model.
- (c) We found that weight is a key predictor, and its relationship with the response is beyond linear. What extension can you come up with? Write down your updated model.



69B32167-8510-47F2-9A9F-B7CB4AC19B2F

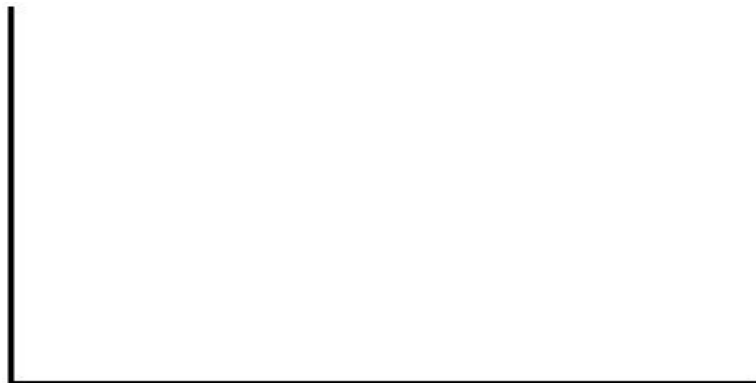
midterm1-b9ace

#28 Page 4 of 10

3. (15 points)

(a) What is bias-variance tradeoff? Explain the meaning of each term in the formula.

(b) Provide a sketch of typical (squared) bias, variance, training error, test error, and Bayes (or irreducible) error curves, on a single plot, as we go from less flexible statistical learning methods towards more flexible approaches. The  $x$ -axis should represent the amount of flexibility in the method, and the  $y$ -axis should represent the values for each curve. There should be five curves. Make sure to label each one.



(c) Explain why the (i) training error and (ii) testing error lines in your drawing have the shape displayed.



4. (10 points) The data for this example come from a study by Stamey et al. (1989). The data comes from a number of clinical measures in men who were about to receive a radical prostatectomy. The goal is to predict the log of PSA (**lpsa**) (a prostate-specific antigen measured in nanograms of PSA per milliliter of blood (ng/mL)) from a number of measurements. The variables are log cancer volume (**lcavol**), log prostate weight (**lweight**), age, log of the amount of benign prostatic hyperplasia (**lbph**), seminal vesicle invasion (**svi**), log of capsular penetration (**lcp**), Gleason score (**gleason**), and percent of Gleason scores 4 or 5 (**pgg45**).

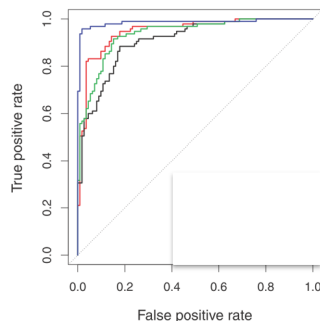
(a) Is this a case of regression or classification? Why?

(b) A linear model is fit to the data set, and the following table was returned.

Term	Coefficient	Std. Error	t-Score	p-value
Intercept	2.46	0.09	27.6	<0.00001
lcavol	0.68	0.13	5.37	<0.00001
lweight	0.26	0.1	2.75	0.00596
age	-0.14	0.1	-1.4	0.16153
lbph	0.21	0.1	2.06	0.039399
svi	0.31	0.12	2.47	0.013511
lcp	-0.29	0.15	-1.87	0.061484
gleason	-0.02	0.15	-0.15	0.880765
pgg45	0.27	0.15	1.74	0.081859

Are all the predictors useful? If yes, explain why. If not, explain what you would try to do next with the data.

5. (5 pts) We are trying to predict whether a patient will have a heart attack within one year. We have tried four different methods on a training dataset and have the following ROC curves. Which curve has the best performance on this training set? Justify your answer.





429A1B27-5F9F-40A7-8480-FCFA7FDF168D

midterm1-b9ace

#28 Page 6 of 10

6. (15 Points) I get way too much email, so I decide to build a logistic regression model to predict whether a new incoming message is spam or not. I decide to just use a few variables:

$$\begin{aligned}X_1 &= \text{Number\_of\_sentences} \\X_2 &= \text{Number\_of\_references\_to\_a\_Nigerian\_Prince}\end{aligned}$$

and am training a logistic model to predict

$$Pr(Y = \text{spam} \mid X_1, X_2)$$

- (a) Write down the equation for the model you would train, using our standard notation with  $\beta_i$ 's.
- (b) If my trained model used  $\beta_0 = -13.1$ ,  $\beta_1 = 1.9$ , and  $\beta_2 = 6.1$ , what is the probability that a 5 sentence email with one reference to a Nigerian prince is spam? .
- (c) We know that if we miss an important email, it may cost a lot, How can you modify your model to avoid this?



7. (15 Points) Sparty generated 100 pair  $x$  and  $y$  via the following relationship:  $Y = 0.3 + 2x + x^2 + \epsilon$  but didn't tell you. You will try to find a good model to fit these data and more importantly to predict future response  $y$  when Sparty gives you another  $x$ . You will start with two models 1)  $Y = \beta_0 + \beta_1x$ ; 2)  $Y = \alpha_0 + \alpha_1x + \alpha_2x^2$ ; and 3)  $Y = \gamma_0 + \gamma_1x + \gamma_2x^2 + \gamma_3x^3$ .
- (a) If you use all 100 data to fit these three model and use the same 100 data to calculate MSE, which model will have the smallest MSE? Justify your answer
- (b) Since we focus on the ability to predict unseen data, you decide use validation set to evaluate the performance of the three models. There are two way to split the data: 1) 80 testing points and 20 training points or 2) 20 testing points and 80 training points. Which one will you choose? Justify your choice
- (c) From the in-class lab, we know validation set is not a good choice to choose model. Explain the drawback of validation set and what procedure will you use to choose the best model?



35963671-DBC7-4816-9BF4-75DEFCC8FC1A

midterm1-b9ace

#28 Page 8 of 10

8. (10 Points) Suppose we have a data set with five predictors,  $X_1 = \text{GPA}$ ,  $X_2 = \text{IQ}$ ,  $X_3 = \text{Level}$  (1 for College and 0 for High School),  $X_4 = \text{Interaction between GPA and IQ}$ , and  $X_5 = \text{Interaction between GPA and Level}$ . The response is starting salary after graduation (in thousands of dollars). Suppose we use least squares to fit the model, and get  $\beta_0 = 50$ ,  $\beta_1 = 20$ ,  $\beta_2 = 0.07$ ,  $\beta_3 = 35$ ,  $\beta_4 = 0.01$ ,  $\beta_5 = -10$ .

(a) Predict the salary of a college graduate with IQ of 110 and a GPA of 4.0.

(b) True or false: Since the coefficient for the GPA/IQ interaction term is very small, there is very little evidence of an interaction effect. Justify your answer.

944C9ACD-A013-4623-AD55-91659FE59E86

midterm1-b9ace

#28 Page 9 of 10



**Scrap Paper**





AFD7938C-1348-47D1-AB1B-3E5A0939B973

midterm1-b9ace

#28 Page 10 of 10



Michigan State University  
CMSE381 - Data Science

Feb 10, 2023  
Dr. Xie

## CMSE381 - Midterm #1

First name (please write as legibly as possible within the boxes)

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

Last name

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

NetID

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

*I will adhere to the Spartan Code of Honor in completing this assignment.*

Signed: \_\_\_\_\_

1. Do not open this test booklet until you are directed to do so.
2. You will have class time (2:40-4:00pm) to complete the exam.
3. This exam is closed book, but you can use the cheatsheet provided by the instructor.
4. Throughout the test, show your work so that your reasoning is clear. Otherwise no credit will be given. **BOX** your answers. Partial credit will be given where warranted.
5. Do not spend too much time on any one problem. Read them all through first and attack them in the order that allows you to make the most progress. Good luck :P



5D76AFDB-25CB-4712-B3E5-491D396D3DC0

midterm1-b9ace

#29 Page 2 of 10

1. (15 points)

(a) Logistic regression is used for regression.

TRUE      FALSE

(b) Any model will never have training error below the irreducible error.

TRUE      FALSE

(c) Increasing your model flexibility always results in a better model.

TRUE      FALSE

(d) A logistic regression model is set up so that the odds are linear.

TRUE      FALSE

(e) Circle all of the following that would represent a qualitative variable.

Age      Year      Dog\_breed      Country\_of\_origin

Student\_(True/False)      Weight      Speed      MPG

(f) What equation would you use to evaluate the result of a regression model?

(g) What equation would you use to evaluate the result of a classification model?



2. (15 Points) I'm building a model to predict amount of a given brand of dog food eaten by a collection of dogs. I have 100 dogs eat this dog food, and I collect information on their height, weight, breed, and whether they live with another dog in the house.
- (a) List all input variables and specify whether they are quantitative or qualitative.
  - (b) Say our dog breeds sampled are Huskies, Terriers, and Spaniels. Write down your prediction model.
  - (c) We found that weight is a key predictor, and its relationship with the response is beyond linear. What extension can you come up with? Write down your updated model.



3. (15 points)

(a) What is bias-variance tradeoff? Explain the meaning of each term in the formula.

(b) Provide a sketch of typical (squared) bias, variance, training error, test error, and Bayes (or irreducible) error curves, on a single plot, as we go from less flexible statistical learning methods towards more flexible approaches. The  $x$ -axis should represent the amount of flexibility in the method, and the  $y$ -axis should represent the values for each curve. There should be five curves. Make sure to label each one.



(c) Explain why the (i) training error and (ii) testing error lines in your drawing have the shape displayed.



4. (10 points) The data for this example come from a study by Stamey et al. (1989). The data comes from a number of clinical measures in men who were about to receive a radical prostatectomy. The goal is to predict the log of PSA (**lpsa**) (a prostate-specific antigen measured in nanograms of PSA per milliliter of blood (ng/mL)) from a number of measurements. The variables are log cancer volume (**lcavol**), log prostate weight (**lweight**), age, log of the amount of benign prostatic hyperplasia (**lbph**), seminal vesicle invasion (**svi**), log of capsular penetration (**lcp**), Gleason score (**gleason**), and percent of Gleason scores 4 or 5 (**pgg45**).

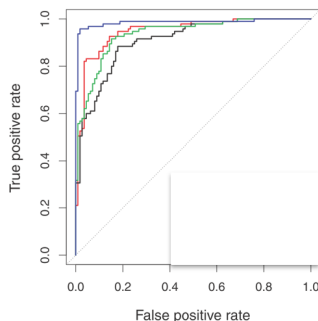
(a) Is this a case of regression or classification? Why?

(b) A linear model is fit to the data set, and the following table was returned.

Term	Coefficient	Std. Error	t-Score	p-value
Intercept	2.46	0.09	27.6	<0.00001
lcavol	0.68	0.13	5.37	<0.00001
lweight	0.26	0.1	2.75	0.00596
age	-0.14	0.1	-1.4	0.16153
lbph	0.21	0.1	2.06	0.039399
svi	0.31	0.12	2.47	0.013511
lcp	-0.29	0.15	-1.87	0.061484
gleason	-0.02	0.15	-0.15	0.880765
pgg45	0.27	0.15	1.74	0.081859

Are all the predictors useful? If yes, explain why. If not, explain what you would try to do next with the data.

5. (5 pts) We are trying to predict whether a patient will have a heart attack within one year. We have tried four different methods on a training dataset and have the following ROC curves. Which curve has the best performance on this training set? Justify your answer.





C69578F6-E745-41C2-9C2D-5E9804DEE69E

midterm1-b9ace

#29 Page 6 of 10

6. (15 Points) I get way too much email, so I decide to build a logistic regression model to predict whether a new incoming message is spam or not. I decide to just use a few variables:

$$\begin{aligned}X_1 &= \text{Number\_of\_sentences} \\X_2 &= \text{Number\_of\_references\_to\_a\_Nigerian\_Prince}\end{aligned}$$

and am training a logistic model to predict

$$Pr(Y = \text{spam} \mid X_1, X_2)$$

- (a) Write down the equation for the model you would train, using our standard notation with  $\beta_i$ 's.
- (b) If my trained model used  $\beta_0 = -13.1$ ,  $\beta_1 = 1.9$ , and  $\beta_2 = 6.1$ , what is the probability that a 5 sentence email with one reference to a Nigerian prince is spam? .
- (c) We know that if we miss an important email, it may cost a lot, How can you modify your model to avoid this?



7. (15 Points) Sparty generated 100 pair  $x$  and  $y$  via the following relationship:  $Y = 0.3 + 2x + x^2 + \epsilon$  but didn't tell you. You will try to find a good model to fit these data and more importantly to predict future response  $y$  when Sparty gives you another  $x$ . You will start with two models 1)  $Y = \beta_0 + \beta_1x$ ; 2)  $Y = \alpha_0 + \alpha_1x + \alpha_2x^2$ ; and 3)  $Y = \gamma_0 + \gamma_1x + \gamma_2x^2 + \gamma_3x^3$ .
- (a) If you use all 100 data to fit these three model and use the same 100 data to calculate MSE, which model will have the smallest MSE? Justify your answer
- (b) Since we focus on the ability to predict unseen data, you decide use validation set to evaluate the performance of the three models. There are two way to split the data: 1) 80 testing points and 20 training points or 2) 20 testing points and 80 training points. Which one will you choose? Justify your choice
- (c) From the in-class lab, we know validation set is not a good choice to choose model. Explain the drawback of validation set and what procedure will you use to choose the best model?





8. (10 Points) Suppose we have a data set with five predictors,  $X_1 = \text{GPA}$ ,  $X_2 = \text{IQ}$ ,  $X_3 = \text{Level}$  (1 for College and 0 for High School),  $X_4 = \text{Interaction between GPA and IQ}$ , and  $X_5 = \text{Interaction between GPA and Level}$ . The response is starting salary after graduation (in thousands of dollars). Suppose we use least squares to fit the model, and get  $\beta_0 = 50$ ,  $\beta_1 = 20$ ,  $\beta_2 = 0.07$ ,  $\beta_3 = 35$ ,  $\beta_4 = 0.01$ ,  $\beta_5 = -10$ .

(a) Predict the salary of a college graduate with IQ of 110 and a GPA of 4.0.

- (b) True or false: Since the coefficient for the GPA/IQ interaction term is very small, there is very little evidence of an interaction effect. Justify your answer.

89F9501B-E1B5-4B66-AF67-BA65E19597C4

midterm1-b9ace

#29 Page 9 of 10



**Scrap Paper**



1EA7DFD2-32DE-4996-85A2-1FF92EC69544

midterm1-b9ace

#29 Page 10 of 10



Michigan State University  
CMSE381 - Data Science

Feb 10, 2023  
Dr. Xie

## CMSE381 - Midterm #1

First name (please write as legibly as possible within the boxes)

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

Last name

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

NetID

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

*I will adhere to the Spartan Code of Honor in completing this assignment.*

Signed: \_\_\_\_\_

1. Do not open this test booklet until you are directed to do so.
2. You will have class time (2:40-4:00pm) to complete the exam.
3. This exam is closed book, but you can use the cheatsheet provided by the instructor.
4. Throughout the test, show your work so that your reasoning is clear. Otherwise no credit will be given. **BOX** your answers. Partial credit will be given where warranted.
5. Do not spend too much time on any one problem. Read them all through first and attack them in the order that allows you to make the most progress. Good luck :P



E221445B-C9EE-40E4-A9DA-A124B490AF56

midterm1-b9ace

#30 Page 2 of 10

1. (15 points)

(a) Logistic regression is used for regression.

TRUE      FALSE

(b) Any model will never have training error below the irreducible error.

TRUE      FALSE

(c) Increasing your model flexibility always results in a better model.

TRUE      FALSE

(d) A logistic regression model is set up so that the odds are linear.

TRUE      FALSE

(e) Circle all of the following that would represent a qualitative variable.

Age      Year      Dog\_breed      Country\_of\_origin

Student\_(True/False)      Weight      Speed      MPG

(f) What equation would you use to evaluate the result of a regression model?

(g) What equation would you use to evaluate the result of a classification model?



2. (15 Points) I'm building a model to predict amount of a given brand of dog food eaten by a collection of dogs. I have 100 dogs eat this dog food, and I collect information on their height, weight, breed, and whether they live with another dog in the house.
- (a) List all input variables and specify whether they are quantitative or qualitative.
- (b) Say our dog breeds sampled are Huskies, Terriers, and Spaniels. Write down your prediction model.
- (c) We found that weight is a key predictor, and its relationship with the response is beyond linear. What extension can you come up with? Write down your updated model.



FB95986F-C9C5-428B-962D-54F144FFAF1D

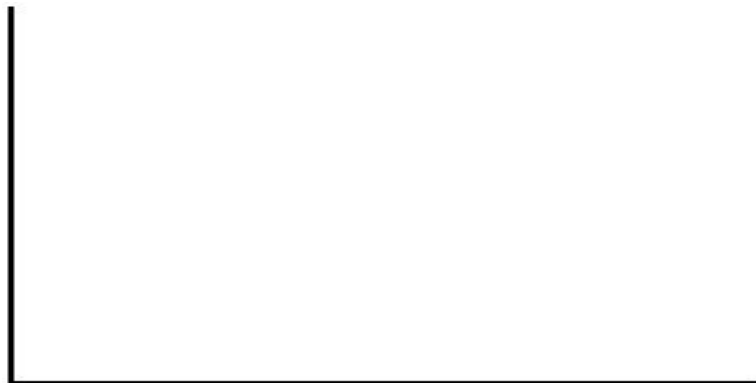
midterm1-b9ace

#30 Page 4 of 10

3. (15 points)

(a) What is bias-variance tradeoff? Explain the meaning of each term in the formula.

(b) Provide a sketch of typical (squared) bias, variance, training error, test error, and Bayes (or irreducible) error curves, on a single plot, as we go from less flexible statistical learning methods towards more flexible approaches. The  $x$ -axis should represent the amount of flexibility in the method, and the  $y$ -axis should represent the values for each curve. There should be five curves. Make sure to label each one.



(c) Explain why the (i) training error and (ii) testing error lines in your drawing have the shape displayed.



4. (10 points) The data for this example come from a study by Stamey et al. (1989). The data comes from a number of clinical measures in men who were about to receive a radical prostatectomy. The goal is to predict the log of PSA (**lpsa**) (a prostate-specific antigen measured in nanograms of PSA per milliliter of blood (ng/mL)) from a number of measurements. The variables are log cancer volume (**lcavol**), log prostate weight (**lweight**), age, log of the amount of benign prostatic hyperplasia (**lbph**), seminal vesicle invasion (**svi**), log of capsular penetration (**lcp**), Gleason score (**gleason**), and percent of Gleason scores 4 or 5 (**pgg45**).

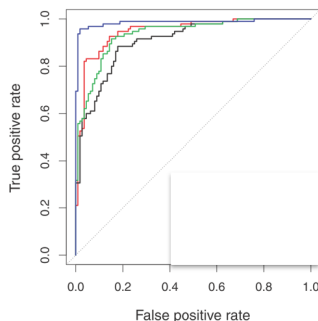
(a) Is this a case of regression or classification? Why?

(b) A linear model is fit to the data set, and the following table was returned.

Term	Coefficient	Std. Error	t-Score	p-value
Intercept	2.46	0.09	27.6	<0.00001
lcavol	0.68	0.13	5.37	<0.00001
lweight	0.26	0.1	2.75	0.00596
age	-0.14	0.1	-1.4	0.16153
lbph	0.21	0.1	2.06	0.039399
svi	0.31	0.12	2.47	0.013511
lcp	-0.29	0.15	-1.87	0.061484
gleason	-0.02	0.15	-0.15	0.880765
pgg45	0.27	0.15	1.74	0.081859

Are all the predictors useful? If yes, explain why. If not, explain what you would try to do next with the data.

5. (5 pts) We are trying to predict whether a patient will have a heart attack within one year. We have tried four different methods on a training dataset and have the following ROC curves. Which curve has the best performance on this training set? Justify your answer.







6. (15 Points) I get way too much email, so I decide to build a logistic regression model to predict whether a new incoming message is spam or not. I decide to just use a few variables:

$$\begin{aligned}X_1 &= \text{Number\_of\_sentences} \\X_2 &= \text{Number\_of\_references\_to\_a\_Nigerian\_Prince}\end{aligned}$$

and am training a logistic model to predict

$$Pr(Y = \text{spam} \mid X_1, X_2)$$

- (a) Write down the equation for the model you would train, using our standard notation with  $\beta_i$ 's.
- (b) If my trained model used  $\beta_0 = -13.1$ ,  $\beta_1 = 1.9$ , and  $\beta_2 = 6.1$ , what is the probability that a 5 sentence email with one reference to a Nigerian prince is spam? .
- (c) We know that if we miss an important email, it may cost a lot, How can you modify your model to avoid this?



7. (15 Points) Sparty generated 100 pair  $x$  and  $y$  via the following relationship:  $Y = 0.3 + 2x + x^2 + \epsilon$  but didn't tell you. You will try to find a good model to fit these data and more importantly to predict future response  $y$  when Sparty gives you another  $x$ . You will start with two models 1)  $Y = \beta_0 + \beta_1x$ ; 2)  $Y = \alpha_0 + \alpha_1x + \alpha_2x^2$ ; and 3)  $Y = \gamma_0 + \gamma_1x + \gamma_2x^2 + \gamma_3x^3$ .
- (a) If you use all 100 data to fit these three model and use the same 100 data to calculate MSE, which model will have the smallest MSE? Justify your answer
- (b) Since we focus on the ability to predict unseen data, you decide use validation set to evaluate the performance of the three models. There are two way to split the data: 1) 80 testing points and 20 training points or 2) 20 testing points and 80 training points. Which one will you choose? Justify your choice
- (c) From the in-class lab, we know validation set is not a good choice to choose model. Explain the drawback of validation set and what procedure will you use to choose the best model?



FBDFAE62-AAC1-4B51-81F1-6823D3BD1DC8

midterm1-b9ace

#30 Page 8 of 10

8. (10 Points) Suppose we have a data set with five predictors,  $X_1 = \text{GPA}$ ,  $X_2 = \text{IQ}$ ,  $X_3 = \text{Level}$  (1 for College and 0 for High School),  $X_4 = \text{Interaction between GPA and IQ}$ , and  $X_5 = \text{Interaction between GPA and Level}$ . The response is starting salary after graduation (in thousands of dollars). Suppose we use least squares to fit the model, and get  $\beta_0 = 50$ ,  $\beta_1 = 20$ ,  $\beta_2 = 0.07$ ,  $\beta_3 = 35$ ,  $\beta_4 = 0.01$ ,  $\beta_5 = -10$ .

(a) Predict the salary of a college graduate with IQ of 110 and a GPA of 4.0.

- (b) True or false: Since the coefficient for the GPA/IQ interaction term is very small, there is very little evidence of an interaction effect. Justify your answer.

6FFF41BA-0AB5-4664-BF37-6FBCA6EA20CC

midterm1-b9ace

#30 Page 9 of 10



## Scrap Paper



00F22792-F76D-4EB7-970B-929F4914ED69

midterm1-b9ace

#30 Page 10 of 10



Michigan State University  
CMSE381 - Data Science

Feb 10, 2023  
Dr. Xie

## CMSE381 - Midterm #1

First name (please write as legibly as possible within the boxes)

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

Last name

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

NetID

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

*I will adhere to the Spartan Code of Honor in completing this assignment.*

Signed: \_\_\_\_\_

1. Do not open this test booklet until you are directed to do so.
2. You will have class time (2:40-4:00pm) to complete the exam.
3. This exam is closed book, but you can use the cheatsheet provided by the instructor.
4. Throughout the test, show your work so that your reasoning is clear. Otherwise no credit will be given. **BOX** your answers. Partial credit will be given where warranted.
5. Do not spend too much time on any one problem. Read them all through first and attack them in the order that allows you to make the most progress. Good luck :P



F67652CA-4D5F-4C60-A385-913DA263679E

midterm1-b9ace

#31 Page 2 of 10

1. (15 points)

(a) Logistic regression is used for regression.

TRUE      FALSE

(b) Any model will never have training error below the irreducible error.

TRUE      FALSE

(c) Increasing your model flexibility always results in a better model.

TRUE      FALSE

(d) A logistic regression model is set up so that the odds are linear.

TRUE      FALSE

(e) Circle all of the following that would represent a qualitative variable.

Age      Year      Dog\_breed      Country\_of\_origin

Student\_(True/False)      Weight      Speed      MPG

(f) What equation would you use to evaluate the result of a regression model?

(g) What equation would you use to evaluate the result of a classification model?



2. (15 Points) I'm building a model to predict amount of a given brand of dog food eaten by a collection of dogs. I have 100 dogs eat this dog food, and I collect information on their height, weight, breed, and whether they live with another dog in the house.
- (a) List all input variables and specify whether they are quantitative or qualitative.
  - (b) Say our dog breeds sampled are Huskies, Terriers, and Spaniels. Write down your prediction model.
  - (c) We found that weight is a key predictor, and its relationship with the response is beyond linear. What extension can you come up with? Write down your updated model.





3. (15 points)

(a) What is bias-variance tradeoff? Explain the meaning of each term in the formula.

(b) Provide a sketch of typical (squared) bias, variance, training error, test error, and Bayes (or irreducible) error curves, on a single plot, as we go from less flexible statistical learning methods towards more flexible approaches. The  $x$ -axis should represent the amount of flexibility in the method, and the  $y$ -axis should represent the values for each curve. There should be five curves. Make sure to label each one.



(c) Explain why the (i) training error and (ii) testing error lines in your drawing have the shape displayed.



4. (10 points) The data for this example come from a study by Stamey et al. (1989). The data comes from a number of clinical measures in men who were about to receive a radical prostatectomy. The goal is to predict the log of PSA (**lpsa**) (a prostate-specific antigen measured in nanograms of PSA per milliliter of blood (ng/mL)) from a number of measurements. The variables are log cancer volume (**lcavol**), log prostate weight (**lweight**), age, log of the amount of benign prostatic hyperplasia (**lbph**), seminal vesicle invasion (**svi**), log of capsular penetration (**lcp**), Gleason score (**gleason**), and percent of Gleason scores 4 or 5 (**pgg45**).

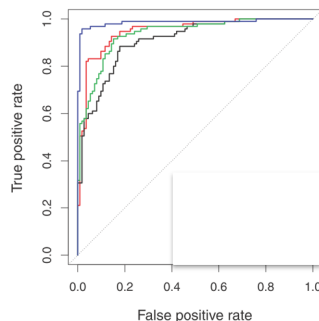
(a) Is this a case of regression or classification? Why?

(b) A linear model is fit to the data set, and the following table was returned.

Term	Coefficient	Std. Error	t-Score	p-value
Intercept	2.46	0.09	27.6	<0.00001
lcavol	0.68	0.13	5.37	<0.00001
lweight	0.26	0.1	2.75	0.00596
age	-0.14	0.1	-1.4	0.16153
lbph	0.21	0.1	2.06	0.039399
svi	0.31	0.12	2.47	0.013511
lcp	-0.29	0.15	-1.87	0.061484
gleason	-0.02	0.15	-0.15	0.880765
pgg45	0.27	0.15	1.74	0.081859

Are all the predictors useful? If yes, explain why. If not, explain what you would try to do next with the data.

5. (5 pts) We are trying to predict whether a patient will have a heart attack within one year. We have tried four different methods on a training dataset and have the following ROC curves. Which curve has the best performance on this training set? Justify your answer.





D2FF281C-32D0-498F-A0F7-3E04F3B0724D

midterm1-b9ace

#31 Page 6 of 10

6. (15 Points) I get way too much email, so I decide to build a logistic regression model to predict whether a new incoming message is spam or not. I decide to just use a few variables:

$$\begin{aligned}X_1 &= \text{Number\_of\_sentences} \\X_2 &= \text{Number\_of\_references\_to\_a\_Nigerian\_Prince}\end{aligned}$$

and am training a logistic model to predict

$$Pr(Y = \text{spam} \mid X_1, X_2)$$

- (a) Write down the equation for the model you would train, using our standard notation with  $\beta_i$ 's.
- (b) If my trained model used  $\beta_0 = -13.1$ ,  $\beta_1 = 1.9$ , and  $\beta_2 = 6.1$ , what is the probability that a 5 sentence email with one reference to a Nigerian prince is spam? .
- (c) We know that if we miss an important email, it may cost a lot, How can you modify your model to avoid this?



7. (15 Points) Sparty generated 100 pair  $x$  and  $y$  via the following relationship:  $Y = 0.3 + 2x + x^2 + \epsilon$  but didn't tell you. You will try to find a good model to fit these data and more importantly to predict future response  $y$  when Sparty gives you another  $x$ . You will start with two models 1)  $Y = \beta_0 + \beta_1x$ ; 2)  $Y = \alpha_0 + \alpha_1x + \alpha_2x^2$ ; and 3)  $Y = \gamma_0 + \gamma_1x + \gamma_2x^2 + \gamma_3x^3$ .
- (a) If you use all 100 data to fit these three model and use the same 100 data to calculate MSE, which model will have the smallest MSE? Justify your answer
- (b) Since we focus on the ability to predict unseen data, you decide use validation set to evaluate the performance of the three models. There are two way to split the data: 1) 80 testing points and 20 training points or 2) 20 testing points and 80 training points. Which one will you choose? Justify your choice
- (c) From the in-class lab, we know validation set is not a good choice to choose model. Explain the drawback of validation set and what procedure will you use to choose the best model?



8. (10 Points) Suppose we have a data set with five predictors,  $X_1 = \text{GPA}$ ,  $X_2 = \text{IQ}$ ,  $X_3 = \text{Level}$  (1 for College and 0 for High School),  $X_4 = \text{Interaction between GPA and IQ}$ , and  $X_5 = \text{Interaction between GPA and Level}$ . The response is starting salary after graduation (in thousands of dollars). Suppose we use least squares to fit the model, and get  $\beta_0 = 50$ ,  $\beta_1 = 20$ ,  $\beta_2 = 0.07$ ,  $\beta_3 = 35$ ,  $\beta_4 = 0.01$ ,  $\beta_5 = -10$ .

(a) Predict the salary of a college graduate with IQ of 110 and a GPA of 4.0.

- (b) True or false: Since the coefficient for the GPA/IQ interaction term is very small, there is very little evidence of an interaction effect. Justify your answer.

36F0DEE4-BDE4-4030-AB1F-F70F39796651

midterm1-b9ace

#31 Page 9 of 10



**Scrap Paper**



3D2F845F-7486-4E32-93DB-2A975616DF3D

midterm1-b9ace

#31 Page 10 of 10



Michigan State University  
CMSE381 - Data Science

Feb 10, 2023  
Dr. Xie

## CMSE381 - Midterm #1

First name (please write as legibly as possible within the boxes)

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

Last name

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

NetID

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

*I will adhere to the Spartan Code of Honor in completing this assignment.*

Signed: \_\_\_\_\_

1. Do not open this test booklet until you are directed to do so.
2. You will have class time (2:40-4:00pm) to complete the exam.
3. This exam is closed book, but you can use the cheatsheet provided by the instructor.
4. Throughout the test, show your work so that your reasoning is clear. Otherwise no credit will be given. **BOX** your answers. Partial credit will be given where warranted.
5. Do not spend too much time on any one problem. Read them all through first and attack them in the order that allows you to make the most progress. Good luck :P





8B3E6725-2044-49BF-9970-80F19EE8157F

midterm1-b9ace

#32 Page 2 of 10

1. (15 points)

(a) Logistic regression is used for regression.

TRUE      FALSE

(b) Any model will never have training error below the irreducible error.

TRUE      FALSE

(c) Increasing your model flexibility always results in a better model.

TRUE      FALSE

(d) A logistic regression model is set up so that the odds are linear.

TRUE      FALSE

(e) Circle all of the following that would represent a qualitative variable.

Age      Year      Dog\_breed      Country\_of\_origin

Student\_(True/False)      Weight      Speed      MPG

(f) What equation would you use to evaluate the result of a regression model?

(g) What equation would you use to evaluate the result of a classification model?



2. (15 Points) I'm building a model to predict amount of a given brand of dog food eaten by a collection of dogs. I have 100 dogs eat this dog food, and I collect information on their height, weight, breed, and whether they live with another dog in the house.
- (a) List all input variables and specify whether they are quantitative or qualitative.
- (b) Say our dog breeds sampled are Huskies, Terriers, and Spaniels. Write down your prediction model.
- (c) We found that weight is a key predictor, and its relationship with the response is beyond linear. What extension can you come up with? Write down your updated model.



A2DC8E36-0BB2-42D7-8734-28CF754341BE

midterm1-b9ace

#32 Page 4 of 10

3. (15 points)

(a) What is bias-variance tradeoff? Explain the meaning of each term in the formula.

(b) Provide a sketch of typical (squared) bias, variance, training error, test error, and Bayes (or irreducible) error curves, on a single plot, as we go from less flexible statistical learning methods towards more flexible approaches. The  $x$ -axis should represent the amount of flexibility in the method, and the  $y$ -axis should represent the values for each curve. There should be five curves. Make sure to label each one.



(c) Explain why the (i) training error and (ii) testing error lines in your drawing have the shape displayed.



4. (10 points) The data for this example come from a study by Stamey et al. (1989). The data comes from a number of clinical measures in men who were about to receive a radical prostatectomy. The goal is to predict the log of PSA (**lpsa**) (a prostate-specific antigen measured in nanograms of PSA per milliliter of blood (ng/mL)) from a number of measurements. The variables are log cancer volume (**lcavol**), log prostate weight (**lweight**), age, log of the amount of benign prostatic hyperplasia (**lbph**), seminal vesicle invasion (**svi**), log of capsular penetration (**lcp**), Gleason score (**gleason**), and percent of Gleason scores 4 or 5 (**pgg45**).

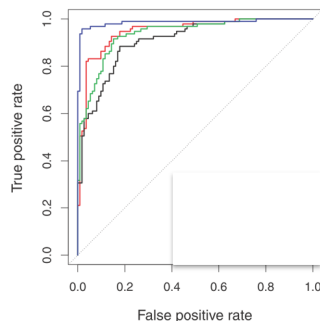
(a) Is this a case of regression or classification? Why?

(b) A linear model is fit to the data set, and the following table was returned.

Term	Coefficient	Std. Error	t-Score	p-value
Intercept	2.46	0.09	27.6	<0.00001
lcavol	0.68	0.13	5.37	<0.00001
lweight	0.26	0.1	2.75	0.00596
age	-0.14	0.1	-1.4	0.16153
lbph	0.21	0.1	2.06	0.039399
svi	0.31	0.12	2.47	0.013511
lcp	-0.29	0.15	-1.87	0.061484
gleason	-0.02	0.15	-0.15	0.880765
pgg45	0.27	0.15	1.74	0.081859

Are all the predictors useful? If yes, explain why. If not, explain what you would try to do next with the data.

5. (5 pts) We are trying to predict whether a patient will have a heart attack within one year. We have tried four different methods on a training dataset and have the following ROC curves. Which curve has the best performance on this training set? Justify your answer.





6. (15 Points) I get way too much email, so I decide to build a logistic regression model to predict whether a new incoming message is spam or not. I decide to just use a few variables:

$$\begin{aligned}X_1 &= \text{Number\_of\_sentences} \\X_2 &= \text{Number\_of\_references\_to\_a\_Nigerian\_Prince}\end{aligned}$$

and am training a logistic model to predict

$$Pr(Y = \text{spam} \mid X_1, X_2)$$

- (a) Write down the equation for the model you would train, using our standard notation with  $\beta_i$ 's.
- (b) If my trained model used  $\beta_0 = -13.1$ ,  $\beta_1 = 1.9$ , and  $\beta_2 = 6.1$ , what is the probability that a 5 sentence email with one reference to a Nigerian prince is spam? .
- (c) We know that if we miss an important email, it may cost a lot, How can you modify your model to avoid this?



7. (15 Points) Sparty generated 100 pair  $x$  and  $y$  via the following relationship:  $Y = 0.3 + 2x + x^2 + \epsilon$  but didn't tell you. You will try to find a good model to fit these data and more importantly to predict future response  $y$  when Sparty gives you another  $x$ . You will start with two models 1)  $Y = \beta_0 + \beta_1x$ ; 2)  $Y = \alpha_0 + \alpha_1x + \alpha_2x^2$ ; and 3)  $Y = \gamma_0 + \gamma_1x + \gamma_2x^2 + \gamma_3x^3$ .
- (a) If you use all 100 data to fit these three model and use the same 100 data to calculate MSE, which model will have the smallest MSE? Justify your answer
- (b) Since we focus on the ability to predict unseen data, you decide use validation set to evaluate the performance of the three models. There are two way to split the data: 1) 80 testing points and 20 training points or 2) 20 testing points and 80 training points. Which one will you choose? Justify your choice
- (c) From the in-class lab, we know validation set is not a good choice to choose model. Explain the drawback of validation set and what procedure will you use to choose the best model?



82F8510A-7FFF-4D1D-A0E2-412D8EAD4B14

midterm1-b9ace

#32 Page 8 of 10

8. (10 Points) Suppose we have a data set with five predictors,  $X_1 = \text{GPA}$ ,  $X_2 = \text{IQ}$ ,  $X_3 = \text{Level}$  (1 for College and 0 for High School),  $X_4 = \text{Interaction between GPA and IQ}$ , and  $X_5 = \text{Interaction between GPA and Level}$ . The response is starting salary after graduation (in thousands of dollars). Suppose we use least squares to fit the model, and get  $\beta_0 = 50$ ,  $\beta_1 = 20$ ,  $\beta_2 = 0.07$ ,  $\beta_3 = 35$ ,  $\beta_4 = 0.01$ ,  $\beta_5 = -10$ .

(a) Predict the salary of a college graduate with IQ of 110 and a GPA of 4.0.

- (b) True or false: Since the coefficient for the GPA/IQ interaction term is very small, there is very little evidence of an interaction effect. Justify your answer.

BDD2914E-5AD6-4C48-BE20-AB41DB4CEAA4

midterm1-b9ace

#32 Page 9 of 10



## Scrap Paper





57F172BE-231B-4DF3-ACD6-3A47107BD24B

midterm1-b9ace

#32 Page 10 of 10



Michigan State University  
CMSE381 - Data Science

Feb 10, 2023  
Dr. Xie

## CMSE381 - Midterm #1

First name (please write as legibly as possible within the boxes)

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

Last name

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

NetID

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

*I will adhere to the Spartan Code of Honor in completing this assignment.*

Signed: \_\_\_\_\_

1. Do not open this test booklet until you are directed to do so.
2. You will have class time (2:40-4:00pm) to complete the exam.
3. This exam is closed book, but you can use the cheatsheet provided by the instructor.
4. Throughout the test, show your work so that your reasoning is clear. Otherwise no credit will be given. **BOX** your answers. Partial credit will be given where warranted.
5. Do not spend too much time on any one problem. Read them all through first and attack them in the order that allows you to make the most progress. Good luck :P



BE6E435F-7F1D-4EB9-9B00-3760EE7F4CE2

midterm1-b9ace

#33 Page 2 of 10

1. (15 points)

(a) Logistic regression is used for regression.

TRUE      FALSE

(b) Any model will never have training error below the irreducible error.

TRUE      FALSE

(c) Increasing your model flexibility always results in a better model.

TRUE      FALSE

(d) A logistic regression model is set up so that the odds are linear.

TRUE      FALSE

(e) Circle all of the following that would represent a qualitative variable.

Age      Year      Dog\_breed      Country\_of\_origin

Student\_(True/False)      Weight      Speed      MPG

(f) What equation would you use to evaluate the result of a regression model?

(g) What equation would you use to evaluate the result of a classification model?



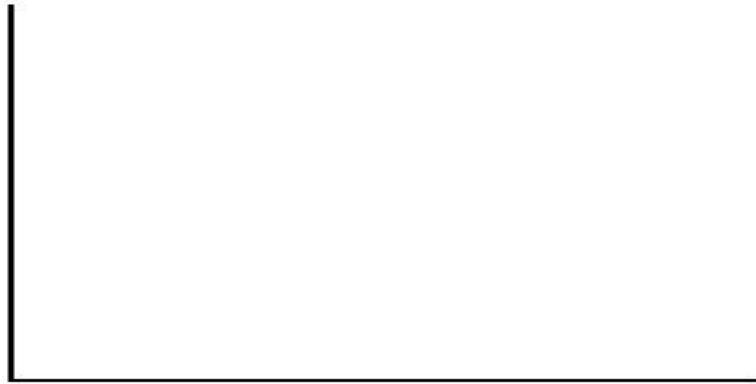
2. (15 Points) I'm building a model to predict amount of a given brand of dog food eaten by a collection of dogs. I have 100 dogs eat this dog food, and I collect information on their height, weight, breed, and whether they live with another dog in the house.
- (a) List all input variables and specify whether they are quantitative or qualitative.
- (b) Say our dog breeds sampled are Huskies, Terriers, and Spaniels. Write down your prediction model.
- (c) We found that weight is a key predictor, and its relationship with the response is beyond linear. What extension can you come up with? Write down your updated model.



3. (15 points)

(a) What is bias-variance tradeoff? Explain the meaning of each term in the formula.

(b) Provide a sketch of typical (squared) bias, variance, training error, test error, and Bayes (or irreducible) error curves, on a single plot, as we go from less flexible statistical learning methods towards more flexible approaches. The  $x$ -axis should represent the amount of flexibility in the method, and the  $y$ -axis should represent the values for each curve. There should be five curves. Make sure to label each one.



(c) Explain why the (i) training error and (ii) testing error lines in your drawing have the shape displayed.



4. (10 points) The data for this example come from a study by Stamey et al. (1989). The data comes from a number of clinical measures in men who were about to receive a radical prostatectomy. The goal is to predict the log of PSA (**lpsa**) (a prostate-specific antigen measured in nanograms of PSA per milliliter of blood (ng/mL)) from a number of measurements. The variables are log cancer volume (**lcavol**), log prostate weight (**lweight**), age, log of the amount of benign prostatic hyperplasia (**lbph**), seminal vesicle invasion (**svi**), log of capsular penetration (**lcp**), Gleason score (**gleason**), and percent of Gleason scores 4 or 5 (**pgg45**).

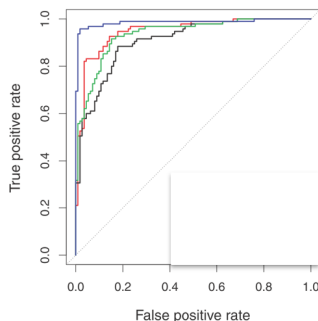
(a) Is this a case of regression or classification? Why?

(b) A linear model is fit to the data set, and the following table was returned.

Term	Coefficient	Std. Error	t-Score	p-value
Intercept	2.46	0.09	27.6	<0.00001
lcavol	0.68	0.13	5.37	<0.00001
lweight	0.26	0.1	2.75	0.00596
age	-0.14	0.1	-1.4	0.16153
lbph	0.21	0.1	2.06	0.039399
svi	0.31	0.12	2.47	0.013511
lcp	-0.29	0.15	-1.87	0.061484
gleason	-0.02	0.15	-0.15	0.880765
pgg45	0.27	0.15	1.74	0.081859

Are all the predictors useful? If yes, explain why. If not, explain what you would try to do next with the data.

5. (5 pts) We are trying to predict whether a patient will have a heart attack within one year. We have tried four different methods on a training dataset and have the following ROC curves. Which curve has the best performance on this training set? Justify your answer.





1126C387-E430-4E2E-85CD-2E25FD8C7F18

midterm1-b9ace

#33 Page 6 of 10

6. (15 Points) I get way too much email, so I decide to build a logistic regression model to predict whether a new incoming message is spam or not. I decide to just use a few variables:

$$\begin{aligned}X_1 &= \text{Number\_of\_sentences} \\X_2 &= \text{Number\_of\_references\_to\_a\_Nigerian\_Prince}\end{aligned}$$

and am training a logistic model to predict

$$Pr(Y = \text{spam} \mid X_1, X_2)$$

- (a) Write down the equation for the model you would train, using our standard notation with  $\beta_i$ 's.
- (b) If my trained model used  $\beta_0 = -13.1$ ,  $\beta_1 = 1.9$ , and  $\beta_2 = 6.1$ , what is the probability that a 5 sentence email with one reference to a Nigerian prince is spam? .
- (c) We know that if we miss an important email, it may cost a lot, How can you modify your model to avoid this?



7. (15 Points) Sparty generated 100 pair  $x$  and  $y$  via the following relationship:  $Y = 0.3 + 2x + x^2 + \epsilon$  but didn't tell you. You will try to find a good model to fit these data and more importantly to predict future response  $y$  when Sparty gives you another  $x$ . You will start with two models 1)  $Y = \beta_0 + \beta_1x$ ; 2)  $Y = \alpha_0 + \alpha_1x + \alpha_2x^2$ ; and 3)  $Y = \gamma_0 + \gamma_1x + \gamma_2x^2 + \gamma_3x^3$ .
- (a) If you use all 100 data to fit these three model and use the same 100 data to calculate MSE, which model will have the smallest MSE? Justify your answer
- (b) Since we focus on the ability to predict unseen data, you decide use validation set to evaluate the performance of the three models. There are two way to split the data: 1) 80 testing points and 20 training points or 2) 20 testing points and 80 training points. Which one will you choose? Justify your choice
- (c) From the in-class lab, we know validation set is not a good choice to choose model. Explain the drawback of validation set and what procedure will you use to choose the best model?





8D8D10C8-56D0-47E3-81EE-B6E091F34025

midterm1-b9ace

#33 Page 8 of 10

8. (10 Points) Suppose we have a data set with five predictors,  $X_1 = \text{GPA}$ ,  $X_2 = \text{IQ}$ ,  $X_3 = \text{Level}$  (1 for College and 0 for High School),  $X_4 = \text{Interaction between GPA and IQ}$ , and  $X_5 = \text{Interaction between GPA and Level}$ . The response is starting salary after graduation (in thousands of dollars). Suppose we use least squares to fit the model, and get  $\beta_0 = 50$ ,  $\beta_1 = 20$ ,  $\beta_2 = 0.07$ ,  $\beta_3 = 35$ ,  $\beta_4 = 0.01$ ,  $\beta_5 = -10$ .

(a) Predict the salary of a college graduate with IQ of 110 and a GPA of 4.0.

- (b) True or false: Since the coefficient for the GPA/IQ interaction term is very small, there is very little evidence of an interaction effect. Justify your answer.

E451978D-7FF2-47D5-ADEB-A94FF82EA21F

midterm1-b9ace

#33 Page 9 of 10



## Scrap Paper



4FC2BD43-3C1F-419E-94D8-F04359F996DF

midterm1-b9ace

#33 Page 10 of 10



Michigan State University  
CMSE381 - Data Science

Feb 10, 2023  
Dr. Xie

## CMSE381 - Midterm #1

First name (please write as legibly as possible within the boxes)

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

Last name

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

NetID

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

*I will adhere to the Spartan Code of Honor in completing this assignment.*

Signed: \_\_\_\_\_

1. Do not open this test booklet until you are directed to do so.
2. You will have class time (2:40-4:00pm) to complete the exam.
3. This exam is closed book, but you can use the cheatsheet provided by the instructor.
4. Throughout the test, show your work so that your reasoning is clear. Otherwise no credit will be given. **BOX** your answers. Partial credit will be given where warranted.
5. Do not spend too much time on any one problem. Read them all through first and attack them in the order that allows you to make the most progress. Good luck :P



1. (15 points)

(a) Logistic regression is used for regression.

TRUE      FALSE

(b) Any model will never have training error below the irreducible error.

TRUE      FALSE

(c) Increasing your model flexibility always results in a better model.

TRUE      FALSE

(d) A logistic regression model is set up so that the odds are linear.

TRUE      FALSE

(e) Circle all of the following that would represent a qualitative variable.

Age      Year      Dog\_breed      Country\_of\_origin

Student\_(True/False)      Weight      Speed      MPG

(f) What equation would you use to evaluate the result of a regression model?

(g) What equation would you use to evaluate the result of a classification model?



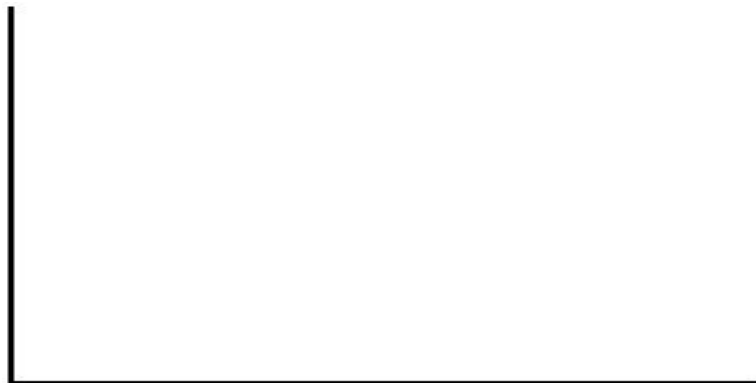
2. (15 Points) I'm building a model to predict amount of a given brand of dog food eaten by a collection of dogs. I have 100 dogs eat this dog food, and I collect information on their height, weight, breed, and whether they live with another dog in the house.
- (a) List all input variables and specify whether they are quantitative or qualitative.
  - (b) Say our dog breeds sampled are Huskies, Terriers, and Spaniels. Write down your prediction model.
  - (c) We found that weight is a key predictor, and its relationship with the response is beyond linear. What extension can you come up with? Write down your updated model.



3. (15 points)

(a) What is bias-variance tradeoff? Explain the meaning of each term in the formula.

(b) Provide a sketch of typical (squared) bias, variance, training error, test error, and Bayes (or irreducible) error curves, on a single plot, as we go from less flexible statistical learning methods towards more flexible approaches. The  $x$ -axis should represent the amount of flexibility in the method, and the  $y$ -axis should represent the values for each curve. There should be five curves. Make sure to label each one.



(c) Explain why the (i) training error and (ii) testing error lines in your drawing have the shape displayed.



4. (10 points) The data for this example come from a study by Stamey et al. (1989). The data comes from a number of clinical measures in men who were about to receive a radical prostatectomy. The goal is to predict the log of PSA (**lpsa**) (a prostate-specific antigen measured in nanograms of PSA per milliliter of blood (ng/mL)) from a number of measurements. The variables are log cancer volume (**lcavol**), log prostate weight (**lweight**), age, log of the amount of benign prostatic hyperplasia (**lbph**), seminal vesicle invasion (**svi**), log of capsular penetration (**lcp**), Gleason score (**gleason**), and percent of Gleason scores 4 or 5 (**pgg45**).

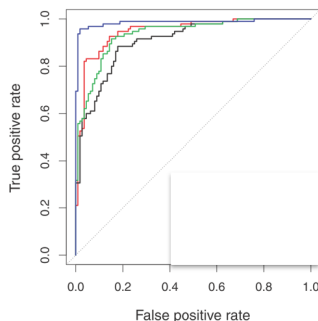
(a) Is this a case of regression or classification? Why?

(b) A linear model is fit to the data set, and the following table was returned.

Term	Coefficient	Std. Error	t-Score	p-value
Intercept	2.46	0.09	27.6	<0.00001
lcavol	0.68	0.13	5.37	<0.00001
lweight	0.26	0.1	2.75	0.00596
age	-0.14	0.1	-1.4	0.16153
lbph	0.21	0.1	2.06	0.039399
svi	0.31	0.12	2.47	0.013511
lcp	-0.29	0.15	-1.87	0.061484
gleason	-0.02	0.15	-0.15	0.880765
pgg45	0.27	0.15	1.74	0.081859

Are all the predictors useful? If yes, explain why. If not, explain what you would try to do next with the data.

5. (5 pts) We are trying to predict whether a patient will have a heart attack within one year. We have tried four different methods on a training dataset and have the following ROC curves. Which curve has the best performance on this training set? Justify your answer.







6. (15 Points) I get way too much email, so I decide to build a logistic regression model to predict whether a new incoming message is spam or not. I decide to just use a few variables:

$$\begin{aligned}X_1 &= \text{Number\_of\_sentences} \\X_2 &= \text{Number\_of\_references\_to\_a\_Nigerian\_Prince}\end{aligned}$$

and am training a logistic model to predict

$$Pr(Y = \text{spam} \mid X_1, X_2)$$

- (a) Write down the equation for the model you would train, using our standard notation with  $\beta_i$ 's.
- (b) If my trained model used  $\beta_0 = -13.1$ ,  $\beta_1 = 1.9$ , and  $\beta_2 = 6.1$ , what is the probability that a 5 sentence email with one reference to a Nigerian prince is spam? .
- (c) We know that if we miss an important email, it may cost a lot, How can you modify your model to avoid this?



7. (15 Points) Sparty generated 100 pair  $x$  and  $y$  via the following relationship:  $Y = 0.3 + 2x + x^2 + \epsilon$  but didn't tell you. You will try to find a good model to fit these data and more importantly to predict future response  $y$  when Sparty gives you another  $x$ . You will start with two models 1)  $Y = \beta_0 + \beta_1x$ ; 2)  $Y = \alpha_0 + \alpha_1x + \alpha_2x^2$ ; and 3)  $Y = \gamma_0 + \gamma_1x + \gamma_2x^2 + \gamma_3x^3$ .
- (a) If you use all 100 data to fit these three model and use the same 100 data to calculate MSE, which model will have the smallest MSE? Justify your answer
- (b) Since we focus on the ability to predict unseen data, you decide use validation set to evaluate the performance of the three models. There are two way to split the data: 1) 80 testing points and 20 training points or 2) 20 testing points and 80 training points. Which one will you choose? Justify your choice
- (c) From the in-class lab, we know validation set is not a good choice to choose model. Explain the drawback of validation set and what procedure will you use to choose the best model?



949FB2A3-B025-4DF8-A919-5B3BBB16A2D0

midterm1-b9ace

#34 Page 8 of 10

8. (10 Points) Suppose we have a data set with five predictors,  $X_1 = \text{GPA}$ ,  $X_2 = \text{IQ}$ ,  $X_3 = \text{Level}$  (1 for College and 0 for High School),  $X_4 = \text{Interaction between GPA and IQ}$ , and  $X_5 = \text{Interaction between GPA and Level}$ . The response is starting salary after graduation (in thousands of dollars). Suppose we use least squares to fit the model, and get  $\beta_0 = 50$ ,  $\beta_1 = 20$ ,  $\beta_2 = 0.07$ ,  $\beta_3 = 35$ ,  $\beta_4 = 0.01$ ,  $\beta_5 = -10$ .

(a) Predict the salary of a college graduate with IQ of 110 and a GPA of 4.0.

- (b) True or false: Since the coefficient for the GPA/IQ interaction term is very small, there is very little evidence of an interaction effect. Justify your answer.

255AA273-4493-475E-A1E0-B21FC50F2BC5

midterm1-b9ace

#34 Page 9 of 10



## Scrap Paper



80E757B1-DCE2-4BAF-B03B-E78DB5B69B54

midterm1-b9ace

#34 Page 10 of 10



Michigan State University  
CMSE381 - Data Science

Feb 10, 2023  
Dr. Xie

## CMSE381 - Midterm #1

First name (please write as legibly as possible within the boxes)

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

Last name

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

NetID

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

*I will adhere to the Spartan Code of Honor in completing this assignment.*

Signed: \_\_\_\_\_

1. Do not open this test booklet until you are directed to do so.
2. You will have class time (2:40-4:00pm) to complete the exam.
3. This exam is closed book, but you can use the cheatsheet provided by the instructor.
4. Throughout the test, show your work so that your reasoning is clear. Otherwise no credit will be given. **BOX** your answers. Partial credit will be given where warranted.
5. Do not spend too much time on any one problem. Read them all through first and attack them in the order that allows you to make the most progress. Good luck :P



E213F1CF-4EBD-4143-839F-484DD8E71B06

midterm1-b9ace

#35 Page 2 of 10

1. (15 points)

(a) Logistic regression is used for regression.

TRUE      FALSE

(b) Any model will never have training error below the irreducible error.

TRUE      FALSE

(c) Increasing your model flexibility always results in a better model.

TRUE      FALSE

(d) A logistic regression model is set up so that the odds are linear.

TRUE      FALSE

(e) Circle all of the following that would represent a qualitative variable.

Age      Year      Dog\_breed      Country\_of\_origin

Student\_(True/False)      Weight      Speed      MPG

(f) What equation would you use to evaluate the result of a regression model?

(g) What equation would you use to evaluate the result of a classification model?



2. (15 Points) I'm building a model to predict amount of a given brand of dog food eaten by a collection of dogs. I have 100 dogs eat this dog food, and I collect information on their height, weight, breed, and whether they live with another dog in the house.
- (a) List all input variables and specify whether they are quantitative or qualitative.
- (b) Say our dog breeds sampled are Huskies, Terriers, and Spaniels. Write down your prediction model.
- (c) We found that weight is a key predictor, and its relationship with the response is beyond linear. What extension can you come up with? Write down your updated model.





514DC596-3BD9-437E-882B-2B63B4262F34

midterm1-b9ace

#35 Page 4 of 10

3. (15 points)

(a) What is bias-variance tradeoff? Explain the meaning of each term in the formula.

(b) Provide a sketch of typical (squared) bias, variance, training error, test error, and Bayes (or irreducible) error curves, on a single plot, as we go from less flexible statistical learning methods towards more flexible approaches. The  $x$ -axis should represent the amount of flexibility in the method, and the  $y$ -axis should represent the values for each curve. There should be five curves. Make sure to label each one.



(c) Explain why the (i) training error and (ii) testing error lines in your drawing have the shape displayed.



4. (10 points) The data for this example come from a study by Stamey et al. (1989). The data comes from a number of clinical measures in men who were about to receive a radical prostatectomy. The goal is to predict the log of PSA (**lpsa**) (a prostate-specific antigen measured in nanograms of PSA per milliliter of blood (ng/mL)) from a number of measurements. The variables are log cancer volume (**lcavol**), log prostate weight (**lweight**), age, log of the amount of benign prostatic hyperplasia (**lbph**), seminal vesicle invasion (**svi**), log of capsular penetration (**lcp**), Gleason score (**gleason**), and percent of Gleason scores 4 or 5 (**pgg45**).

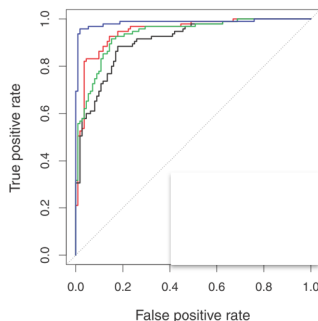
(a) Is this a case of regression or classification? Why?

(b) A linear model is fit to the data set, and the following table was returned.

Term	Coefficient	Std. Error	t-Score	p-value
Intercept	2.46	0.09	27.6	<0.00001
lcavol	0.68	0.13	5.37	<0.00001
lweight	0.26	0.1	2.75	0.00596
age	-0.14	0.1	-1.4	0.16153
lbph	0.21	0.1	2.06	0.039399
svi	0.31	0.12	2.47	0.013511
lcp	-0.29	0.15	-1.87	0.061484
gleason	-0.02	0.15	-0.15	0.880765
pgg45	0.27	0.15	1.74	0.081859

Are all the predictors useful? If yes, explain why. If not, explain what you would try to do next with the data.

5. (5 pts) We are trying to predict whether a patient will have a heart attack within one year. We have tried four different methods on a training dataset and have the following ROC curves. Which curve has the best performance on this training set? Justify your answer.





6. (15 Points) I get way too much email, so I decide to build a logistic regression model to predict whether a new incoming message is spam or not. I decide to just use a few variables:

$$\begin{aligned}X_1 &= \text{Number\_of\_sentences} \\X_2 &= \text{Number\_of\_references\_to\_a\_Nigerian\_Prince}\end{aligned}$$

and am training a logistic model to predict

$$Pr(Y = \text{spam} \mid X_1, X_2)$$

- (a) Write down the equation for the model you would train, using our standard notation with  $\beta_i$ 's.
- (b) If my trained model used  $\beta_0 = -13.1$ ,  $\beta_1 = 1.9$ , and  $\beta_2 = 6.1$ , what is the probability that a 5 sentence email with one reference to a Nigerian prince is spam? .
- (c) We know that if we miss an important email, it may cost a lot, How can you modify your model to avoid this?



7. (15 Points) Sparty generated 100 pair  $x$  and  $y$  via the following relationship:  $Y = 0.3 + 2x + x^2 + \epsilon$  but didn't tell you. You will try to find a good model to fit these data and more importantly to predict future response  $y$  when Sparty gives you another  $x$ . You will start with two models 1)  $Y = \beta_0 + \beta_1x$ ; 2)  $Y = \alpha_0 + \alpha_1x + \alpha_2x^2$ ; and 3)  $Y = \gamma_0 + \gamma_1x + \gamma_2x^2 + \gamma_3x^3$ .
- (a) If you use all 100 data to fit these three model and use the same 100 data to calculate MSE, which model will have the smallest MSE? Justify your answer
- (b) Since we focus on the ability to predict unseen data, you decide use validation set to evaluate the performance of the three models. There are two way to split the data: 1) 80 testing points and 20 training points or 2) 20 testing points and 80 training points. Which one will you choose? Justify your choice
- (c) From the in-class lab, we know validation set is not a good choice to choose model. Explain the drawback of validation set and what procedure will you use to choose the best model?



8. (10 Points) Suppose we have a data set with five predictors,  $X_1 = \text{GPA}$ ,  $X_2 = \text{IQ}$ ,  $X_3 = \text{Level}$  (1 for College and 0 for High School),  $X_4 = \text{Interaction between GPA and IQ}$ , and  $X_5 = \text{Interaction between GPA and Level}$ . The response is starting salary after graduation (in thousands of dollars). Suppose we use least squares to fit the model, and get  $\beta_0 = 50$ ,  $\beta_1 = 20$ ,  $\beta_2 = 0.07$ ,  $\beta_3 = 35$ ,  $\beta_4 = 0.01$ ,  $\beta_5 = -10$ .

(a) Predict the salary of a college graduate with IQ of 110 and a GPA of 4.0.

- (b) True or false: Since the coefficient for the GPA/IQ interaction term is very small, there is very little evidence of an interaction effect. Justify your answer.

DEABA7B4-14FF-4E26-94C9-39EF00B3EE54

midterm1-b9ace

#35 Page 9 of 10



## Scrap Paper



6CE8E75D-4A54-4EF7-915A-3859C6163EAB

midterm1-b9ace

#35 Page 10 of 10



Michigan State University  
CMSE381 - Data Science

Feb 10, 2023  
Dr. Xie

## CMSE381 - Midterm #1

First name (please write as legibly as possible within the boxes)

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

Last name

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

NetID

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

*I will adhere to the Spartan Code of Honor in completing this assignment.*

Signed: \_\_\_\_\_

1. Do not open this test booklet until you are directed to do so.
2. You will have class time (2:40-4:00pm) to complete the exam.
3. This exam is closed book, but you can use the cheatsheet provided by the instructor.
4. Throughout the test, show your work so that your reasoning is clear. Otherwise no credit will be given. **BOX** your answers. Partial credit will be given where warranted.
5. Do not spend too much time on any one problem. Read them all through first and attack them in the order that allows you to make the most progress. Good luck :P





9EB71CB6-6EE9-4E61-9AE2-AB13C1014720

midterm1-b9ace

#36 Page 2 of 10

1. (15 points)

(a) Logistic regression is used for regression.

TRUE      FALSE

(b) Any model will never have training error below the irreducible error.

TRUE      FALSE

(c) Increasing your model flexibility always results in a better model.

TRUE      FALSE

(d) A logistic regression model is set up so that the odds are linear.

TRUE      FALSE

(e) Circle all of the following that would represent a qualitative variable.

Age      Year      Dog\_breed      Country\_of\_origin

Student\_(True/False)      Weight      Speed      MPG

(f) What equation would you use to evaluate the result of a regression model?

(g) What equation would you use to evaluate the result of a classification model?



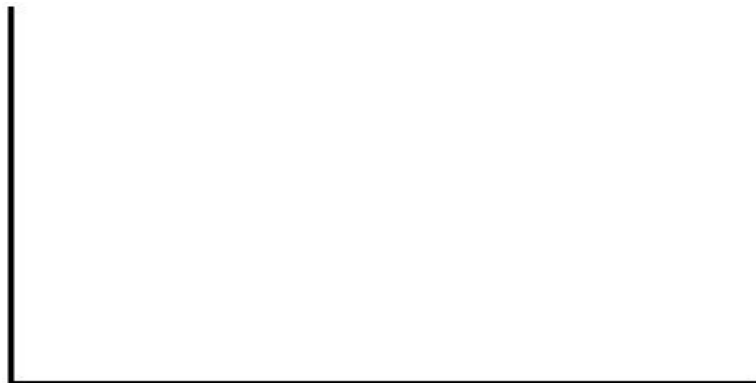
2. (15 Points) I'm building a model to predict amount of a given brand of dog food eaten by a collection of dogs. I have 100 dogs eat this dog food, and I collect information on their height, weight, breed, and whether they live with another dog in the house.
- (a) List all input variables and specify whether they are quantitative or qualitative.
  - (b) Say our dog breeds sampled are Huskies, Terriers, and Spaniels. Write down your prediction model.
  - (c) We found that weight is a key predictor, and its relationship with the response is beyond linear. What extension can you come up with? Write down your updated model.



3. (15 points)

(a) What is bias-variance tradeoff? Explain the meaning of each term in the formula.

(b) Provide a sketch of typical (squared) bias, variance, training error, test error, and Bayes (or irreducible) error curves, on a single plot, as we go from less flexible statistical learning methods towards more flexible approaches. The  $x$ -axis should represent the amount of flexibility in the method, and the  $y$ -axis should represent the values for each curve. There should be five curves. Make sure to label each one.



(c) Explain why the (i) training error and (ii) testing error lines in your drawing have the shape displayed.



4. (10 points) The data for this example come from a study by Stamey et al. (1989). The data comes from a number of clinical measures in men who were about to receive a radical prostatectomy. The goal is to predict the log of PSA (**lpsa**) (a prostate-specific antigen measured in nanograms of PSA per milliliter of blood (ng/mL)) from a number of measurements. The variables are log cancer volume (**lcavol**), log prostate weight (**lweight**), age, log of the amount of benign prostatic hyperplasia (**lbph**), seminal vesicle invasion (**svi**), log of capsular penetration (**lcp**), Gleason score (**gleason**), and percent of Gleason scores 4 or 5 (**pgg45**).

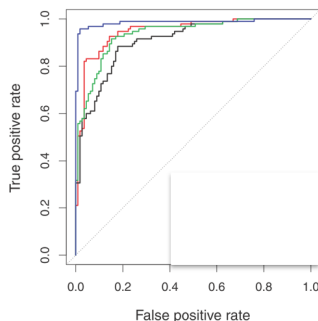
(a) Is this a case of regression or classification? Why?

(b) A linear model is fit to the data set, and the following table was returned.

Term	Coefficient	Std. Error	t-Score	p-value
Intercept	2.46	0.09	27.6	<0.00001
lcavol	0.68	0.13	5.37	<0.00001
lweight	0.26	0.1	2.75	0.00596
age	-0.14	0.1	-1.4	0.16153
lbph	0.21	0.1	2.06	0.039399
svi	0.31	0.12	2.47	0.013511
lcp	-0.29	0.15	-1.87	0.061484
gleason	-0.02	0.15	-0.15	0.880765
pgg45	0.27	0.15	1.74	0.081859

Are all the predictors useful? If yes, explain why. If not, explain what you would try to do next with the data.

5. (5 pts) We are trying to predict whether a patient will have a heart attack within one year. We have tried four different methods on a training dataset and have the following ROC curves. Which curve has the best performance on this training set? Justify your answer.





256E6D5E-C0B7-4BC3-BB6D-0AF4BEFEFC5F

midterm1-b9ace

#36 Page 6 of 10

6. (15 Points) I get way too much email, so I decide to build a logistic regression model to predict whether a new incoming message is spam or not. I decide to just use a few variables:

$$\begin{aligned} X_1 &= \text{Number\_of\_sentences} \\ X_2 &= \text{Number\_of\_references\_to\_a\_Nigerian\_Prince} \end{aligned}$$

and am training a logistic model to predict

$$Pr(Y = \text{spam} \mid X_1, X_2)$$

- (a) Write down the equation for the model you would train, using our standard notation with  $\beta_i$ 's.
- (b) If my trained model used  $\beta_0 = -13.1$ ,  $\beta_1 = 1.9$ , and  $\beta_2 = 6.1$ , what is the probability that a 5 sentence email with one reference to a Nigerian prince is spam? .
- (c) We know that if we miss an important email, it may cost a lot, How can you modify your model to avoid this?



7. (15 Points) Sparty generated 100 pair  $x$  and  $y$  via the following relationship:  $Y = 0.3 + 2x + x^2 + \epsilon$  but didn't tell you. You will try to find a good model to fit these data and more importantly to predict future response  $y$  when Sparty gives you another  $x$ . You will start with two models 1)  $Y = \beta_0 + \beta_1x$ ; 2)  $Y = \alpha_0 + \alpha_1x + \alpha_2x^2$ ; and 3)  $Y = \gamma_0 + \gamma_1x + \gamma_2x^2 + \gamma_3x^3$ .
- (a) If you use all 100 data to fit these three model and use the same 100 data to calculate MSE, which model will have the smallest MSE? Justify your answer
- (b) Since we focus on the ability to predict unseen data, you decide use validation set to evaluate the performance of the three models. There are two way to split the data: 1) 80 testing points and 20 training points or 2) 20 testing points and 80 training points. Which one will you choose? Justify your choice
- (c) From the in-class lab, we know validation set is not a good choice to choose model. Explain the drawback of validation set and what procedure will you use to choose the best model?



04899721-8153-4F50-A60A-54C91E3F4A8C

midterm1-b9ace

#36 Page 8 of 10

8. (10 Points) Suppose we have a data set with five predictors,  $X_1 = \text{GPA}$ ,  $X_2 = \text{IQ}$ ,  $X_3 = \text{Level}$  (1 for College and 0 for High School),  $X_4 = \text{Interaction between GPA and IQ}$ , and  $X_5 = \text{Interaction between GPA and Level}$ . The response is starting salary after graduation (in thousands of dollars). Suppose we use least squares to fit the model, and get  $\beta_0 = 50$ ,  $\beta_1 = 20$ ,  $\beta_2 = 0.07$ ,  $\beta_3 = 35$ ,  $\beta_4 = 0.01$ ,  $\beta_5 = -10$ .

(a) Predict the salary of a college graduate with IQ of 110 and a GPA of 4.0.

- (b) True or false: Since the coefficient for the GPA/IQ interaction term is very small, there is very little evidence of an interaction effect. Justify your answer.

AAB88229-4B45-4BAB-BE83-12666BB1FA80

midterm1-b9ace

#36 Page 9 of 10



**Scrap Paper**





1A0DD3C6-5D64-4CA3-BBEB-59955CE89489

midterm1-b9ace

#36 Page 10 of 10



Michigan State University  
CMSE381 - Data Science

Feb 10, 2023  
Dr. Xie

## CMSE381 - Midterm #1

First name (please write as legibly as possible within the boxes)

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

Last name

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

NetID

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

*I will adhere to the Spartan Code of Honor in completing this assignment.*

Signed: \_\_\_\_\_

1. Do not open this test booklet until you are directed to do so.
2. You will have class time (2:40-4:00pm) to complete the exam.
3. This exam is closed book, but you can use the cheatsheet provided by the instructor.
4. Throughout the test, show your work so that your reasoning is clear. Otherwise no credit will be given. **BOX** your answers. Partial credit will be given where warranted.
5. Do not spend too much time on any one problem. Read them all through first and attack them in the order that allows you to make the most progress. Good luck :P



06F9B7BD-C901-4A04-B378-918241F23FA0

midterm1-b9ace

#37 Page 2 of 10

1. (15 points)

(a) Logistic regression is used for regression.

TRUE      FALSE

(b) Any model will never have training error below the irreducible error.

TRUE      FALSE

(c) Increasing your model flexibility always results in a better model.

TRUE      FALSE

(d) A logistic regression model is set up so that the odds are linear.

TRUE      FALSE

(e) Circle all of the following that would represent a qualitative variable.

Age      Year      Dog\_breed      Country\_of\_origin

Student\_(True/False)      Weight      Speed      MPG

(f) What equation would you use to evaluate the result of a regression model?

(g) What equation would you use to evaluate the result of a classification model?



2. (15 Points) I'm building a model to predict amount of a given brand of dog food eaten by a collection of dogs. I have 100 dogs eat this dog food, and I collect information on their height, weight, breed, and whether they live with another dog in the house.
- (a) List all input variables and specify whether they are quantitative or qualitative.
- (b) Say our dog breeds sampled are Huskies, Terriers, and Spaniels. Write down your prediction model.
- (c) We found that weight is a key predictor, and its relationship with the response is beyond linear. What extension can you come up with? Write down your updated model.



B475CB29-6BAF-4EC9-A1ED-36DBC5ECB16B

midterm1-b9ace

#37 Page 4 of 10

3. (15 points)

(a) What is bias-variance tradeoff? Explain the meaning of each term in the formula.

(b) Provide a sketch of typical (squared) bias, variance, training error, test error, and Bayes (or irreducible) error curves, on a single plot, as we go from less flexible statistical learning methods towards more flexible approaches. The  $x$ -axis should represent the amount of flexibility in the method, and the  $y$ -axis should represent the values for each curve. There should be five curves. Make sure to label each one.



(c) Explain why the (i) training error and (ii) testing error lines in your drawing have the shape displayed.



4. (10 points) The data for this example come from a study by Stamey et al. (1989). The data comes from a number of clinical measures in men who were about to receive a radical prostatectomy. The goal is to predict the log of PSA (**lpsa**) (a prostate-specific antigen measured in nanograms of PSA per milliliter of blood (ng/mL)) from a number of measurements. The variables are log cancer volume (**lcavol**), log prostate weight (**lweight**), age, log of the amount of benign prostatic hyperplasia (**lbph**), seminal vesicle invasion (**svi**), log of capsular penetration (**lcp**), Gleason score (**gleason**), and percent of Gleason scores 4 or 5 (**pgg45**).

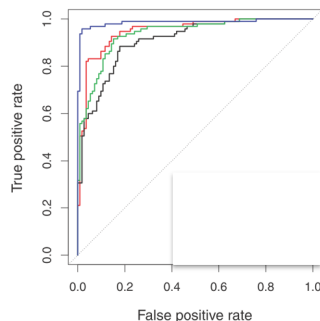
(a) Is this a case of regression or classification? Why?

(b) A linear model is fit to the data set, and the following table was returned.

Term	Coefficient	Std. Error	t-Score	p-value
Intercept	2.46	0.09	27.6	<0.00001
lcavol	0.68	0.13	5.37	<0.00001
lweight	0.26	0.1	2.75	0.00596
age	-0.14	0.1	-1.4	0.16153
lbph	0.21	0.1	2.06	0.039399
svi	0.31	0.12	2.47	0.013511
lcp	-0.29	0.15	-1.87	0.061484
gleason	-0.02	0.15	-0.15	0.880765
pgg45	0.27	0.15	1.74	0.081859

Are all the predictors useful? If yes, explain why. If not, explain what you would try to do next with the data.

5. (5 pts) We are trying to predict whether a patient will have a heart attack within one year. We have tried four different methods on a training dataset and have the following ROC curves. Which curve has the best performance on this training set? Justify your answer.





6. (15 Points) I get way too much email, so I decide to build a logistic regression model to predict whether a new incoming message is spam or not. I decide to just use a few variables:

$$\begin{aligned}X_1 &= \text{Number\_of\_sentences} \\X_2 &= \text{Number\_of\_references\_to\_a\_Nigerian\_Prince}\end{aligned}$$

and am training a logistic model to predict

$$Pr(Y = \text{spam} \mid X_1, X_2)$$

- (a) Write down the equation for the model you would train, using our standard notation with  $\beta_i$ 's.
- (b) If my trained model used  $\beta_0 = -13.1$ ,  $\beta_1 = 1.9$ , and  $\beta_2 = 6.1$ , what is the probability that a 5 sentence email with one reference to a Nigerian prince is spam? .
- (c) We know that if we miss an important email, it may cost a lot, How can you modify your model to avoid this?



7. (15 Points) Sparty generated 100 pair  $x$  and  $y$  via the following relationship:  $Y = 0.3 + 2x + x^2 + \epsilon$  but didn't tell you. You will try to find a good model to fit these data and more importantly to predict future response  $y$  when Sparty gives you another  $x$ . You will start with two models 1)  $Y = \beta_0 + \beta_1x$ ; 2)  $Y = \alpha_0 + \alpha_1x + \alpha_2x^2$ ; and 3)  $Y = \gamma_0 + \gamma_1x + \gamma_2x^2 + \gamma_3x^3$ .
- (a) If you use all 100 data to fit these three model and use the same 100 data to calculate MSE, which model will have the smallest MSE? Justify your answer
- (b) Since we focus on the ability to predict unseen data, you decide use validation set to evaluate the performance of the three models. There are two way to split the data: 1) 80 testing points and 20 training points or 2) 20 testing points and 80 training points. Which one will you choose? Justify your choice
- (c) From the in-class lab, we know validation set is not a good choice to choose model. Explain the drawback of validation set and what procedure will you use to choose the best model?





F51ED66A-E79C-4C06-AE9E-27A39B447934

midterm1-b9ace

#37 Page 8 of 10

8. (10 Points) Suppose we have a data set with five predictors,  $X_1 = \text{GPA}$ ,  $X_2 = \text{IQ}$ ,  $X_3 = \text{Level}$  (1 for College and 0 for High School),  $X_4 = \text{Interaction between GPA and IQ}$ , and  $X_5 = \text{Interaction between GPA and Level}$ . The response is starting salary after graduation (in thousands of dollars). Suppose we use least squares to fit the model, and get  $\beta_0 = 50$ ,  $\beta_1 = 20$ ,  $\beta_2 = 0.07$ ,  $\beta_3 = 35$ ,  $\beta_4 = 0.01$ ,  $\beta_5 = -10$ .

(a) Predict the salary of a college graduate with IQ of 110 and a GPA of 4.0.

- (b) True or false: Since the coefficient for the GPA/IQ interaction term is very small, there is very little evidence of an interaction effect. Justify your answer.

1FF199B8-B5F8-4BF5-9597-FB8286E01D56

midterm1-b9ace

#37 Page 9 of 10



**Scrap Paper**



B65CC2A5-A318-46AA-90F8-BA9190B984B2

midterm1-b9ace

#37 Page 10 of 10



Michigan State University  
CMSE381 - Data Science

Feb 10, 2023  
Dr. Xie

## CMSE381 - Midterm #1

First name (please write as legibly as possible within the boxes)

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

Last name

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

NetID

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

*I will adhere to the Spartan Code of Honor in completing this assignment.*

Signed: \_\_\_\_\_

1. Do not open this test booklet until you are directed to do so.
2. You will have class time (2:40-4:00pm) to complete the exam.
3. This exam is closed book, but you can use the cheatsheet provided by the instructor.
4. Throughout the test, show your work so that your reasoning is clear. Otherwise no credit will be given. **BOX** your answers. Partial credit will be given where warranted.
5. Do not spend too much time on any one problem. Read them all through first and attack them in the order that allows you to make the most progress. Good luck :P



C245557B-007B-41F3-92A1-7011E1DED6D7

midterm1-b9ace

#38 Page 2 of 10

1. (15 points)

(a) Logistic regression is used for regression.

TRUE      FALSE

(b) Any model will never have training error below the irreducible error.

TRUE      FALSE

(c) Increasing your model flexibility always results in a better model.

TRUE      FALSE

(d) A logistic regression model is set up so that the odds are linear.

TRUE      FALSE

(e) Circle all of the following that would represent a qualitative variable.

Age      Year      Dog\_breed      Country\_of\_origin

Student\_(True/False)      Weight      Speed      MPG

(f) What equation would you use to evaluate the result of a regression model?

(g) What equation would you use to evaluate the result of a classification model?



2. (15 Points) I'm building a model to predict amount of a given brand of dog food eaten by a collection of dogs. I have 100 dogs eat this dog food, and I collect information on their height, weight, breed, and whether they live with another dog in the house.
- (a) List all input variables and specify whether they are quantitative or qualitative.
  - (b) Say our dog breeds sampled are Huskies, Terriers, and Spaniels. Write down your prediction model.
  - (c) We found that weight is a key predictor, and its relationship with the response is beyond linear. What extension can you come up with? Write down your updated model.



49E001D5-AE9D-4F8E-99AF-FC2BD81DC7C6

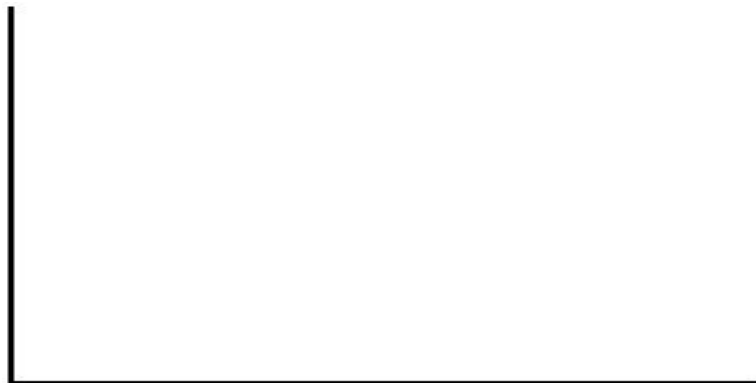
midterm1-b9ace

#38 Page 4 of 10

3. (15 points)

(a) What is bias-variance tradeoff? Explain the meaning of each term in the formula.

(b) Provide a sketch of typical (squared) bias, variance, training error, test error, and Bayes (or irreducible) error curves, on a single plot, as we go from less flexible statistical learning methods towards more flexible approaches. The  $x$ -axis should represent the amount of flexibility in the method, and the  $y$ -axis should represent the values for each curve. There should be five curves. Make sure to label each one.



(c) Explain why the (i) training error and (ii) testing error lines in your drawing have the shape displayed.



4. (10 points) The data for this example come from a study by Stamey et al. (1989). The data comes from a number of clinical measures in men who were about to receive a radical prostatectomy. The goal is to predict the log of PSA (**lpsa**) (a prostate-specific antigen measured in nanograms of PSA per milliliter of blood (ng/mL)) from a number of measurements. The variables are log cancer volume (**lcavol**), log prostate weight (**lweight**), age, log of the amount of benign prostatic hyperplasia (**lbph**), seminal vesicle invasion (**svi**), log of capsular penetration (**lcp**), Gleason score (**gleason**), and percent of Gleason scores 4 or 5 (**pgg45**).

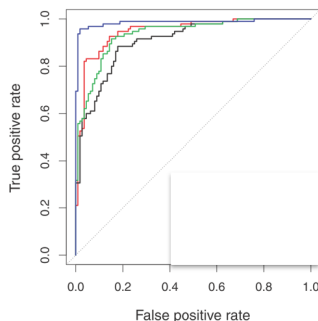
(a) Is this a case of regression or classification? Why?

(b) A linear model is fit to the data set, and the following table was returned.

Term	Coefficient	Std. Error	t-Score	p-value
Intercept	2.46	0.09	27.6	<0.00001
lcavol	0.68	0.13	5.37	<0.00001
lweight	0.26	0.1	2.75	0.00596
age	-0.14	0.1	-1.4	0.16153
lbph	0.21	0.1	2.06	0.039399
svi	0.31	0.12	2.47	0.013511
lcp	-0.29	0.15	-1.87	0.061484
gleason	-0.02	0.15	-0.15	0.880765
pgg45	0.27	0.15	1.74	0.081859

Are all the predictors useful? If yes, explain why. If not, explain what you would try to do next with the data.

5. (5 pts) We are trying to predict whether a patient will have a heart attack within one year. We have tried four different methods on a training dataset and have the following ROC curves. Which curve has the best performance on this training set? Justify your answer.







5C871C55-EA03-4B87-A339-2AEC922A93AD

midterm1-b9ace

#38 Page 6 of 10

6. (15 Points) I get way too much email, so I decide to build a logistic regression model to predict whether a new incoming message is spam or not. I decide to just use a few variables:

$$\begin{aligned}X_1 &= \text{Number\_of\_sentences} \\X_2 &= \text{Number\_of\_references\_to\_a\_Nigerian\_Prince}\end{aligned}$$

and am training a logistic model to predict

$$Pr(Y = \text{spam} \mid X_1, X_2)$$

- (a) Write down the equation for the model you would train, using our standard notation with  $\beta_i$ 's.
- (b) If my trained model used  $\beta_0 = -13.1$ ,  $\beta_1 = 1.9$ , and  $\beta_2 = 6.1$ , what is the probability that a 5 sentence email with one reference to a Nigerian prince is spam? .
- (c) We know that if we miss an important email, it may cost a lot, How can you modify your model to avoid this?



7. (15 Points) Sparty generated 100 pair  $x$  and  $y$  via the following relationship:  $Y = 0.3 + 2x + x^2 + \epsilon$  but didn't tell you. You will try to find a good model to fit these data and more importantly to predict future response  $y$  when Sparty gives you another  $x$ . You will start with two models 1)  $Y = \beta_0 + \beta_1x$ ; 2)  $Y = \alpha_0 + \alpha_1x + \alpha_2x^2$ ; and 3)  $Y = \gamma_0 + \gamma_1x + \gamma_2x^2 + \gamma_3x^3$ .
- (a) If you use all 100 data to fit these three model and use the same 100 data to calculate MSE, which model will have the smallest MSE? Justify your answer
- (b) Since we focus on the ability to predict unseen data, you decide use validation set to evaluate the performance of the three models. There are two way to split the data: 1) 80 testing points and 20 training points or 2) 20 testing points and 80 training points. Which one will you choose? Justify your choice
- (c) From the in-class lab, we know validation set is not a good choice to choose model. Explain the drawback of validation set and what procedure will you use to choose the best model?



8. (10 Points) Suppose we have a data set with five predictors,  $X_1 = \text{GPA}$ ,  $X_2 = \text{IQ}$ ,  $X_3 = \text{Level}$  (1 for College and 0 for High School),  $X_4 = \text{Interaction between GPA and IQ}$ , and  $X_5 = \text{Interaction between GPA and Level}$ . The response is starting salary after graduation (in thousands of dollars). Suppose we use least squares to fit the model, and get  $\beta_0 = 50$ ,  $\beta_1 = 20$ ,  $\beta_2 = 0.07$ ,  $\beta_3 = 35$ ,  $\beta_4 = 0.01$ ,  $\beta_5 = -10$ .

(a) Predict the salary of a college graduate with IQ of 110 and a GPA of 4.0.

(b) True or false: Since the coefficient for the GPA/IQ interaction term is very small, there is very little evidence of an interaction effect. Justify your answer.

BEB51A7E-55F8-4D56-AC5E-484DA79D98B4

midterm1-b9ace

#38 Page 9 of 10



**Scrap Paper**



30E263C6-D188-49ED-B51F-2A890C665D1F

midterm1-b9ace

#38 Page 10 of 10



Michigan State University  
CMSE381 - Data Science

Feb 10, 2023  
Dr. Xie

## CMSE381 - Midterm #1

First name (please write as legibly as possible within the boxes)

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

Last name

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

NetID

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

*I will adhere to the Spartan Code of Honor in completing this assignment.*

Signed: \_\_\_\_\_

1. Do not open this test booklet until you are directed to do so.
2. You will have class time (2:40-4:00pm) to complete the exam.
3. This exam is closed book, but you can use the cheatsheet provided by the instructor.
4. Throughout the test, show your work so that your reasoning is clear. Otherwise no credit will be given. **BOX** your answers. Partial credit will be given where warranted.
5. Do not spend too much time on any one problem. Read them all through first and attack them in the order that allows you to make the most progress. Good luck :P



1. (15 points)

(a) Logistic regression is used for regression.

TRUE      FALSE

(b) Any model will never have training error below the irreducible error.

TRUE      FALSE

(c) Increasing your model flexibility always results in a better model.

TRUE      FALSE

(d) A logistic regression model is set up so that the odds are linear.

TRUE      FALSE

(e) Circle all of the following that would represent a qualitative variable.

Age      Year      Dog\_breed      Country\_of\_origin

Student\_(True/False)      Weight      Speed      MPG

(f) What equation would you use to evaluate the result of a regression model?

(g) What equation would you use to evaluate the result of a classification model?



2. (15 Points) I'm building a model to predict amount of a given brand of dog food eaten by a collection of dogs. I have 100 dogs eat this dog food, and I collect information on their height, weight, breed, and whether they live with another dog in the house.
- (a) List all input variables and specify whether they are quantitative or qualitative.
  - (b) Say our dog breeds sampled are Huskies, Terriers, and Spaniels. Write down your prediction model.
  - (c) We found that weight is a key predictor, and its relationship with the response is beyond linear. What extension can you come up with? Write down your updated model.





93AB9104-FF10-4307-8FDB-6A06B204337B

midterm1-b9ace

#39 Page 4 of 10

3. (15 points)

(a) What is bias-variance tradeoff? Explain the meaning of each term in the formula.

(b) Provide a sketch of typical (squared) bias, variance, training error, test error, and Bayes (or irreducible) error curves, on a single plot, as we go from less flexible statistical learning methods towards more flexible approaches. The  $x$ -axis should represent the amount of flexibility in the method, and the  $y$ -axis should represent the values for each curve. There should be five curves. Make sure to label each one.



(c) Explain why the (i) training error and (ii) testing error lines in your drawing have the shape displayed.



4. (10 points) The data for this example come from a study by Stamey et al. (1989). The data comes from a number of clinical measures in men who were about to receive a radical prostatectomy. The goal is to predict the log of PSA (**lpsa**) (a prostate-specific antigen measured in nanograms of PSA per milliliter of blood (ng/mL)) from a number of measurements. The variables are log cancer volume (**lcavol**), log prostate weight (**lweight**), age, log of the amount of benign prostatic hyperplasia (**lbph**), seminal vesicle invasion (**svi**), log of capsular penetration (**lcp**), Gleason score (**gleason**), and percent of Gleason scores 4 or 5 (**pgg45**).

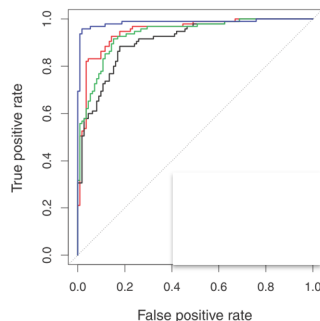
(a) Is this a case of regression or classification? Why?

(b) A linear model is fit to the data set, and the following table was returned.

Term	Coefficient	Std. Error	t-Score	p-value
Intercept	2.46	0.09	27.6	<0.00001
lcavol	0.68	0.13	5.37	<0.00001
lweight	0.26	0.1	2.75	0.00596
age	-0.14	0.1	-1.4	0.16153
lbph	0.21	0.1	2.06	0.039399
svi	0.31	0.12	2.47	0.013511
lcp	-0.29	0.15	-1.87	0.061484
gleason	-0.02	0.15	-0.15	0.880765
pgg45	0.27	0.15	1.74	0.081859

Are all the predictors useful? If yes, explain why. If not, explain what you would try to do next with the data.

5. (5 pts) We are trying to predict whether a patient will have a heart attack within one year. We have tried four different methods on a training dataset and have the following ROC curves. Which curve has the best performance on this training set? Justify your answer.





E2C4D1E5-F67D-4B12-877E-F561B4F86745

midterm1-b9ace

#39 Page 6 of 10

6. (15 Points) I get way too much email, so I decide to build a logistic regression model to predict whether a new incoming message is spam or not. I decide to just use a few variables:

$$\begin{aligned}X_1 &= \text{Number\_of\_sentences} \\X_2 &= \text{Number\_of\_references\_to\_a\_Nigerian\_Prince}\end{aligned}$$

and am training a logistic model to predict

$$Pr(Y = \text{spam} \mid X_1, X_2)$$

- (a) Write down the equation for the model you would train, using our standard notation with  $\beta_i$ 's.
- (b) If my trained model used  $\beta_0 = -13.1$ ,  $\beta_1 = 1.9$ , and  $\beta_2 = 6.1$ , what is the probability that a 5 sentence email with one reference to a Nigerian prince is spam? .
- (c) We know that if we miss an important email, it may cost a lot, How can you modify your model to avoid this?



7. (15 Points) Sparty generated 100 pair  $x$  and  $y$  via the following relationship:  $Y = 0.3 + 2x + x^2 + \epsilon$  but didn't tell you. You will try to find a good model to fit these data and more importantly to predict future response  $y$  when Sparty gives you another  $x$ . You will start with two models 1)  $Y = \beta_0 + \beta_1x$ ; 2)  $Y = \alpha_0 + \alpha_1x + \alpha_2x^2$ ; and 3)  $Y = \gamma_0 + \gamma_1x + \gamma_2x^2 + \gamma_3x^3$ .
- (a) If you use all 100 data to fit these three model and use the same 100 data to calculate MSE, which model will have the smallest MSE? Justify your answer
- (b) Since we focus on the ability to predict unseen data, you decide use validation set to evaluate the performance of the three models. There are two way to split the data: 1) 80 testing points and 20 training points or 2) 20 testing points and 80 training points. Which one will you choose? Justify your choice
- (c) From the in-class lab, we know validation set is not a good choice to choose model. Explain the drawback of validation set and what procedure will you use to choose the best model?



A91E02BF-192F-4C42-B5ED-3D041DE2E059

midterm1-b9ace

#39 Page 8 of 10

8. (10 Points) Suppose we have a data set with five predictors,  $X_1 = \text{GPA}$ ,  $X_2 = \text{IQ}$ ,  $X_3 = \text{Level}$  (1 for College and 0 for High School),  $X_4 = \text{Interaction between GPA and IQ}$ , and  $X_5 = \text{Interaction between GPA and Level}$ . The response is starting salary after graduation (in thousands of dollars). Suppose we use least squares to fit the model, and get  $\beta_0 = 50$ ,  $\beta_1 = 20$ ,  $\beta_2 = 0.07$ ,  $\beta_3 = 35$ ,  $\beta_4 = 0.01$ ,  $\beta_5 = -10$ .

(a) Predict the salary of a college graduate with IQ of 110 and a GPA of 4.0.

(b) True or false: Since the coefficient for the GPA/IQ interaction term is very small, there is very little evidence of an interaction effect. Justify your answer.

38602AFC-9DA0-4840-87E1-4B3331E9ECB8

midterm1-b9ace

#39 Page 9 of 10



**Scrap Paper**



9DC274BD-BAF8-4C08-BAD6-F7952F1615BB

midterm1-b9ace

#39 Page 10 of 10



Michigan State University  
CMSE381 - Data Science

Feb 10, 2023  
Dr. Xie

## CMSE381 - Midterm #1

First name (please write as legibly as possible within the boxes)

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

Last name

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

NetID

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

*I will adhere to the Spartan Code of Honor in completing this assignment.*

Signed: \_\_\_\_\_

1. Do not open this test booklet until you are directed to do so.
2. You will have class time (2:40-4:00pm) to complete the exam.
3. This exam is closed book, but you can use the cheatsheet provided by the instructor.
4. Throughout the test, show your work so that your reasoning is clear. Otherwise no credit will be given. **BOX** your answers. Partial credit will be given where warranted.
5. Do not spend too much time on any one problem. Read them all through first and attack them in the order that allows you to make the most progress. Good luck :P





2A6C86A5-9124-45D8-8878-C1CBBD811EA4

midterm1-b9ace

#40 Page 2 of 10

1. (15 points)

(a) Logistic regression is used for regression.

TRUE      FALSE

(b) Any model will never have training error below the irreducible error.

TRUE      FALSE

(c) Increasing your model flexibility always results in a better model.

TRUE      FALSE

(d) A logistic regression model is set up so that the odds are linear.

TRUE      FALSE

(e) Circle all of the following that would represent a qualitative variable.

Age      Year      Dog\_breed      Country\_of\_origin

Student\_(True/False)      Weight      Speed      MPG

(f) What equation would you use to evaluate the result of a regression model?

(g) What equation would you use to evaluate the result of a classification model?



2. (15 Points) I'm building a model to predict amount of a given brand of dog food eaten by a collection of dogs. I have 100 dogs eat this dog food, and I collect information on their height, weight, breed, and whether they live with another dog in the house.
- (a) List all input variables and specify whether they are quantitative or qualitative.
  - (b) Say our dog breeds sampled are Huskies, Terriers, and Spaniels. Write down your prediction model.
  - (c) We found that weight is a key predictor, and its relationship with the response is beyond linear. What extension can you come up with? Write down your updated model.



12EC453E-45B1-4727-8643-E97A328C0F50

midterm1-b9ace

#40 Page 4 of 10

3. (15 points)

(a) What is bias-variance tradeoff? Explain the meaning of each term in the formula.

(b) Provide a sketch of typical (squared) bias, variance, training error, test error, and Bayes (or irreducible) error curves, on a single plot, as we go from less flexible statistical learning methods towards more flexible approaches. The  $x$ -axis should represent the amount of flexibility in the method, and the  $y$ -axis should represent the values for each curve. There should be five curves. Make sure to label each one.



(c) Explain why the (i) training error and (ii) testing error lines in your drawing have the shape displayed.



4. (10 points) The data for this example come from a study by Stamey et al. (1989). The data comes from a number of clinical measures in men who were about to receive a radical prostatectomy. The goal is to predict the log of PSA (**lpsa**) (a prostate-specific antigen measured in nanograms of PSA per milliliter of blood (ng/mL)) from a number of measurements. The variables are log cancer volume (**lcavol**), log prostate weight (**lweight**), age, log of the amount of benign prostatic hyperplasia (**lbph**), seminal vesicle invasion (**svi**), log of capsular penetration (**lcp**), Gleason score (**gleason**), and percent of Gleason scores 4 or 5 (**pgg45**).

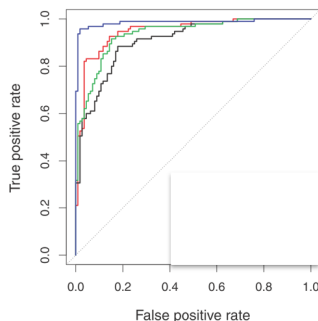
(a) Is this a case of regression or classification? Why?

(b) A linear model is fit to the data set, and the following table was returned.

Term	Coefficient	Std. Error	t-Score	p-value
Intercept	2.46	0.09	27.6	<0.00001
lcavol	0.68	0.13	5.37	<0.00001
lweight	0.26	0.1	2.75	0.00596
age	-0.14	0.1	-1.4	0.16153
lbph	0.21	0.1	2.06	0.039399
svi	0.31	0.12	2.47	0.013511
lcp	-0.29	0.15	-1.87	0.061484
gleason	-0.02	0.15	-0.15	0.880765
pgg45	0.27	0.15	1.74	0.081859

Are all the predictors useful? If yes, explain why. If not, explain what you would try to do next with the data.

5. (5 pts) We are trying to predict whether a patient will have a heart attack within one year. We have tried four different methods on a training dataset and have the following ROC curves. Which curve has the best performance on this training set? Justify your answer.





6. (15 Points) I get way too much email, so I decide to build a logistic regression model to predict whether a new incoming message is spam or not. I decide to just use a few variables:

$$\begin{aligned}X_1 &= \text{Number\_of\_sentences} \\X_2 &= \text{Number\_of\_references\_to\_a\_Nigerian\_Prince}\end{aligned}$$

and am training a logistic model to predict

$$Pr(Y = \text{spam} \mid X_1, X_2)$$

- (a) Write down the equation for the model you would train, using our standard notation with  $\beta_i$ 's.
- (b) If my trained model used  $\beta_0 = -13.1$ ,  $\beta_1 = 1.9$ , and  $\beta_2 = 6.1$ , what is the probability that a 5 sentence email with one reference to a Nigerian prince is spam? .
- (c) We know that if we miss an important email, it may cost a lot, How can you modify your model to avoid this?



7. (15 Points) Sparty generated 100 pair  $x$  and  $y$  via the following relationship:  $Y = 0.3 + 2x + x^2 + \epsilon$  but didn't tell you. You will try to find a good model to fit these data and more importantly to predict future response  $y$  when Sparty gives you another  $x$ . You will start with two models 1)  $Y = \beta_0 + \beta_1 x$ ; 2)  $Y = \alpha_0 + \alpha_1 x + \alpha_2 x^2$ ; and 3)  $Y = \gamma_0 + \gamma_1 x + \gamma_2 x^2 + \gamma_3 x^3$ .
- (a) If you use all 100 data to fit these three model and use the same 100 data to calculate MSE, which model will have the smallest MSE? Justify your answer
- (b) Since we focus on the ability to predict unseen data, you decide use validation set to evaluate the performance of the three models. There are two way to split the data: 1) 80 testing points and 20 training points or 2) 20 testing points and 80 training points. Which one will you choose? Justify your choice
- (c) From the in-class lab, we know validation set is not a good choice to choose model. Explain the drawback of validation set and what procedure will you use to choose the best model?



8. (10 Points) Suppose we have a data set with five predictors,  $X_1 = \text{GPA}$ ,  $X_2 = \text{IQ}$ ,  $X_3 = \text{Level}$  (1 for College and 0 for High School),  $X_4 = \text{Interaction between GPA and IQ}$ , and  $X_5 = \text{Interaction between GPA and Level}$ . The response is starting salary after graduation (in thousands of dollars). Suppose we use least squares to fit the model, and get  $\beta_0 = 50$ ,  $\beta_1 = 20$ ,  $\beta_2 = 0.07$ ,  $\beta_3 = 35$ ,  $\beta_4 = 0.01$ ,  $\beta_5 = -10$ .

(a) Predict the salary of a college graduate with IQ of 110 and a GPA of 4.0.

- (b) True or false: Since the coefficient for the GPA/IQ interaction term is very small, there is very little evidence of an interaction effect. Justify your answer.

2EEB5C86-AC2A-4AFF-95DC-7A9967CA7BFD

midterm1-b9ace

#40 Page 9 of 10



## Scrap Paper





30AB64AC-6E6E-4C77-BE20-F5C6F52FEB30

midterm1-b9ace

#40 Page 10 of 10



Michigan State University  
CMSE381 - Data Science

Feb 10, 2023  
Dr. Xie

## CMSE381 - Midterm #1

First name (please write as legibly as possible within the boxes)

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

Last name

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

NetID

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

*I will adhere to the Spartan Code of Honor in completing this assignment.*

Signed: \_\_\_\_\_

1. Do not open this test booklet until you are directed to do so.
2. You will have class time (2:40-4:00pm) to complete the exam.
3. This exam is closed book, but you can use the cheatsheet provided by the instructor.
4. Throughout the test, show your work so that your reasoning is clear. Otherwise no credit will be given. **BOX** your answers. Partial credit will be given where warranted.
5. Do not spend too much time on any one problem. Read them all through first and attack them in the order that allows you to make the most progress. Good luck :P



65A2F89B-FC81-4054-8D89-FADC9B7C168D

midterm1-b9ace

#41 Page 2 of 10

1. (15 points)

(a) Logistic regression is used for regression.

TRUE      FALSE

(b) Any model will never have training error below the irreducible error.

TRUE      FALSE

(c) Increasing your model flexibility always results in a better model.

TRUE      FALSE

(d) A logistic regression model is set up so that the odds are linear.

TRUE      FALSE

(e) Circle all of the following that would represent a qualitative variable.

Age      Year      Dog\_breed      Country\_of\_origin

Student\_(True/False)      Weight      Speed      MPG

(f) What equation would you use to evaluate the result of a regression model?

(g) What equation would you use to evaluate the result of a classification model?



2. (15 Points) I'm building a model to predict amount of a given brand of dog food eaten by a collection of dogs. I have 100 dogs eat this dog food, and I collect information on their height, weight, breed, and whether they live with another dog in the house.
- (a) List all input variables and specify whether they are quantitative or qualitative.
  - (b) Say our dog breeds sampled are Huskies, Terriers, and Spaniels. Write down your prediction model.
  - (c) We found that weight is a key predictor, and its relationship with the response is beyond linear. What extension can you come up with? Write down your updated model.



3. (15 points)

(a) What is bias-variance tradeoff? Explain the meaning of each term in the formula.

(b) Provide a sketch of typical (squared) bias, variance, training error, test error, and Bayes (or irreducible) error curves, on a single plot, as we go from less flexible statistical learning methods towards more flexible approaches. The  $x$ -axis should represent the amount of flexibility in the method, and the  $y$ -axis should represent the values for each curve. There should be five curves. Make sure to label each one.



(c) Explain why the (i) training error and (ii) testing error lines in your drawing have the shape displayed.



4. (10 points) The data for this example come from a study by Stamey et al. (1989). The data comes from a number of clinical measures in men who were about to receive a radical prostatectomy. The goal is to predict the log of PSA (**lpsa**) (a prostate-specific antigen measured in nanograms of PSA per milliliter of blood (ng/mL)) from a number of measurements. The variables are log cancer volume (**lcavol**), log prostate weight (**lweight**), age, log of the amount of benign prostatic hyperplasia (**lbph**), seminal vesicle invasion (**svi**), log of capsular penetration (**lcp**), Gleason score (**gleason**), and percent of Gleason scores 4 or 5 (**pgg45**).

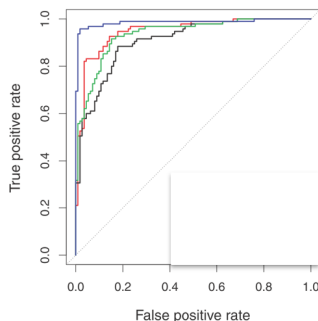
(a) Is this a case of regression or classification? Why?

(b) A linear model is fit to the data set, and the following table was returned.

Term	Coefficient	Std. Error	t-Score	p-value
Intercept	2.46	0.09	27.6	<0.00001
lcavol	0.68	0.13	5.37	<0.00001
lweight	0.26	0.1	2.75	0.00596
age	-0.14	0.1	-1.4	0.16153
lbph	0.21	0.1	2.06	0.039399
svi	0.31	0.12	2.47	0.013511
lcp	-0.29	0.15	-1.87	0.061484
gleason	-0.02	0.15	-0.15	0.880765
pgg45	0.27	0.15	1.74	0.081859

Are all the predictors useful? If yes, explain why. If not, explain what you would try to do next with the data.

5. (5 pts) We are trying to predict whether a patient will have a heart attack within one year. We have tried four different methods on a training dataset and have the following ROC curves. Which curve has the best performance on this training set? Justify your answer.





6. (15 Points) I get way too much email, so I decide to build a logistic regression model to predict whether a new incoming message is spam or not. I decide to just use a few variables:

$$\begin{aligned}X_1 &= \text{Number\_of\_sentences} \\X_2 &= \text{Number\_of\_references\_to\_a\_Nigerian\_Prince}\end{aligned}$$

and am training a logistic model to predict

$$Pr(Y = \text{spam} \mid X_1, X_2)$$

- (a) Write down the equation for the model you would train, using our standard notation with  $\beta_i$ 's.
- (b) If my trained model used  $\beta_0 = -13.1$ ,  $\beta_1 = 1.9$ , and  $\beta_2 = 6.1$ , what is the probability that a 5 sentence email with one reference to a Nigerian prince is spam? .
- (c) We know that if we miss an important email, it may cost a lot, How can you modify your model to avoid this?



7. (15 Points) Sparty generated 100 pair  $x$  and  $y$  via the following relationship:  $Y = 0.3 + 2x + x^2 + \epsilon$  but didn't tell you. You will try to find a good model to fit these data and more importantly to predict future response  $y$  when Sparty gives you another  $x$ . You will start with two models 1)  $Y = \beta_0 + \beta_1x$ ; 2)  $Y = \alpha_0 + \alpha_1x + \alpha_2x^2$ ; and 3)  $Y = \gamma_0 + \gamma_1x + \gamma_2x^2 + \gamma_3x^3$ .
- (a) If you use all 100 data to fit these three model and use the same 100 data to calculate MSE, which model will have the smallest MSE? Justify your answer
- (b) Since we focus on the ability to predict unseen data, you decide use validation set to evaluate the performance of the three models. There are two way to split the data: 1) 80 testing points and 20 training points or 2) 20 testing points and 80 training points. Which one will you choose? Justify your choice
- (c) From the in-class lab, we know validation set is not a good choice to choose model. Explain the drawback of validation set and what procedure will you use to choose the best model?





8. (10 Points) Suppose we have a data set with five predictors,  $X_1 = \text{GPA}$ ,  $X_2 = \text{IQ}$ ,  $X_3 = \text{Level}$  (1 for College and 0 for High School),  $X_4 = \text{Interaction between GPA and IQ}$ , and  $X_5 = \text{Interaction between GPA and Level}$ . The response is starting salary after graduation (in thousands of dollars). Suppose we use least squares to fit the model, and get  $\beta_0 = 50$ ,  $\beta_1 = 20$ ,  $\beta_2 = 0.07$ ,  $\beta_3 = 35$ ,  $\beta_4 = 0.01$ ,  $\beta_5 = -10$ .

(a) Predict the salary of a college graduate with IQ of 110 and a GPA of 4.0.

- (b) True or false: Since the coefficient for the GPA/IQ interaction term is very small, there is very little evidence of an interaction effect. Justify your answer.

A783760D-2BE0-4D52-AD54-82A82338EB84

midterm1-b9ace

#41 Page 9 of 10



**Scrap Paper**



6C27671A-5039-47A2-8A93-740A3857948F

midterm1-b9ace

#41 Page 10 of 10



Michigan State University  
CMSE381 - Data Science

Feb 10, 2023  
Dr. Xie

## CMSE381 - Midterm #1

First name (please write as legibly as possible within the boxes)

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

Last name

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

NetID

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

*I will adhere to the Spartan Code of Honor in completing this assignment.*

Signed: \_\_\_\_\_

1. Do not open this test booklet until you are directed to do so.
2. You will have class time (2:40-4:00pm) to complete the exam.
3. This exam is closed book, but you can use the cheatsheet provided by the instructor.
4. Throughout the test, show your work so that your reasoning is clear. Otherwise no credit will be given. **BOX** your answers. Partial credit will be given where warranted.
5. Do not spend too much time on any one problem. Read them all through first and attack them in the order that allows you to make the most progress. Good luck :P



1. (15 points)

(a) Logistic regression is used for regression.

TRUE      FALSE

(b) Any model will never have training error below the irreducible error.

TRUE      FALSE

(c) Increasing your model flexibility always results in a better model.

TRUE      FALSE

(d) A logistic regression model is set up so that the odds are linear.

TRUE      FALSE

(e) Circle all of the following that would represent a qualitative variable.

Age	Year	Dog_breed	Country_of_origin
Student_(True/False)	Weight	Speed	MPG

(f) What equation would you use to evaluate the result of a regression model?

(g) What equation would you use to evaluate the result of a classification model?



2. (15 Points) I'm building a model to predict amount of a given brand of dog food eaten by a collection of dogs. I have 100 dogs eat this dog food, and I collect information on their height, weight, breed, and whether they live with another dog in the house.
- (a) List all input variables and specify whether they are quantitative or qualitative.
- (b) Say our dog breeds sampled are Huskies, Terriers, and Spaniels. Write down your prediction model.
- (c) We found that weight is a key predictor, and its relationship with the response is beyond linear. What extension can you come up with? Write down your updated model.



52B75405-8DF4-4D45-9026-BE53685E0FA7

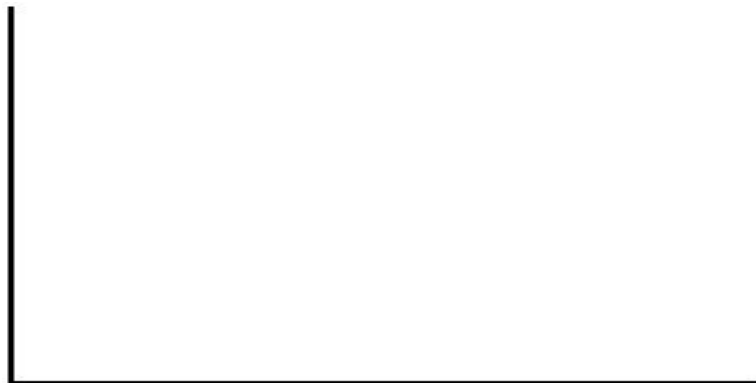
midterm1-b9ace

#42 Page 4 of 10

3. (15 points)

(a) What is bias-variance tradeoff? Explain the meaning of each term in the formula.

(b) Provide a sketch of typical (squared) bias, variance, training error, test error, and Bayes (or irreducible) error curves, on a single plot, as we go from less flexible statistical learning methods towards more flexible approaches. The  $x$ -axis should represent the amount of flexibility in the method, and the  $y$ -axis should represent the values for each curve. There should be five curves. Make sure to label each one.



(c) Explain why the (i) training error and (ii) testing error lines in your drawing have the shape displayed.



4. (10 points) The data for this example come from a study by Stamey et al. (1989). The data comes from a number of clinical measures in men who were about to receive a radical prostatectomy. The goal is to predict the log of PSA (**lpsa**) (a prostate-specific antigen measured in nanograms of PSA per milliliter of blood (ng/mL)) from a number of measurements. The variables are log cancer volume (**lcavol**), log prostate weight (**lweight**), age, log of the amount of benign prostatic hyperplasia (**lbph**), seminal vesicle invasion (**svi**), log of capsular penetration (**lcp**), Gleason score (**gleason**), and percent of Gleason scores 4 or 5 (**pgg45**).

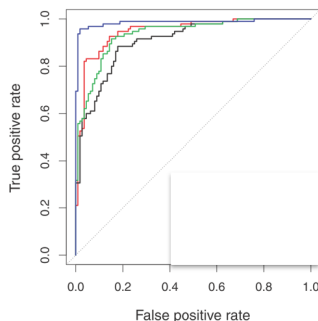
(a) Is this a case of regression or classification? Why?

(b) A linear model is fit to the data set, and the following table was returned.

Term	Coefficient	Std. Error	t-Score	p-value
Intercept	2.46	0.09	27.6	<0.00001
lcavol	0.68	0.13	5.37	<0.00001
lweight	0.26	0.1	2.75	0.00596
age	-0.14	0.1	-1.4	0.16153
lbph	0.21	0.1	2.06	0.039399
svi	0.31	0.12	2.47	0.013511
lcp	-0.29	0.15	-1.87	0.061484
gleason	-0.02	0.15	-0.15	0.880765
pgg45	0.27	0.15	1.74	0.081859

Are all the predictors useful? If yes, explain why. If not, explain what you would try to do next with the data.

5. (5 pts) We are trying to predict whether a patient will have a heart attack within one year. We have tried four different methods on a training dataset and have the following ROC curves. Which curve has the best performance on this training set? Justify your answer.







B197D9EE-C58D-4DDA-90D6-F8E2AC0800A6

midterm1-b9ace

#42 Page 6 of 10

6. (15 Points) I get way too much email, so I decide to build a logistic regression model to predict whether a new incoming message is spam or not. I decide to just use a few variables:

$$\begin{aligned}X_1 &= \text{Number\_of\_sentences} \\X_2 &= \text{Number\_of\_references\_to\_a\_Nigerian\_Prince}\end{aligned}$$

and am training a logistic model to predict

$$Pr(Y = \text{spam} \mid X_1, X_2)$$

- (a) Write down the equation for the model you would train, using our standard notation with  $\beta_i$ 's.
- (b) If my trained model used  $\beta_0 = -13.1$ ,  $\beta_1 = 1.9$ , and  $\beta_2 = 6.1$ , what is the probability that a 5 sentence email with one reference to a Nigerian prince is spam? .
- (c) We know that if we miss an important email, it may cost a lot, How can you modify your model to avoid this?



7. (15 Points) Sparty generated 100 pair  $x$  and  $y$  via the following relationship:  $Y = 0.3 + 2x + x^2 + \epsilon$  but didn't tell you. You will try to find a good model to fit these data and more importantly to predict future response  $y$  when Sparty gives you another  $x$ . You will start with two models 1)  $Y = \beta_0 + \beta_1x$ ; 2)  $Y = \alpha_0 + \alpha_1x + \alpha_2x^2$ ; and 3)  $Y = \gamma_0 + \gamma_1x + \gamma_2x^2 + \gamma_3x^3$ .
- (a) If you use all 100 data to fit these three model and use the same 100 data to calculate MSE, which model will have the smallest MSE? Justify your answer
- (b) Since we focus on the ability to predict unseen data, you decide use validation set to evaluate the performance of the three models. There are two way to split the data: 1) 80 testing points and 20 training points or 2) 20 testing points and 80 training points. Which one will you choose? Justify your choice
- (c) From the in-class lab, we know validation set is not a good choice to choose model. Explain the drawback of validation set and what procedure will you use to choose the best model?



8. (10 Points) Suppose we have a data set with five predictors,  $X_1 = \text{GPA}$ ,  $X_2 = \text{IQ}$ ,  $X_3 = \text{Level}$  (1 for College and 0 for High School),  $X_4 = \text{Interaction between GPA and IQ}$ , and  $X_5 = \text{Interaction between GPA and Level}$ . The response is starting salary after graduation (in thousands of dollars). Suppose we use least squares to fit the model, and get  $\beta_0 = 50$ ,  $\beta_1 = 20$ ,  $\beta_2 = 0.07$ ,  $\beta_3 = 35$ ,  $\beta_4 = 0.01$ ,  $\beta_5 = -10$ .

(a) Predict the salary of a college graduate with IQ of 110 and a GPA of 4.0.

- (b) True or false: Since the coefficient for the GPA/IQ interaction term is very small, there is very little evidence of an interaction effect. Justify your answer.

B45B24FA-DF72-4434-9E45-B8F870F5805C

midterm1-b9ace

#42 Page 9 of 10



**Scrap Paper**



D347B3F6-E02F-46BF-9F5B-FB7CDA77AA86

midterm1-b9ace

#42 Page 10 of 10



Michigan State University  
CMSE381 - Data Science

Feb 10, 2023  
Dr. Xie

## CMSE381 - Midterm #1

First name (please write as legibly as possible within the boxes)

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

Last name

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

NetID

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

*I will adhere to the Spartan Code of Honor in completing this assignment.*

Signed: \_\_\_\_\_

1. Do not open this test booklet until you are directed to do so.
2. You will have class time (2:40-4:00pm) to complete the exam.
3. This exam is closed book, but you can use the cheatsheet provided by the instructor.
4. Throughout the test, show your work so that your reasoning is clear. Otherwise no credit will be given. **BOX** your answers. Partial credit will be given where warranted.
5. Do not spend too much time on any one problem. Read them all through first and attack them in the order that allows you to make the most progress. Good luck :P



B011D3E8-8C5D-4548-9BCF-4B04ADD096E1

midterm1-b9ace

#43 Page 2 of 10

1. (15 points)

(a) Logistic regression is used for regression.

TRUE      FALSE

(b) Any model will never have training error below the irreducible error.

TRUE      FALSE

(c) Increasing your model flexibility always results in a better model.

TRUE      FALSE

(d) A logistic regression model is set up so that the odds are linear.

TRUE      FALSE

(e) Circle all of the following that would represent a qualitative variable.

Age      Year      Dog\_breed      Country\_of\_origin

Student\_(True/False)      Weight      Speed      MPG

(f) What equation would you use to evaluate the result of a regression model?

(g) What equation would you use to evaluate the result of a classification model?



2. (15 Points) I'm building a model to predict amount of a given brand of dog food eaten by a collection of dogs. I have 100 dogs eat this dog food, and I collect information on their height, weight, breed, and whether they live with another dog in the house.
- (a) List all input variables and specify whether they are quantitative or qualitative.
  - (b) Say our dog breeds sampled are Huskies, Terriers, and Spaniels. Write down your prediction model.
  - (c) We found that weight is a key predictor, and its relationship with the response is beyond linear. What extension can you come up with? Write down your updated model.





D6AD62B0-814A-4498-8B79-91A454AFE437

midterm1-b9ace

#43 Page 4 of 10

3. (15 points)

(a) What is bias-variance tradeoff? Explain the meaning of each term in the formula.

(b) Provide a sketch of typical (squared) bias, variance, training error, test error, and Bayes (or irreducible) error curves, on a single plot, as we go from less flexible statistical learning methods towards more flexible approaches. The  $x$ -axis should represent the amount of flexibility in the method, and the  $y$ -axis should represent the values for each curve. There should be five curves. Make sure to label each one.



(c) Explain why the (i) training error and (ii) testing error lines in your drawing have the shape displayed.



4. (10 points) The data for this example come from a study by Stamey et al. (1989). The data comes from a number of clinical measures in men who were about to receive a radical prostatectomy. The goal is to predict the log of PSA (**lpsa**) (a prostate-specific antigen measured in nanograms of PSA per milliliter of blood (ng/mL)) from a number of measurements. The variables are log cancer volume (**lcavol**), log prostate weight (**lweight**), age, log of the amount of benign prostatic hyperplasia (**lbph**), seminal vesicle invasion (**svi**), log of capsular penetration (**lcp**), Gleason score (**gleason**), and percent of Gleason scores 4 or 5 (**pgg45**).

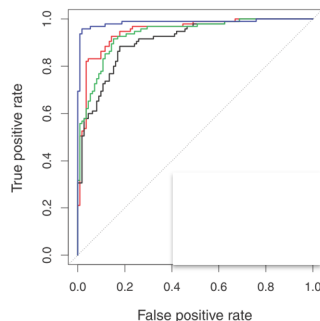
(a) Is this a case of regression or classification? Why?

(b) A linear model is fit to the data set, and the following table was returned.

Term	Coefficient	Std. Error	t-Score	p-value
Intercept	2.46	0.09	27.6	<0.00001
lcavol	0.68	0.13	5.37	<0.00001
lweight	0.26	0.1	2.75	0.00596
age	-0.14	0.1	-1.4	0.16153
lbph	0.21	0.1	2.06	0.039399
svi	0.31	0.12	2.47	0.013511
lcp	-0.29	0.15	-1.87	0.061484
gleason	-0.02	0.15	-0.15	0.880765
pgg45	0.27	0.15	1.74	0.081859

Are all the predictors useful? If yes, explain why. If not, explain what you would try to do next with the data.

5. (5 pts) We are trying to predict whether a patient will have a heart attack within one year. We have tried four different methods on a training dataset and have the following ROC curves. Which curve has the best performance on this training set? Justify your answer.





6. (15 Points) I get way too much email, so I decide to build a logistic regression model to predict whether a new incoming message is spam or not. I decide to just use a few variables:

$$\begin{aligned}X_1 &= \text{Number\_of\_sentences} \\X_2 &= \text{Number\_of\_references\_to\_a\_Nigerian\_Prince}\end{aligned}$$

and am training a logistic model to predict

$$Pr(Y = \text{spam} \mid X_1, X_2)$$

- (a) Write down the equation for the model you would train, using our standard notation with  $\beta_i$ 's.
- (b) If my trained model used  $\beta_0 = -13.1$ ,  $\beta_1 = 1.9$ , and  $\beta_2 = 6.1$ , what is the probability that a 5 sentence email with one reference to a Nigerian prince is spam? .
- (c) We know that if we miss an important email, it may cost a lot, How can you modify your model to avoid this?



7. (15 Points) Sparty generated 100 pair  $x$  and  $y$  via the following relationship:  $Y = 0.3 + 2x + x^2 + \epsilon$  but didn't tell you. You will try to find a good model to fit these data and more importantly to predict future response  $y$  when Sparty gives you another  $x$ . You will start with two models 1)  $Y = \beta_0 + \beta_1x$ ; 2)  $Y = \alpha_0 + \alpha_1x + \alpha_2x^2$ ; and 3)  $Y = \gamma_0 + \gamma_1x + \gamma_2x^2 + \gamma_3x^3$ .
- (a) If you use all 100 data to fit these three model and use the same 100 data to calculate MSE, which model will have the smallest MSE? Justify your answer
- (b) Since we focus on the ability to predict unseen data, you decide use validation set to evaluate the performance of the three models. There are two way to split the data: 1) 80 testing points and 20 training points or 2) 20 testing points and 80 training points. Which one will you choose? Justify your choice
- (c) From the in-class lab, we know validation set is not a good choice to choose model. Explain the drawback of validation set and what procedure will you use to choose the best model?



8. (10 Points) Suppose we have a data set with five predictors,  $X_1 = \text{GPA}$ ,  $X_2 = \text{IQ}$ ,  $X_3 = \text{Level}$  (1 for College and 0 for High School),  $X_4 = \text{Interaction between GPA and IQ}$ , and  $X_5 = \text{Interaction between GPA and Level}$ . The response is starting salary after graduation (in thousands of dollars). Suppose we use least squares to fit the model, and get  $\beta_0 = 50$ ,  $\beta_1 = 20$ ,  $\beta_2 = 0.07$ ,  $\beta_3 = 35$ ,  $\beta_4 = 0.01$ ,  $\beta_5 = -10$ .

(a) Predict the salary of a college graduate with IQ of 110 and a GPA of 4.0.

- (b) True or false: Since the coefficient for the GPA/IQ interaction term is very small, there is very little evidence of an interaction effect. Justify your answer.

BC2DD47E-6060-41A6-9C5E-5DC9781ED7B9

midterm1-b9ace

#43 Page 9 of 10



**Scrap Paper**



3148454C-B911-4504-B3D5-9E7422983D50

midterm1-b9ace

#43 Page 10 of 10



Michigan State University  
CMSE381 - Data Science

Feb 10, 2023  
Dr. Xie

## CMSE381 - Midterm #1

First name (please write as legibly as possible within the boxes)

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

Last name

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

NetID

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

*I will adhere to the Spartan Code of Honor in completing this assignment.*

Signed: \_\_\_\_\_

1. Do not open this test booklet until you are directed to do so.
2. You will have class time (2:40-4:00pm) to complete the exam.
3. This exam is closed book, but you can use the cheatsheet provided by the instructor.
4. Throughout the test, show your work so that your reasoning is clear. Otherwise no credit will be given. **BOX** your answers. Partial credit will be given where warranted.
5. Do not spend too much time on any one problem. Read them all through first and attack them in the order that allows you to make the most progress. Good luck :P





1. (15 points)

(a) Logistic regression is used for regression.

TRUE      FALSE

(b) Any model will never have training error below the irreducible error.

TRUE      FALSE

(c) Increasing your model flexibility always results in a better model.

TRUE      FALSE

(d) A logistic regression model is set up so that the odds are linear.

TRUE      FALSE

(e) Circle all of the following that would represent a qualitative variable.

Age      Year      Dog\_breed      Country\_of\_origin

Student\_(True/False)      Weight      Speed      MPG

(f) What equation would you use to evaluate the result of a regression model?

(g) What equation would you use to evaluate the result of a classification model?



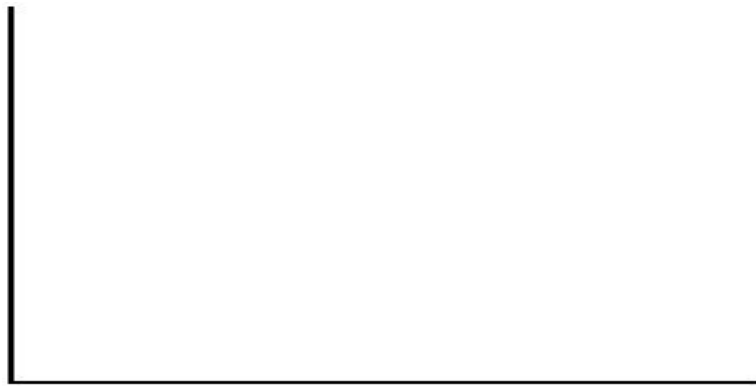
2. (15 Points) I'm building a model to predict amount of a given brand of dog food eaten by a collection of dogs. I have 100 dogs eat this dog food, and I collect information on their height, weight, breed, and whether they live with another dog in the house.
- (a) List all input variables and specify whether they are quantitative or qualitative.
  - (b) Say our dog breeds sampled are Huskies, Terriers, and Spaniels. Write down your prediction model.
  - (c) We found that weight is a key predictor, and its relationship with the response is beyond linear. What extension can you come up with? Write down your updated model.



3. (15 points)

(a) What is bias-variance tradeoff? Explain the meaning of each term in the formula.

(b) Provide a sketch of typical (squared) bias, variance, training error, test error, and Bayes (or irreducible) error curves, on a single plot, as we go from less flexible statistical learning methods towards more flexible approaches. The  $x$ -axis should represent the amount of flexibility in the method, and the  $y$ -axis should represent the values for each curve. There should be five curves. Make sure to label each one.



(c) Explain why the (i) training error and (ii) testing error lines in your drawing have the shape displayed.



4. (10 points) The data for this example come from a study by Stamey et al. (1989). The data comes from a number of clinical measures in men who were about to receive a radical prostatectomy. The goal is to predict the log of PSA (**lpsa**) (a prostate-specific antigen measured in nanograms of PSA per milliliter of blood (ng/mL)) from a number of measurements. The variables are log cancer volume (**lcavol**), log prostate weight (**lweight**), age, log of the amount of benign prostatic hyperplasia (**lbph**), seminal vesicle invasion (**svi**), log of capsular penetration (**lcp**), Gleason score (**gleason**), and percent of Gleason scores 4 or 5 (**pgg45**).

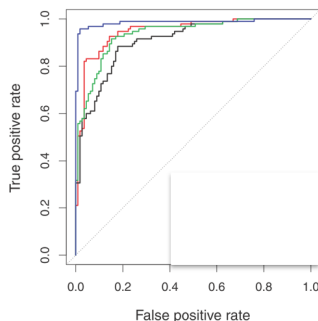
(a) Is this a case of regression or classification? Why?

(b) A linear model is fit to the data set, and the following table was returned.

Term	Coefficient	Std. Error	t-Score	p-value
Intercept	2.46	0.09	27.6	<0.00001
lcavol	0.68	0.13	5.37	<0.00001
lweight	0.26	0.1	2.75	0.00596
age	-0.14	0.1	-1.4	0.16153
lbph	0.21	0.1	2.06	0.039399
svi	0.31	0.12	2.47	0.013511
lcp	-0.29	0.15	-1.87	0.061484
gleason	-0.02	0.15	-0.15	0.880765
pgg45	0.27	0.15	1.74	0.081859

Are all the predictors useful? If yes, explain why. If not, explain what you would try to do next with the data.

5. (5 pts) We are trying to predict whether a patient will have a heart attack within one year. We have tried four different methods on a training dataset and have the following ROC curves. Which curve has the best performance on this training set? Justify your answer.





15E6148B-9CBD-436A-A1D5-DE47F8DB2FEB

midterm1-b9ace

#44 Page 6 of 10

6. (15 Points) I get way too much email, so I decide to build a logistic regression model to predict whether a new incoming message is spam or not. I decide to just use a few variables:

$$\begin{aligned}X_1 &= \text{Number\_of\_sentences} \\X_2 &= \text{Number\_of\_references\_to\_a\_Nigerian\_Prince}\end{aligned}$$

and am training a logistic model to predict

$$Pr(Y = \text{spam} \mid X_1, X_2)$$

- (a) Write down the equation for the model you would train, using our standard notation with  $\beta_i$ 's.
- (b) If my trained model used  $\beta_0 = -13.1$ ,  $\beta_1 = 1.9$ , and  $\beta_2 = 6.1$ , what is the probability that a 5 sentence email with one reference to a Nigerian prince is spam? .
- (c) We know that if we miss an important email, it may cost a lot, How can you modify your model to avoid this?



7. (15 Points) Sparty generated 100 pair  $x$  and  $y$  via the following relationship:  $Y = 0.3 + 2x + x^2 + \epsilon$  but didn't tell you. You will try to find a good model to fit these data and more importantly to predict future response  $y$  when Sparty gives you another  $x$ . You will start with two models 1)  $Y = \beta_0 + \beta_1x$ ; 2)  $Y = \alpha_0 + \alpha_1x + \alpha_2x^2$ ; and 3)  $Y = \gamma_0 + \gamma_1x + \gamma_2x^2 + \gamma_3x^3$ .
- (a) If you use all 100 data to fit these three model and use the same 100 data to calculate MSE, which model will have the smallest MSE? Justify your answer
- (b) Since we focus on the ability to predict unseen data, you decide use validation set to evaluate the performance of the three models. There are two way to split the data: 1) 80 testing points and 20 training points or 2) 20 testing points and 80 training points. Which one will you choose? Justify your choice
- (c) From the in-class lab, we know validation set is not a good choice to choose model. Explain the drawback of validation set and what procedure will you use to choose the best model?



6A1C8F65-9AAE-4E5D-95F9-3E8F15D8CB34

midterm1-b9ace

#44 Page 8 of 10

8. (10 Points) Suppose we have a data set with five predictors,  $X_1 = \text{GPA}$ ,  $X_2 = \text{IQ}$ ,  $X_3 = \text{Level}$  (1 for College and 0 for High School),  $X_4 = \text{Interaction between GPA and IQ}$ , and  $X_5 = \text{Interaction between GPA and Level}$ . The response is starting salary after graduation (in thousands of dollars). Suppose we use least squares to fit the model, and get  $\beta_0 = 50$ ,  $\beta_1 = 20$ ,  $\beta_2 = 0.07$ ,  $\beta_3 = 35$ ,  $\beta_4 = 0.01$ ,  $\beta_5 = -10$ .

(a) Predict the salary of a college graduate with IQ of 110 and a GPA of 4.0.

- (b) True or false: Since the coefficient for the GPA/IQ interaction term is very small, there is very little evidence of an interaction effect. Justify your answer.

A1A7E9B3-FACB-42FD-8588-177BE447A068

midterm1-b9ace

#44 Page 9 of 10



**Scrap Paper**





5584FF92-22F7-44A6-8C2F-10526A77F121

midterm1-b9ace

#44 Page 10 of 10



Michigan State University  
CMSE381 - Data Science

Feb 10, 2023  
Dr. Xie

## CMSE381 - Midterm #1

First name (please write as legibly as possible within the boxes)

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

Last name

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

NetID

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

*I will adhere to the Spartan Code of Honor in completing this assignment.*

Signed: \_\_\_\_\_

1. Do not open this test booklet until you are directed to do so.
2. You will have class time (2:40-4:00pm) to complete the exam.
3. This exam is closed book, but you can use the cheatsheet provided by the instructor.
4. Throughout the test, show your work so that your reasoning is clear. Otherwise no credit will be given. **BOX** your answers. Partial credit will be given where warranted.
5. Do not spend too much time on any one problem. Read them all through first and attack them in the order that allows you to make the most progress. Good luck :P



11A26651-9635-44A3-ABFC-435D1FEE716A

midterm1-b9ace

#45 Page 2 of 10

1. (15 points)

(a) Logistic regression is used for regression.

TRUE      FALSE

(b) Any model will never have training error below the irreducible error.

TRUE      FALSE

(c) Increasing your model flexibility always results in a better model.

TRUE      FALSE

(d) A logistic regression model is set up so that the odds are linear.

TRUE      FALSE

(e) Circle all of the following that would represent a qualitative variable.

Age      Year      Dog\_breed      Country\_of\_origin

Student\_(True/False)      Weight      Speed      MPG

(f) What equation would you use to evaluate the result of a regression model?

(g) What equation would you use to evaluate the result of a classification model?



2. (15 Points) I'm building a model to predict amount of a given brand of dog food eaten by a collection of dogs. I have 100 dogs eat this dog food, and I collect information on their height, weight, breed, and whether they live with another dog in the house.
- (a) List all input variables and specify whether they are quantitative or qualitative.
- (b) Say our dog breeds sampled are Huskies, Terriers, and Spaniels. Write down your prediction model.
- (c) We found that weight is a key predictor, and its relationship with the response is beyond linear. What extension can you come up with? Write down your updated model.



69E5B4BE-A600-4C24-8F5E-5C572AE309DB

midterm1-b9ace

#45 Page 4 of 10

3. (15 points)

(a) What is bias-variance tradeoff? Explain the meaning of each term in the formula.

(b) Provide a sketch of typical (squared) bias, variance, training error, test error, and Bayes (or irreducible) error curves, on a single plot, as we go from less flexible statistical learning methods towards more flexible approaches. The  $x$ -axis should represent the amount of flexibility in the method, and the  $y$ -axis should represent the values for each curve. There should be five curves. Make sure to label each one.



(c) Explain why the (i) training error and (ii) testing error lines in your drawing have the shape displayed.



4. (10 points) The data for this example come from a study by Stamey et al. (1989). The data comes from a number of clinical measures in men who were about to receive a radical prostatectomy. The goal is to predict the log of PSA (**lpsa**) (a prostate-specific antigen measured in nanograms of PSA per milliliter of blood (ng/mL)) from a number of measurements. The variables are log cancer volume (**lcavol**), log prostate weight (**lweight**), age, log of the amount of benign prostatic hyperplasia (**lbph**), seminal vesicle invasion (**svi**), log of capsular penetration (**lcp**), Gleason score (**gleason**), and percent of Gleason scores 4 or 5 (**pgg45**).

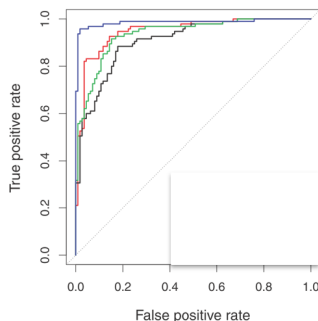
(a) Is this a case of regression or classification? Why?

(b) A linear model is fit to the data set, and the following table was returned.

Term	Coefficient	Std. Error	t-Score	p-value
Intercept	2.46	0.09	27.6	<0.00001
lcavol	0.68	0.13	5.37	<0.00001
lweight	0.26	0.1	2.75	0.00596
age	-0.14	0.1	-1.4	0.16153
lbph	0.21	0.1	2.06	0.039399
svi	0.31	0.12	2.47	0.013511
lcp	-0.29	0.15	-1.87	0.061484
gleason	-0.02	0.15	-0.15	0.880765
pgg45	0.27	0.15	1.74	0.081859

Are all the predictors useful? If yes, explain why. If not, explain what you would try to do next with the data.

5. (5 pts) We are trying to predict whether a patient will have a heart attack within one year. We have tried four different methods on a training dataset and have the following ROC curves. Which curve has the best performance on this training set? Justify your answer.





6. (15 Points) I get way too much email, so I decide to build a logistic regression model to predict whether a new incoming message is spam or not. I decide to just use a few variables:

$$\begin{aligned}X_1 &= \text{Number\_of\_sentences} \\X_2 &= \text{Number\_of\_references\_to\_a\_Nigerian\_Prince}\end{aligned}$$

and am training a logistic model to predict

$$Pr(Y = \text{spam} \mid X_1, X_2)$$

- (a) Write down the equation for the model you would train, using our standard notation with  $\beta_i$ 's.
- (b) If my trained model used  $\beta_0 = -13.1$ ,  $\beta_1 = 1.9$ , and  $\beta_2 = 6.1$ , what is the probability that a 5 sentence email with one reference to a Nigerian prince is spam? .
- (c) We know that if we miss an important email, it may cost a lot, How can you modify your model to avoid this?



7. (15 Points) Sparty generated 100 pair  $x$  and  $y$  via the following relationship:  $Y = 0.3 + 2x + x^2 + \epsilon$  but didn't tell you. You will try to find a good model to fit these data and more importantly to predict future response  $y$  when Sparty gives you another  $x$ . You will start with two models 1)  $Y = \beta_0 + \beta_1x$ ; 2)  $Y = \alpha_0 + \alpha_1x + \alpha_2x^2$ ; and 3)  $Y = \gamma_0 + \gamma_1x + \gamma_2x^2 + \gamma_3x^3$ .
- (a) If you use all 100 data to fit these three model and use the same 100 data to calculate MSE, which model will have the smallest MSE? Justify your answer
- (b) Since we focus on the ability to predict unseen data, you decide use validation set to evaluate the performance of the three models. There are two way to split the data: 1) 80 testing points and 20 training points or 2) 20 testing points and 80 training points. Which one will you choose? Justify your choice
- (c) From the in-class lab, we know validation set is not a good choice to choose model. Explain the drawback of validation set and what procedure will you use to choose the best model?





8. (10 Points) Suppose we have a data set with five predictors,  $X_1 = \text{GPA}$ ,  $X_2 = \text{IQ}$ ,  $X_3 = \text{Level}$  (1 for College and 0 for High School),  $X_4 = \text{Interaction between GPA and IQ}$ , and  $X_5 = \text{Interaction between GPA and Level}$ . The response is starting salary after graduation (in thousands of dollars). Suppose we use least squares to fit the model, and get  $\beta_0 = 50$ ,  $\beta_1 = 20$ ,  $\beta_2 = 0.07$ ,  $\beta_3 = 35$ ,  $\beta_4 = 0.01$ ,  $\beta_5 = -10$ .

(a) Predict the salary of a college graduate with IQ of 110 and a GPA of 4.0.

(b) True or false: Since the coefficient for the GPA/IQ interaction term is very small, there is very little evidence of an interaction effect. Justify your answer.

E8A1A3FE-AAF4-4BD7-AEF7-1557319AFC38

midterm1-b9ace

#45 Page 9 of 10



**Scrap Paper**



5342585F-6D88-409E-9976-87F580B6E972

midterm1-b9ace

#45 Page 10 of 10



Michigan State University  
CMSE381 - Data Science

Feb 10, 2023  
Dr. Xie

## CMSE381 - Midterm #1

First name (please write as legibly as possible within the boxes)

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

Last name

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

NetID

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

*I will adhere to the Spartan Code of Honor in completing this assignment.*

Signed: \_\_\_\_\_

1. Do not open this test booklet until you are directed to do so.
2. You will have class time (2:40-4:00pm) to complete the exam.
3. This exam is closed book, but you can use the cheatsheet provided by the instructor.
4. Throughout the test, show your work so that your reasoning is clear. Otherwise no credit will be given. **BOX** your answers. Partial credit will be given where warranted.
5. Do not spend too much time on any one problem. Read them all through first and attack them in the order that allows you to make the most progress. Good luck :P



6216425F-44F5-49DA-BAFB-D7CC1366C356

midterm1-b9ace

#46 Page 2 of 10

1. (15 points)

(a) Logistic regression is used for regression.

TRUE      FALSE

(b) Any model will never have training error below the irreducible error.

TRUE      FALSE

(c) Increasing your model flexibility always results in a better model.

TRUE      FALSE

(d) A logistic regression model is set up so that the odds are linear.

TRUE      FALSE

(e) Circle all of the following that would represent a qualitative variable.

Age      Year      Dog\_breed      Country\_of\_origin

Student\_(True/False)      Weight      Speed      MPG

(f) What equation would you use to evaluate the result of a regression model?

(g) What equation would you use to evaluate the result of a classification model?



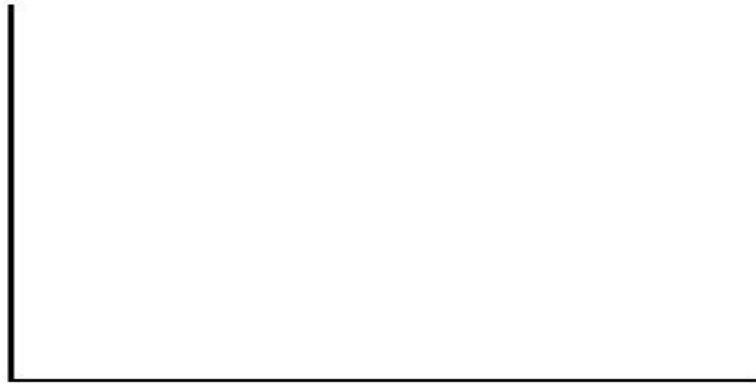
2. (15 Points) I'm building a model to predict amount of a given brand of dog food eaten by a collection of dogs. I have 100 dogs eat this dog food, and I collect information on their height, weight, breed, and whether they live with another dog in the house.
- (a) List all input variables and specify whether they are quantitative or qualitative.
  - (b) Say our dog breeds sampled are Huskies, Terriers, and Spaniels. Write down your prediction model.
  - (c) We found that weight is a key predictor, and its relationship with the response is beyond linear. What extension can you come up with? Write down your updated model.



3. (15 points)

(a) What is bias-variance tradeoff? Explain the meaning of each term in the formula.

(b) Provide a sketch of typical (squared) bias, variance, training error, test error, and Bayes (or irreducible) error curves, on a single plot, as we go from less flexible statistical learning methods towards more flexible approaches. The  $x$ -axis should represent the amount of flexibility in the method, and the  $y$ -axis should represent the values for each curve. There should be five curves. Make sure to label each one.



(c) Explain why the (i) training error and (ii) testing error lines in your drawing have the shape displayed.



4. (10 points) The data for this example come from a study by Stamey et al. (1989). The data comes from a number of clinical measures in men who were about to receive a radical prostatectomy. The goal is to predict the log of PSA (**lpsa**) (a prostate-specific antigen measured in nanograms of PSA per milliliter of blood (ng/mL)) from a number of measurements. The variables are log cancer volume (**lcavol**), log prostate weight (**lweight**), age, log of the amount of benign prostatic hyperplasia (**lbph**), seminal vesicle invasion (**svi**), log of capsular penetration (**lcp**), Gleason score (**gleason**), and percent of Gleason scores 4 or 5 (**pgg45**).

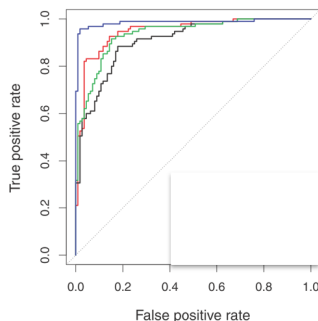
(a) Is this a case of regression or classification? Why?

(b) A linear model is fit to the data set, and the following table was returned.

Term	Coefficient	Std. Error	t-Score	p-value
Intercept	2.46	0.09	27.6	<0.00001
lcavol	0.68	0.13	5.37	<0.00001
lweight	0.26	0.1	2.75	0.00596
age	-0.14	0.1	-1.4	0.16153
lbph	0.21	0.1	2.06	0.039399
svi	0.31	0.12	2.47	0.013511
lcp	-0.29	0.15	-1.87	0.061484
gleason	-0.02	0.15	-0.15	0.880765
pgg45	0.27	0.15	1.74	0.081859

Are all the predictors useful? If yes, explain why. If not, explain what you would try to do next with the data.

5. (5 pts) We are trying to predict whether a patient will have a heart attack within one year. We have tried four different methods on a training dataset and have the following ROC curves. Which curve has the best performance on this training set? Justify your answer.







BE4A158D-59BB-4891-B881-631CADEA485E

midterm1-b9ace

#46 Page 6 of 10

6. (15 Points) I get way too much email, so I decide to build a logistic regression model to predict whether a new incoming message is spam or not. I decide to just use a few variables:

$$\begin{aligned}X_1 &= \text{Number\_of\_sentences} \\X_2 &= \text{Number\_of\_references\_to\_a\_Nigerian\_Prince}\end{aligned}$$

and am training a logistic model to predict

$$Pr(Y = \text{spam} \mid X_1, X_2)$$

- (a) Write down the equation for the model you would train, using our standard notation with  $\beta_i$ 's.
- (b) If my trained model used  $\beta_0 = -13.1$ ,  $\beta_1 = 1.9$ , and  $\beta_2 = 6.1$ , what is the probability that a 5 sentence email with one reference to a Nigerian prince is spam? .
- (c) We know that if we miss an important email, it may cost a lot, How can you modify your model to avoid this?



7. (15 Points) Sparty generated 100 pair  $x$  and  $y$  via the following relationship:  $Y = 0.3 + 2x + x^2 + \epsilon$  but didn't tell you. You will try to find a good model to fit these data and more importantly to predict future response  $y$  when Sparty gives you another  $x$ . You will start with two models 1)  $Y = \beta_0 + \beta_1x$ ; 2)  $Y = \alpha_0 + \alpha_1x + \alpha_2x^2$ ; and 3)  $Y = \gamma_0 + \gamma_1x + \gamma_2x^2 + \gamma_3x^3$ .
- (a) If you use all 100 data to fit these three model and use the same 100 data to calculate MSE, which model will have the smallest MSE? Justify your answer
- (b) Since we focus on the ability to predict unseen data, you decide use validation set to evaluate the performance of the three models. There are two way to split the data: 1) 80 testing points and 20 training points or 2) 20 testing points and 80 training points. Which one will you choose? Justify your choice
- (c) From the in-class lab, we know validation set is not a good choice to choose model. Explain the drawback of validation set and what procedure will you use to choose the best model?



8. (10 Points) Suppose we have a data set with five predictors,  $X_1 = \text{GPA}$ ,  $X_2 = \text{IQ}$ ,  $X_3 = \text{Level}$  (1 for College and 0 for High School),  $X_4 = \text{Interaction between GPA and IQ}$ , and  $X_5 = \text{Interaction between GPA and Level}$ . The response is starting salary after graduation (in thousands of dollars). Suppose we use least squares to fit the model, and get  $\beta_0 = 50$ ,  $\beta_1 = 20$ ,  $\beta_2 = 0.07$ ,  $\beta_3 = 35$ ,  $\beta_4 = 0.01$ ,  $\beta_5 = -10$ .

(a) Predict the salary of a college graduate with IQ of 110 and a GPA of 4.0.

- (b) True or false: Since the coefficient for the GPA/IQ interaction term is very small, there is very little evidence of an interaction effect. Justify your answer.

3CEEB73B-587F-439F-9042-6509502EDE36

midterm1-b9ace

#46 Page 9 of 10



**Scrap Paper**



B5FAC395-86F1-43B2-B8A4-6FDBC30EE163

midterm1-b9ace

#46 Page 10 of 10



Michigan State University  
CMSE381 - Data Science

Feb 10, 2023  
Dr. Xie

## CMSE381 - Midterm #1

First name (please write as legibly as possible within the boxes)

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

Last name

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

NetID

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

*I will adhere to the Spartan Code of Honor in completing this assignment.*

Signed: \_\_\_\_\_

1. Do not open this test booklet until you are directed to do so.
2. You will have class time (2:40-4:00pm) to complete the exam.
3. This exam is closed book, but you can use the cheatsheet provided by the instructor.
4. Throughout the test, show your work so that your reasoning is clear. Otherwise no credit will be given. **BOX** your answers. Partial credit will be given where warranted.
5. Do not spend too much time on any one problem. Read them all through first and attack them in the order that allows you to make the most progress. Good luck :P



0DCEF0C5-72BF-42B3-A510-3756F5D2067E

midterm1-b9ace

#47 Page 2 of 10

1. (15 points)

(a) Logistic regression is used for regression.

TRUE      FALSE

(b) Any model will never have training error below the irreducible error.

TRUE      FALSE

(c) Increasing your model flexibility always results in a better model.

TRUE      FALSE

(d) A logistic regression model is set up so that the odds are linear.

TRUE      FALSE

(e) Circle all of the following that would represent a qualitative variable.

Age      Year      Dog\_breed      Country\_of\_origin

Student\_(True/False)      Weight      Speed      MPG

(f) What equation would you use to evaluate the result of a regression model?

(g) What equation would you use to evaluate the result of a classification model?



2. (15 Points) I'm building a model to predict amount of a given brand of dog food eaten by a collection of dogs. I have 100 dogs eat this dog food, and I collect information on their height, weight, breed, and whether they live with another dog in the house.
  - (a) List all input variables and specify whether they are quantitative or qualitative.
  - (b) Say our dog breeds sampled are Huskies, Terriers, and Spaniels. Write down your prediction model.
  - (c) We found that weight is a key predictor, and its relationship with the response is beyond linear. What extension can you come up with? Write down your updated model.





F36C3D8D-FC90-47A5-8FF8-E8E0F392FF43

midterm1-b9ace

#47 Page 4 of 10

3. (15 points)

(a) What is bias-variance tradeoff? Explain the meaning of each term in the formula.

(b) Provide a sketch of typical (squared) bias, variance, training error, test error, and Bayes (or irreducible) error curves, on a single plot, as we go from less flexible statistical learning methods towards more flexible approaches. The  $x$ -axis should represent the amount of flexibility in the method, and the  $y$ -axis should represent the values for each curve. There should be five curves. Make sure to label each one.



(c) Explain why the (i) training error and (ii) testing error lines in your drawing have the shape displayed.



4. (10 points) The data for this example come from a study by Stamey et al. (1989). The data comes from a number of clinical measures in men who were about to receive a radical prostatectomy. The goal is to predict the log of PSA (**lpsa**) (a prostate-specific antigen measured in nanograms of PSA per milliliter of blood (ng/mL)) from a number of measurements. The variables are log cancer volume (**lcavol**), log prostate weight (**lweight**), age, log of the amount of benign prostatic hyperplasia (**lbph**), seminal vesicle invasion (**svi**), log of capsular penetration (**lcp**), Gleason score (**gleason**), and percent of Gleason scores 4 or 5 (**pgg45**).

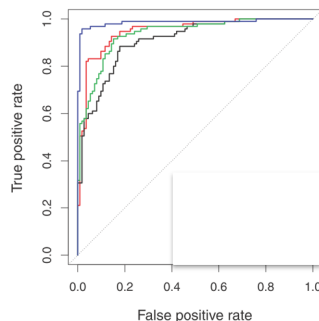
(a) Is this a case of regression or classification? Why?

(b) A linear model is fit to the data set, and the following table was returned.

Term	Coefficient	Std. Error	t-Score	p-value
Intercept	2.46	0.09	27.6	<0.00001
lcavol	0.68	0.13	5.37	<0.00001
lweight	0.26	0.1	2.75	0.00596
age	-0.14	0.1	-1.4	0.16153
lbph	0.21	0.1	2.06	0.039399
svi	0.31	0.12	2.47	0.013511
lcp	-0.29	0.15	-1.87	0.061484
gleason	-0.02	0.15	-0.15	0.880765
pgg45	0.27	0.15	1.74	0.081859

Are all the predictors useful? If yes, explain why. If not, explain what you would try to do next with the data.

5. (5 pts) We are trying to predict whether a patient will have a heart attack within one year. We have tried four different methods on a training dataset and have the following ROC curves. Which curve has the best performance on this training set? Justify your answer.





9489E03C-4D61-48BB-A940-DA0D8AD2DF03

midterm1-b9ace

#47 Page 6 of 10

6. (15 Points) I get way too much email, so I decide to build a logistic regression model to predict whether a new incoming message is spam or not. I decide to just use a few variables:

$$\begin{aligned} X_1 &= \text{Number\_of\_sentences} \\ X_2 &= \text{Number\_of\_references\_to\_a\_Nigerian\_Prince} \end{aligned}$$

and am training a logistic model to predict

$$Pr(Y = \text{spam} \mid X_1, X_2)$$

- (a) Write down the equation for the model you would train, using our standard notation with  $\beta_i$ 's.
- (b) If my trained model used  $\beta_0 = -13.1$ ,  $\beta_1 = 1.9$ , and  $\beta_2 = 6.1$ , what is the probability that a 5 sentence email with one reference to a Nigerian prince is spam? .
- (c) We know that if we miss an important email, it may cost a lot, How can you modify your model to avoid this?



7. (15 Points) Sparty generated 100 pair  $x$  and  $y$  via the following relationship:  $Y = 0.3 + 2x + x^2 + \epsilon$  but didn't tell you. You will try to find a good model to fit these data and more importantly to predict future response  $y$  when Sparty gives you another  $x$ . You will start with two models 1)  $Y = \beta_0 + \beta_1x$ ; 2)  $Y = \alpha_0 + \alpha_1x + \alpha_2x^2$ ; and 3)  $Y = \gamma_0 + \gamma_1x + \gamma_2x^2 + \gamma_3x^3$ .
- (a) If you use all 100 data to fit these three model and use the same 100 data to calculate MSE, which model will have the smallest MSE? Justify your answer
- (b) Since we focus on the ability to predict unseen data, you decide use validation set to evaluate the performance of the three models. There are two way to split the data: 1) 80 testing points and 20 training points or 2) 20 testing points and 80 training points. Which one will you choose? Justify your choice
- (c) From the in-class lab, we know validation set is not a good choice to choose model. Explain the drawback of validation set and what procedure will you use to choose the best model?



04BF0B0D-FC50-48AD-9801-E5CBD4262F77

midterm1-b9ace

#47 Page 8 of 10

8. (10 Points) Suppose we have a data set with five predictors,  $X_1 = \text{GPA}$ ,  $X_2 = \text{IQ}$ ,  $X_3 = \text{Level}$  (1 for College and 0 for High School),  $X_4 = \text{Interaction between GPA and IQ}$ , and  $X_5 = \text{Interaction between GPA and Level}$ . The response is starting salary after graduation (in thousands of dollars). Suppose we use least squares to fit the model, and get  $\beta_0 = 50$ ,  $\beta_1 = 20$ ,  $\beta_2 = 0.07$ ,  $\beta_3 = 35$ ,  $\beta_4 = 0.01$ ,  $\beta_5 = -10$ .

(a) Predict the salary of a college graduate with IQ of 110 and a GPA of 4.0.

- (b) True or false: Since the coefficient for the GPA/IQ interaction term is very small, there is very little evidence of an interaction effect. Justify your answer.

1E82EC0B-120E-48FE-9F0A-639A04AB4C28

midterm1-b9ace

#47 Page 9 of 10



## Scrap Paper



94B5741E-CBEB-4658-A353-B382C7DF7B01

midterm1-b9ace

#47 Page 10 of 10



Michigan State University  
CMSE381 - Data Science

Feb 10, 2023  
Dr. Xie

## CMSE381 - Midterm #1

First name (please write as legibly as possible within the boxes)

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

Last name

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

NetID

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

*I will adhere to the Spartan Code of Honor in completing this assignment.*

Signed: \_\_\_\_\_

1. Do not open this test booklet until you are directed to do so.
2. You will have class time (2:40-4:00pm) to complete the exam.
3. This exam is closed book, but you can use the cheatsheet provided by the instructor.
4. Throughout the test, show your work so that your reasoning is clear. Otherwise no credit will be given. **BOX** your answers. Partial credit will be given where warranted.
5. Do not spend too much time on any one problem. Read them all through first and attack them in the order that allows you to make the most progress. Good luck :P





4AD71E1E-A61E-4E21-AB08-EF98C2660C04

midterm1-b9ace

#48 Page 2 of 10

1. (15 points)

(a) Logistic regression is used for regression.

TRUE      FALSE

(b) Any model will never have training error below the irreducible error.

TRUE      FALSE

(c) Increasing your model flexibility always results in a better model.

TRUE      FALSE

(d) A logistic regression model is set up so that the odds are linear.

TRUE      FALSE

(e) Circle all of the following that would represent a qualitative variable.

Age      Year      Dog\_breed      Country\_of\_origin

Student\_(True/False)      Weight      Speed      MPG

(f) What equation would you use to evaluate the result of a regression model?

(g) What equation would you use to evaluate the result of a classification model?



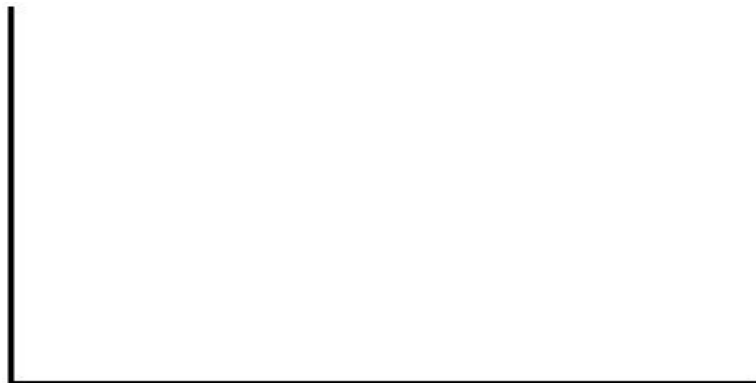
2. (15 Points) I'm building a model to predict amount of a given brand of dog food eaten by a collection of dogs. I have 100 dogs eat this dog food, and I collect information on their height, weight, breed, and whether they live with another dog in the house.
- (a) List all input variables and specify whether they are quantitative or qualitative.
- (b) Say our dog breeds sampled are Huskies, Terriers, and Spaniels. Write down your prediction model.
- (c) We found that weight is a key predictor, and its relationship with the response is beyond linear. What extension can you come up with? Write down your updated model.



3. (15 points)

(a) What is bias-variance tradeoff? Explain the meaning of each term in the formula.

(b) Provide a sketch of typical (squared) bias, variance, training error, test error, and Bayes (or irreducible) error curves, on a single plot, as we go from less flexible statistical learning methods towards more flexible approaches. The  $x$ -axis should represent the amount of flexibility in the method, and the  $y$ -axis should represent the values for each curve. There should be five curves. Make sure to label each one.



(c) Explain why the (i) training error and (ii) testing error lines in your drawing have the shape displayed.



4. (10 points) The data for this example come from a study by Stamey et al. (1989). The data comes from a number of clinical measures in men who were about to receive a radical prostatectomy. The goal is to predict the log of PSA (**lpsa**) (a prostate-specific antigen measured in nanograms of PSA per milliliter of blood (ng/mL)) from a number of measurements. The variables are log cancer volume (**lcavol**), log prostate weight (**lweight**), age, log of the amount of benign prostatic hyperplasia (**lbph**), seminal vesicle invasion (**svi**), log of capsular penetration (**lcp**), Gleason score (**gleason**), and percent of Gleason scores 4 or 5 (**pgg45**).

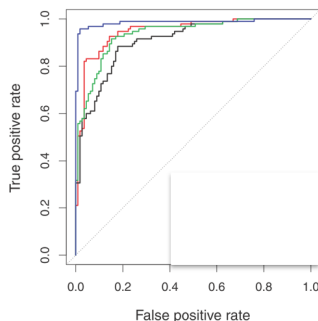
(a) Is this a case of regression or classification? Why?

(b) A linear model is fit to the data set, and the following table was returned.

Term	Coefficient	Std. Error	t-Score	p-value
Intercept	2.46	0.09	27.6	<0.00001
lcavol	0.68	0.13	5.37	<0.00001
lweight	0.26	0.1	2.75	0.00596
age	-0.14	0.1	-1.4	0.16153
lbph	0.21	0.1	2.06	0.039399
svi	0.31	0.12	2.47	0.013511
lcp	-0.29	0.15	-1.87	0.061484
gleason	-0.02	0.15	-0.15	0.880765
pgg45	0.27	0.15	1.74	0.081859

Are all the predictors useful? If yes, explain why. If not, explain what you would try to do next with the data.

5. (5 pts) We are trying to predict whether a patient will have a heart attack within one year. We have tried four different methods on a training dataset and have the following ROC curves. Which curve has the best performance on this training set? Justify your answer.





1F833330-33D7-4904-9269-C779AEB84E3B

midterm1-b9ace

#48 Page 6 of 10

6. (15 Points) I get way too much email, so I decide to build a logistic regression model to predict whether a new incoming message is spam or not. I decide to just use a few variables:

$$\begin{aligned}X_1 &= \text{Number\_of\_sentences} \\X_2 &= \text{Number\_of\_references\_to\_a\_Nigerian\_Prince}\end{aligned}$$

and am training a logistic model to predict

$$Pr(Y = \text{spam} \mid X_1, X_2)$$

- (a) Write down the equation for the model you would train, using our standard notation with  $\beta_i$ 's.
- (b) If my trained model used  $\beta_0 = -13.1$ ,  $\beta_1 = 1.9$ , and  $\beta_2 = 6.1$ , what is the probability that a 5 sentence email with one reference to a Nigerian prince is spam? .
- (c) We know that if we miss an important email, it may cost a lot, How can you modify your model to avoid this?



7. (15 Points) Sparty generated 100 pair  $x$  and  $y$  via the following relationship:  $Y = 0.3 + 2x + x^2 + \epsilon$  but didn't tell you. You will try to find a good model to fit these data and more importantly to predict future response  $y$  when Sparty gives you another  $x$ . You will start with two models 1)  $Y = \beta_0 + \beta_1x$ ; 2)  $Y = \alpha_0 + \alpha_1x + \alpha_2x^2$ ; and 3)  $Y = \gamma_0 + \gamma_1x + \gamma_2x^2 + \gamma_3x^3$ .
- (a) If you use all 100 data to fit these three model and use the same 100 data to calculate MSE, which model will have the smallest MSE? Justify your answer
- (b) Since we focus on the ability to predict unseen data, you decide use validation set to evaluate the performance of the three models. There are two way to split the data: 1) 80 testing points and 20 training points or 2) 20 testing points and 80 training points. Which one will you choose? Justify your choice
- (c) From the in-class lab, we know validation set is not a good choice to choose model. Explain the drawback of validation set and what procedure will you use to choose the best model?



880459A9-8B45-4008-9D70-54CACD209F86

midterm1-b9ace

#48 Page 8 of 10

8. (10 Points) Suppose we have a data set with five predictors,  $X_1 = \text{GPA}$ ,  $X_2 = \text{IQ}$ ,  $X_3 = \text{Level}$  (1 for College and 0 for High School),  $X_4 = \text{Interaction between GPA and IQ}$ , and  $X_5 = \text{Interaction between GPA and Level}$ . The response is starting salary after graduation (in thousands of dollars). Suppose we use least squares to fit the model, and get  $\beta_0 = 50$ ,  $\beta_1 = 20$ ,  $\beta_2 = 0.07$ ,  $\beta_3 = 35$ ,  $\beta_4 = 0.01$ ,  $\beta_5 = -10$ .

(a) Predict the salary of a college graduate with IQ of 110 and a GPA of 4.0.

(b) True or false: Since the coefficient for the GPA/IQ interaction term is very small, there is very little evidence of an interaction effect. Justify your answer.

7FE089CC-906D-4F60-948A-B0F4F9409731

midterm1-b9ace

#48 Page 9 of 10



**Scrap Paper**





47CB92C8-6135-483A-A5A7-7FD30664B117

midterm1-b9ace

#48 Page 10 of 10



Michigan State University  
CMSE381 - Data Science

Feb 10, 2023  
Dr. Xie

## CMSE381 - Midterm #1

First name (please write as legibly as possible within the boxes)

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

Last name

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

NetID

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

*I will adhere to the Spartan Code of Honor in completing this assignment.*

Signed: \_\_\_\_\_

1. Do not open this test booklet until you are directed to do so.
2. You will have class time (2:40-4:00pm) to complete the exam.
3. This exam is closed book, but you can use the cheatsheet provided by the instructor.
4. Throughout the test, show your work so that your reasoning is clear. Otherwise no credit will be given. **BOX** your answers. Partial credit will be given where warranted.
5. Do not spend too much time on any one problem. Read them all through first and attack them in the order that allows you to make the most progress. Good luck :P



16939696-6322-46CD-8503-B18B67EB4F5C

midterm1-b9ace

#49 Page 2 of 10

1. (15 points)

(a) Logistic regression is used for regression.

TRUE      FALSE

(b) Any model will never have training error below the irreducible error.

TRUE      FALSE

(c) Increasing your model flexibility always results in a better model.

TRUE      FALSE

(d) A logistic regression model is set up so that the odds are linear.

TRUE      FALSE

(e) Circle all of the following that would represent a qualitative variable.

Age      Year      Dog\_breed      Country\_of\_origin

Student\_(True/False)      Weight      Speed      MPG

(f) What equation would you use to evaluate the result of a regression model?

(g) What equation would you use to evaluate the result of a classification model?



2. (15 Points) I'm building a model to predict amount of a given brand of dog food eaten by a collection of dogs. I have 100 dogs eat this dog food, and I collect information on their height, weight, breed, and whether they live with another dog in the house.
- (a) List all input variables and specify whether they are quantitative or qualitative.
- (b) Say our dog breeds sampled are Huskies, Terriers, and Spaniels. Write down your prediction model.
- (c) We found that weight is a key predictor, and its relationship with the response is beyond linear. What extension can you come up with? Write down your updated model.



651E329C-5A69-40D3-8846-C59901024B08

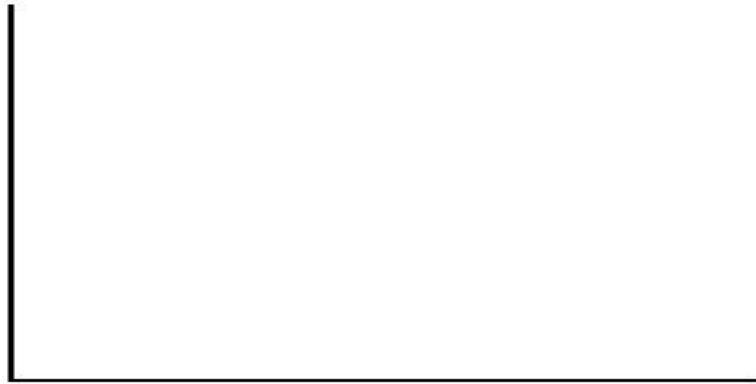
midterm1-b9ace

#49 Page 4 of 10

3. (15 points)

(a) What is bias-variance tradeoff? Explain the meaning of each term in the formula.

(b) Provide a sketch of typical (squared) bias, variance, training error, test error, and Bayes (or irreducible) error curves, on a single plot, as we go from less flexible statistical learning methods towards more flexible approaches. The  $x$ -axis should represent the amount of flexibility in the method, and the  $y$ -axis should represent the values for each curve. There should be five curves. Make sure to label each one.



(c) Explain why the (i) training error and (ii) testing error lines in your drawing have the shape displayed.



4. (10 points) The data for this example come from a study by Stamey et al. (1989). The data comes from a number of clinical measures in men who were about to receive a radical prostatectomy. The goal is to predict the log of PSA (**lpsa**) (a prostate-specific antigen measured in nanograms of PSA per milliliter of blood (ng/mL)) from a number of measurements. The variables are log cancer volume (**lcavol**), log prostate weight (**lweight**), age, log of the amount of benign prostatic hyperplasia (**lbph**), seminal vesicle invasion (**svi**), log of capsular penetration (**lcp**), Gleason score (**gleason**), and percent of Gleason scores 4 or 5 (**pgg45**).

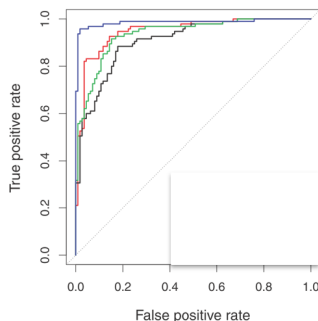
(a) Is this a case of regression or classification? Why?

(b) A linear model is fit to the data set, and the following table was returned.

Term	Coefficient	Std. Error	t-Score	p-value
Intercept	2.46	0.09	27.6	<0.00001
lcavol	0.68	0.13	5.37	<0.00001
lweight	0.26	0.1	2.75	0.00596
age	-0.14	0.1	-1.4	0.16153
lbph	0.21	0.1	2.06	0.039399
svi	0.31	0.12	2.47	0.013511
lcp	-0.29	0.15	-1.87	0.061484
gleason	-0.02	0.15	-0.15	0.880765
pgg45	0.27	0.15	1.74	0.081859

Are all the predictors useful? If yes, explain why. If not, explain what you would try to do next with the data.

5. (5 pts) We are trying to predict whether a patient will have a heart attack within one year. We have tried four different methods on a training dataset and have the following ROC curves. Which curve has the best performance on this training set? Justify your answer.





6. (15 Points) I get way too much email, so I decide to build a logistic regression model to predict whether a new incoming message is spam or not. I decide to just use a few variables:

$$\begin{aligned}X_1 &= \text{Number\_of\_sentences} \\X_2 &= \text{Number\_of\_references\_to\_a\_Nigerian\_Prince}\end{aligned}$$

and am training a logistic model to predict

$$Pr(Y = \text{spam} \mid X_1, X_2)$$

- (a) Write down the equation for the model you would train, using our standard notation with  $\beta_i$ 's.
- (b) If my trained model used  $\beta_0 = -13.1$ ,  $\beta_1 = 1.9$ , and  $\beta_2 = 6.1$ , what is the probability that a 5 sentence email with one reference to a Nigerian prince is spam? .
- (c) We know that if we miss an important email, it may cost a lot, How can you modify your model to avoid this?



7. (15 Points) Sparty generated 100 pair  $x$  and  $y$  via the following relationship:  $Y = 0.3 + 2x + x^2 + \epsilon$  but didn't tell you. You will try to find a good model to fit these data and more importantly to predict future response  $y$  when Sparty gives you another  $x$ . You will start with two models 1)  $Y = \beta_0 + \beta_1x$ ; 2)  $Y = \alpha_0 + \alpha_1x + \alpha_2x^2$ ; and 3)  $Y = \gamma_0 + \gamma_1x + \gamma_2x^2 + \gamma_3x^3$ .
- (a) If you use all 100 data to fit these three model and use the same 100 data to calculate MSE, which model will have the smallest MSE? Justify your answer
- (b) Since we focus on the ability to predict unseen data, you decide use validation set to evaluate the performance of the three models. There are two way to split the data: 1) 80 testing points and 20 training points or 2) 20 testing points and 80 training points. Which one will you choose? Justify your choice
- (c) From the in-class lab, we know validation set is not a good choice to choose model. Explain the drawback of validation set and what procedure will you use to choose the best model?





8. (10 Points) Suppose we have a data set with five predictors,  $X_1 = \text{GPA}$ ,  $X_2 = \text{IQ}$ ,  $X_3 = \text{Level}$  (1 for College and 0 for High School),  $X_4 = \text{Interaction between GPA and IQ}$ , and  $X_5 = \text{Interaction between GPA and Level}$ . The response is starting salary after graduation (in thousands of dollars). Suppose we use least squares to fit the model, and get  $\beta_0 = 50$ ,  $\beta_1 = 20$ ,  $\beta_2 = 0.07$ ,  $\beta_3 = 35$ ,  $\beta_4 = 0.01$ ,  $\beta_5 = -10$ .

(a) Predict the salary of a college graduate with IQ of 110 and a GPA of 4.0.

- (b) True or false: Since the coefficient for the GPA/IQ interaction term is very small, there is very little evidence of an interaction effect. Justify your answer.

EFA3F332-19CF-4A57-B7C4-31AD0B7EABE2

midterm1-b9ace

#49 Page 9 of 10



## Scrap Paper



E6BA8EA8-C344-432C-9A66-752A4A7589DB

midterm1-b9ace

#49 Page 10 of 10



Michigan State University  
CMSE381 - Data Science

Feb 10, 2023  
Dr. Xie

## CMSE381 - Midterm #1

First name (please write as legibly as possible within the boxes)

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

Last name

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

NetID

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

*I will adhere to the Spartan Code of Honor in completing this assignment.*

Signed: \_\_\_\_\_

1. Do not open this test booklet until you are directed to do so.
2. You will have class time (2:40-4:00pm) to complete the exam.
3. This exam is closed book, but you can use the cheatsheet provided by the instructor.
4. Throughout the test, show your work so that your reasoning is clear. Otherwise no credit will be given. **BOX** your answers. Partial credit will be given where warranted.
5. Do not spend too much time on any one problem. Read them all through first and attack them in the order that allows you to make the most progress. Good luck :P



1ADC08AA-58DD-4920-BFB8-2E6CD7BE6068

midterm1-b9ace

#50 Page 2 of 10

1. (15 points)

(a) Logistic regression is used for regression.

TRUE      FALSE

(b) Any model will never have training error below the irreducible error.

TRUE      FALSE

(c) Increasing your model flexibility always results in a better model.

TRUE      FALSE

(d) A logistic regression model is set up so that the odds are linear.

TRUE      FALSE

(e) Circle all of the following that would represent a qualitative variable.

Age      Year      Dog\_breed      Country\_of\_origin

Student\_(True/False)      Weight      Speed      MPG

(f) What equation would you use to evaluate the result of a regression model?

(g) What equation would you use to evaluate the result of a classification model?



2. (15 Points) I'm building a model to predict amount of a given brand of dog food eaten by a collection of dogs. I have 100 dogs eat this dog food, and I collect information on their height, weight, breed, and whether they live with another dog in the house.
- (a) List all input variables and specify whether they are quantitative or qualitative.
- (b) Say our dog breeds sampled are Huskies, Terriers, and Spaniels. Write down your prediction model.
- (c) We found that weight is a key predictor, and its relationship with the response is beyond linear. What extension can you come up with? Write down your updated model.



7CAE864B-D13E-432D-8223-CE5C41669B27

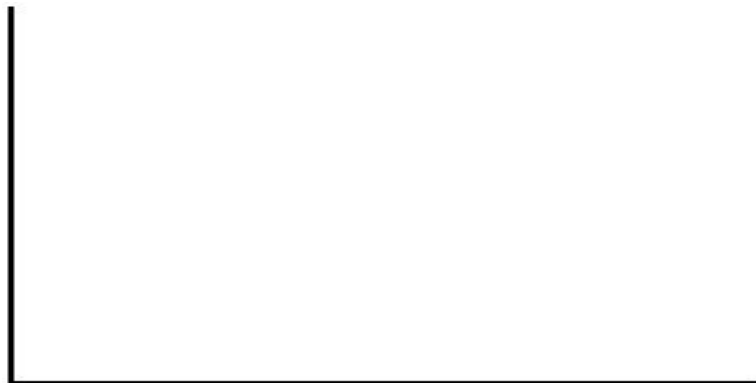
midterm1-b9ace

#50 Page 4 of 10

3. (15 points)

(a) What is bias-variance tradeoff? Explain the meaning of each term in the formula.

(b) Provide a sketch of typical (squared) bias, variance, training error, test error, and Bayes (or irreducible) error curves, on a single plot, as we go from less flexible statistical learning methods towards more flexible approaches. The  $x$ -axis should represent the amount of flexibility in the method, and the  $y$ -axis should represent the values for each curve. There should be five curves. Make sure to label each one.



(c) Explain why the (i) training error and (ii) testing error lines in your drawing have the shape displayed.



4. (10 points) The data for this example come from a study by Stamey et al. (1989). The data comes from a number of clinical measures in men who were about to receive a radical prostatectomy. The goal is to predict the log of PSA (**lpsa**) (a prostate-specific antigen measured in nanograms of PSA per milliliter of blood (ng/mL)) from a number of measurements. The variables are log cancer volume (**lcavol**), log prostate weight (**lweight**), age, log of the amount of benign prostatic hyperplasia (**lbph**), seminal vesicle invasion (**svi**), log of capsular penetration (**lcp**), Gleason score (**gleason**), and percent of Gleason scores 4 or 5 (**pgg45**).

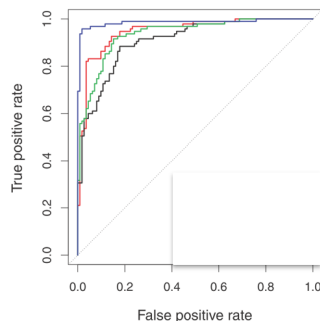
(a) Is this a case of regression or classification? Why?

(b) A linear model is fit to the data set, and the following table was returned.

Term	Coefficient	Std. Error	t-Score	p-value
Intercept	2.46	0.09	27.6	<0.00001
lcavol	0.68	0.13	5.37	<0.00001
lweight	0.26	0.1	2.75	0.00596
age	-0.14	0.1	-1.4	0.16153
lbph	0.21	0.1	2.06	0.039399
svi	0.31	0.12	2.47	0.013511
lcp	-0.29	0.15	-1.87	0.061484
gleason	-0.02	0.15	-0.15	0.880765
pgg45	0.27	0.15	1.74	0.081859

Are all the predictors useful? If yes, explain why. If not, explain what you would try to do next with the data.

5. (5 pts) We are trying to predict whether a patient will have a heart attack within one year. We have tried four different methods on a training dataset and have the following ROC curves. Which curve has the best performance on this training set? Justify your answer.







3B794A4E-23E9-43E8-A06D-A1CCE92FF7F6

midterm1-b9ace

#50 Page 6 of 10

6. (15 Points) I get way too much email, so I decide to build a logistic regression model to predict whether a new incoming message is spam or not. I decide to just use a few variables:

$$\begin{aligned} X_1 &= \text{Number\_of\_sentences} \\ X_2 &= \text{Number\_of\_references\_to\_a\_Nigerian\_Prince} \end{aligned}$$

and am training a logistic model to predict

$$Pr(Y = \text{spam} \mid X_1, X_2)$$

- (a) Write down the equation for the model you would train, using our standard notation with  $\beta_i$ 's.
- (b) If my trained model used  $\beta_0 = -13.1$ ,  $\beta_1 = 1.9$ , and  $\beta_2 = 6.1$ , what is the probability that a 5 sentence email with one reference to a Nigerian prince is spam? .
- (c) We know that if we miss an important email, it may cost a lot, How can you modify your model to avoid this?



7. (15 Points) Sparty generated 100 pair  $x$  and  $y$  via the following relationship:  $Y = 0.3 + 2x + x^2 + \epsilon$  but didn't tell you. You will try to find a good model to fit these data and more importantly to predict future response  $y$  when Sparty gives you another  $x$ . You will start with two models 1)  $Y = \beta_0 + \beta_1x$ ; 2)  $Y = \alpha_0 + \alpha_1x + \alpha_2x^2$ ; and 3)  $Y = \gamma_0 + \gamma_1x + \gamma_2x^2 + \gamma_3x^3$ .
- (a) If you use all 100 data to fit these three model and use the same 100 data to calculate MSE, which model will have the smallest MSE? Justify your answer
- (b) Since we focus on the ability to predict unseen data, you decide use validation set to evaluate the performance of the three models. There are two way to split the data: 1) 80 testing points and 20 training points or 2) 20 testing points and 80 training points. Which one will you choose? Justify your choice
- (c) From the in-class lab, we know validation set is not a good choice to choose model. Explain the drawback of validation set and what procedure will you use to choose the best model?



8. (10 Points) Suppose we have a data set with five predictors,  $X_1 = \text{GPA}$ ,  $X_2 = \text{IQ}$ ,  $X_3 = \text{Level}$  (1 for College and 0 for High School),  $X_4 = \text{Interaction between GPA and IQ}$ , and  $X_5 = \text{Interaction between GPA and Level}$ . The response is starting salary after graduation (in thousands of dollars). Suppose we use least squares to fit the model, and get  $\beta_0 = 50$ ,  $\beta_1 = 20$ ,  $\beta_2 = 0.07$ ,  $\beta_3 = 35$ ,  $\beta_4 = 0.01$ ,  $\beta_5 = -10$ .

(a) Predict the salary of a college graduate with IQ of 110 and a GPA of 4.0.

- (b) True or false: Since the coefficient for the GPA/IQ interaction term is very small, there is very little evidence of an interaction effect. Justify your answer.

22A59D8B-5D7D-446B-A9C5-3F4192959A9F

midterm1-b9ace

#50 Page 9 of 10



**Scrap Paper**



3D0ED42A-3C4B-4BDA-8EDD-D53290662DAC

midterm1-b9ace

#50 Page 10 of 10



Michigan State University  
CMSE381 - Data Science

Feb 10, 2023  
Dr. Xie

## CMSE381 - Midterm #1

First name (please write as legibly as possible within the boxes)

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

Last name

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

NetID

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

*I will adhere to the Spartan Code of Honor in completing this assignment.*

Signed: \_\_\_\_\_

1. Do not open this test booklet until you are directed to do so.
2. You will have class time (2:40-4:00pm) to complete the exam.
3. This exam is closed book, but you can use the cheatsheet provided by the instructor.
4. Throughout the test, show your work so that your reasoning is clear. Otherwise no credit will be given. **BOX** your answers. Partial credit will be given where warranted.
5. Do not spend too much time on any one problem. Read them all through first and attack them in the order that allows you to make the most progress. Good luck :P



A6AE3B3E-788D-4231-A78C-9728C4279B25

midterm1-b9ace

#51 Page 2 of 10

1. (15 points)

(a) Logistic regression is used for regression.

TRUE      FALSE

(b) Any model will never have training error below the irreducible error.

TRUE      FALSE

(c) Increasing your model flexibility always results in a better model.

TRUE      FALSE

(d) A logistic regression model is set up so that the odds are linear.

TRUE      FALSE

(e) Circle all of the following that would represent a qualitative variable.

Age      Year      Dog\_breed      Country\_of\_origin

Student\_(True/False)      Weight      Speed      MPG

(f) What equation would you use to evaluate the result of a regression model?

(g) What equation would you use to evaluate the result of a classification model?

2. (15 Points) I'm building a model to predict amount of a given brand of dog food eaten by a collection of dogs. I have 100 dogs eat this dog food, and I collect information on their height, weight, breed, and whether they live with another dog in the house.
  - (a) List all input variables and specify whether they are quantitative or qualitative.
  - (b) Say our dog breeds sampled are Huskies, Terriers, and Spaniels. Write down your prediction model.
  - (c) We found that weight is a key predictor, and its relationship with the response is beyond linear. What extension can you come up with? Write down your updated model.





3B721953-7035-4F46-8FCC-F37753940B06

midterm1-b9ace

#51 Page 4 of 10

3. (15 points)

(a) What is bias-variance tradeoff? Explain the meaning of each term in the formula.

(b) Provide a sketch of typical (squared) bias, variance, training error, test error, and Bayes (or irreducible) error curves, on a single plot, as we go from less flexible statistical learning methods towards more flexible approaches. The  $x$ -axis should represent the amount of flexibility in the method, and the  $y$ -axis should represent the values for each curve. There should be five curves. Make sure to label each one.



(c) Explain why the (i) training error and (ii) testing error lines in your drawing have the shape displayed.



4. (10 points) The data for this example come from a study by Stamey et al. (1989). The data comes from a number of clinical measures in men who were about to receive a radical prostatectomy. The goal is to predict the log of PSA (**lpsa**) (a prostate-specific antigen measured in nanograms of PSA per milliliter of blood (ng/mL)) from a number of measurements. The variables are log cancer volume (**lcavol**), log prostate weight (**lweight**), age, log of the amount of benign prostatic hyperplasia (**lbph**), seminal vesicle invasion (**svi**), log of capsular penetration (**lcp**), Gleason score (**gleason**), and percent of Gleason scores 4 or 5 (**pgg45**).

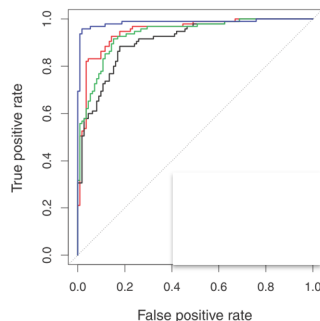
(a) Is this a case of regression or classification? Why?

(b) A linear model is fit to the data set, and the following table was returned.

Term	Coefficient	Std. Error	t-Score	p-value
Intercept	2.46	0.09	27.6	<0.00001
lcavol	0.68	0.13	5.37	<0.00001
lweight	0.26	0.1	2.75	0.00596
age	-0.14	0.1	-1.4	0.16153
lbph	0.21	0.1	2.06	0.039399
svi	0.31	0.12	2.47	0.013511
lcp	-0.29	0.15	-1.87	0.061484
gleason	-0.02	0.15	-0.15	0.880765
pgg45	0.27	0.15	1.74	0.081859

Are all the predictors useful? If yes, explain why. If not, explain what you would try to do next with the data.

5. (5 pts) We are trying to predict whether a patient will have a heart attack within one year. We have tried four different methods on a training dataset and have the following ROC curves. Which curve has the best performance on this training set? Justify your answer.





6. (15 Points) I get way too much email, so I decide to build a logistic regression model to predict whether a new incoming message is spam or not. I decide to just use a few variables:

$$\begin{aligned}X_1 &= \text{Number\_of\_sentences} \\X_2 &= \text{Number\_of\_references\_to\_a\_Nigerian\_Prince}\end{aligned}$$

and am training a logistic model to predict

$$Pr(Y = \text{spam} \mid X_1, X_2)$$

- (a) Write down the equation for the model you would train, using our standard notation with  $\beta_i$ 's.
- (b) If my trained model used  $\beta_0 = -13.1$ ,  $\beta_1 = 1.9$ , and  $\beta_2 = 6.1$ , what is the probability that a 5 sentence email with one reference to a Nigerian prince is spam? .
- (c) We know that if we miss an important email, it may cost a lot, How can you modify your model to avoid this?



7. (15 Points) Sparty generated 100 pair  $x$  and  $y$  via the following relationship:  $Y = 0.3 + 2x + x^2 + \epsilon$  but didn't tell you. You will try to find a good model to fit these data and more importantly to predict future response  $y$  when Sparty gives you another  $x$ . You will start with two models 1)  $Y = \beta_0 + \beta_1x$ ; 2)  $Y = \alpha_0 + \alpha_1x + \alpha_2x^2$ ; and 3)  $Y = \gamma_0 + \gamma_1x + \gamma_2x^2 + \gamma_3x^3$ .
- (a) If you use all 100 data to fit these three model and use the same 100 data to calculate MSE, which model will have the smallest MSE? Justify your answer
- (b) Since we focus on the ability to predict unseen data, you decide use validation set to evaluate the performance of the three models. There are two way to split the data: 1) 80 testing points and 20 training points or 2) 20 testing points and 80 training points. Which one will you choose? Justify your choice
- (c) From the in-class lab, we know validation set is not a good choice to choose model. Explain the drawback of validation set and what procedure will you use to choose the best model?



E4C8C134-A6EE-4E90-9200-C1244794D813

midterm1-b9ace

#51 Page 8 of 10

8. (10 Points) Suppose we have a data set with five predictors,  $X_1 = \text{GPA}$ ,  $X_2 = \text{IQ}$ ,  $X_3 = \text{Level}$  (1 for College and 0 for High School),  $X_4 = \text{Interaction between GPA and IQ}$ , and  $X_5 = \text{Interaction between GPA and Level}$ . The response is starting salary after graduation (in thousands of dollars). Suppose we use least squares to fit the model, and get  $\beta_0 = 50$ ,  $\beta_1 = 20$ ,  $\beta_2 = 0.07$ ,  $\beta_3 = 35$ ,  $\beta_4 = 0.01$ ,  $\beta_5 = -10$ .

(a) Predict the salary of a college graduate with IQ of 110 and a GPA of 4.0.

- (b) True or false: Since the coefficient for the GPA/IQ interaction term is very small, there is very little evidence of an interaction effect. Justify your answer.

29DFA151-CC52-4060-B59D-88D6EF356712

midterm1-b9ace

#51 Page 9 of 10



**Scrap Paper**



3E287F3C-05BC-4385-B42E-AB8951A3C755

midterm1-b9ace

#51 Page 10 of 10



Michigan State University  
CMSE381 - Data Science

Feb 10, 2023  
Dr. Xie

## CMSE381 - Midterm #1

First name (please write as legibly as possible within the boxes)

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

Last name

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

NetID

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

*I will adhere to the Spartan Code of Honor in completing this assignment.*

Signed: \_\_\_\_\_

1. Do not open this test booklet until you are directed to do so.
2. You will have class time (2:40-4:00pm) to complete the exam.
3. This exam is closed book, but you can use the cheatsheet provided by the instructor.
4. Throughout the test, show your work so that your reasoning is clear. Otherwise no credit will be given. **BOX** your answers. Partial credit will be given where warranted.
5. Do not spend too much time on any one problem. Read them all through first and attack them in the order that allows you to make the most progress. Good luck :P





29EA15CB-8361-416C-A4BF-940318795C3C

midterm1-b9ace

#52 Page 2 of 10

1. (15 points)

(a) Logistic regression is used for regression.

TRUE      FALSE

(b) Any model will never have training error below the irreducible error.

TRUE      FALSE

(c) Increasing your model flexibility always results in a better model.

TRUE      FALSE

(d) A logistic regression model is set up so that the odds are linear.

TRUE      FALSE

(e) Circle all of the following that would represent a qualitative variable.

Age      Year      Dog\_breed      Country\_of\_origin

Student\_(True/False)      Weight      Speed      MPG

(f) What equation would you use to evaluate the result of a regression model?

(g) What equation would you use to evaluate the result of a classification model?



2. (15 Points) I'm building a model to predict amount of a given brand of dog food eaten by a collection of dogs. I have 100 dogs eat this dog food, and I collect information on their height, weight, breed, and whether they live with another dog in the house.
- (a) List all input variables and specify whether they are quantitative or qualitative.
- (b) Say our dog breeds sampled are Huskies, Terriers, and Spaniels. Write down your prediction model.
- (c) We found that weight is a key predictor, and its relationship with the response is beyond linear. What extension can you come up with? Write down your updated model.



8CC0BD5B-7BFF-4672-A645-39E14A9D04BD

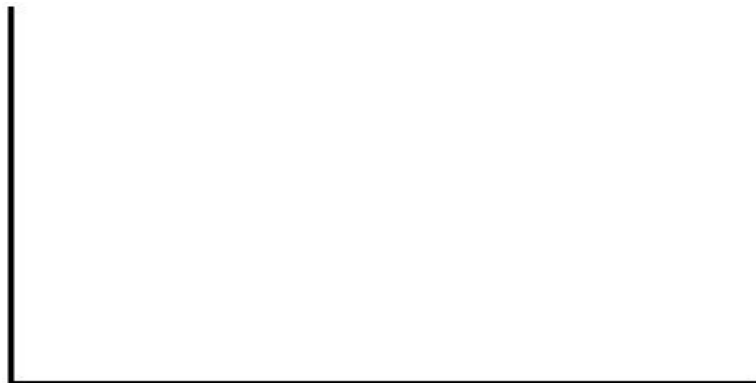
midterm1-b9ace

#52 Page 4 of 10

3. (15 points)

(a) What is bias-variance tradeoff? Explain the meaning of each term in the formula.

(b) Provide a sketch of typical (squared) bias, variance, training error, test error, and Bayes (or irreducible) error curves, on a single plot, as we go from less flexible statistical learning methods towards more flexible approaches. The  $x$ -axis should represent the amount of flexibility in the method, and the  $y$ -axis should represent the values for each curve. There should be five curves. Make sure to label each one.



(c) Explain why the (i) training error and (ii) testing error lines in your drawing have the shape displayed.



4. (10 points) The data for this example come from a study by Stamey et al. (1989). The data comes from a number of clinical measures in men who were about to receive a radical prostatectomy. The goal is to predict the log of PSA (**lpsa**) (a prostate-specific antigen measured in nanograms of PSA per milliliter of blood (ng/mL)) from a number of measurements. The variables are log cancer volume (**lcavol**), log prostate weight (**lweight**), age, log of the amount of benign prostatic hyperplasia (**lbph**), seminal vesicle invasion (**svi**), log of capsular penetration (**lcp**), Gleason score (**gleason**), and percent of Gleason scores 4 or 5 (**pgg45**).

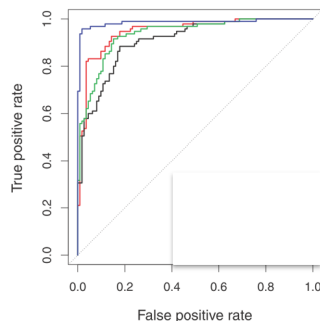
(a) Is this a case of regression or classification? Why?

(b) A linear model is fit to the data set, and the following table was returned.

Term	Coefficient	Std. Error	t-Score	p-value
Intercept	2.46	0.09	27.6	<0.00001
lcavol	0.68	0.13	5.37	<0.00001
lweight	0.26	0.1	2.75	0.00596
age	-0.14	0.1	-1.4	0.16153
lbph	0.21	0.1	2.06	0.039399
svi	0.31	0.12	2.47	0.013511
lcp	-0.29	0.15	-1.87	0.061484
gleason	-0.02	0.15	-0.15	0.880765
pgg45	0.27	0.15	1.74	0.081859

Are all the predictors useful? If yes, explain why. If not, explain what you would try to do next with the data.

5. (5 pts) We are trying to predict whether a patient will have a heart attack within one year. We have tried four different methods on a training dataset and have the following ROC curves. Which curve has the best performance on this training set? Justify your answer.





55434AD1-0B9D-4166-80ED-48C0E33B66FE

midterm1-b9ace

#52 Page 6 of 10

6. (15 Points) I get way too much email, so I decide to build a logistic regression model to predict whether a new incoming message is spam or not. I decide to just use a few variables:

$$\begin{aligned}X_1 &= \text{Number\_of\_sentences} \\X_2 &= \text{Number\_of\_references\_to\_a\_Nigerian\_Prince}\end{aligned}$$

and am training a logistic model to predict

$$Pr(Y = \text{spam} \mid X_1, X_2)$$

- (a) Write down the equation for the model you would train, using our standard notation with  $\beta_i$ 's.
- (b) If my trained model used  $\beta_0 = -13.1$ ,  $\beta_1 = 1.9$ , and  $\beta_2 = 6.1$ , what is the probability that a 5 sentence email with one reference to a Nigerian prince is spam? .
- (c) We know that if we miss an important email, it may cost a lot, How can you modify your model to avoid this?



7. (15 Points) Sparty generated 100 pair  $x$  and  $y$  via the following relationship:  $Y = 0.3 + 2x + x^2 + \epsilon$  but didn't tell you. You will try to find a good model to fit these data and more importantly to predict future response  $y$  when Sparty gives you another  $x$ . You will start with two models 1)  $Y = \beta_0 + \beta_1x$ ; 2)  $Y = \alpha_0 + \alpha_1x + \alpha_2x^2$ ; and 3)  $Y = \gamma_0 + \gamma_1x + \gamma_2x^2 + \gamma_3x^3$ .
- (a) If you use all 100 data to fit these three model and use the same 100 data to calculate MSE, which model will have the smallest MSE? Justify your answer
- (b) Since we focus on the ability to predict unseen data, you decide use validation set to evaluate the performance of the three models. There are two way to split the data: 1) 80 testing points and 20 training points or 2) 20 testing points and 80 training points. Which one will you choose? Justify your choice
- (c) From the in-class lab, we know validation set is not a good choice to choose model. Explain the drawback of validation set and what procedure will you use to choose the best model?



8. (10 Points) Suppose we have a data set with five predictors,  $X_1 = \text{GPA}$ ,  $X_2 = \text{IQ}$ ,  $X_3 = \text{Level}$  (1 for College and 0 for High School),  $X_4 = \text{Interaction between GPA and IQ}$ , and  $X_5 = \text{Interaction between GPA and Level}$ . The response is starting salary after graduation (in thousands of dollars). Suppose we use least squares to fit the model, and get  $\beta_0 = 50$ ,  $\beta_1 = 20$ ,  $\beta_2 = 0.07$ ,  $\beta_3 = 35$ ,  $\beta_4 = 0.01$ ,  $\beta_5 = -10$ .

(a) Predict the salary of a college graduate with IQ of 110 and a GPA of 4.0.

(b) True or false: Since the coefficient for the GPA/IQ interaction term is very small, there is very little evidence of an interaction effect. Justify your answer.

61EFB773-D5E7-4D80-974F-A87B1D7C0722

midterm1-b9ace

#52 Page 9 of 10



## Scrap Paper





D1E67E2E-077A-44B5-91E9-71A43F1F0085

midterm1-b9ace

#52 Page 10 of 10



Michigan State University  
CMSE381 - Data Science

Feb 10, 2023  
Dr. Xie

## CMSE381 - Midterm #1

First name (please write as legibly as possible within the boxes)

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

Last name

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

NetID

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

*I will adhere to the Spartan Code of Honor in completing this assignment.*

Signed: \_\_\_\_\_

1. Do not open this test booklet until you are directed to do so.
2. You will have class time (2:40-4:00pm) to complete the exam.
3. This exam is closed book, but you can use the cheatsheet provided by the instructor.
4. Throughout the test, show your work so that your reasoning is clear. Otherwise no credit will be given. **BOX** your answers. Partial credit will be given where warranted.
5. Do not spend too much time on any one problem. Read them all through first and attack them in the order that allows you to make the most progress. Good luck :P



E38EF60A-41B1-4D85-B4E8-2B4D639E1188

midterm1-b9ace

#53 Page 2 of 10

1. (15 points)

(a) Logistic regression is used for regression.

TRUE      FALSE

(b) Any model will never have training error below the irreducible error.

TRUE      FALSE

(c) Increasing your model flexibility always results in a better model.

TRUE      FALSE

(d) A logistic regression model is set up so that the odds are linear.

TRUE      FALSE

(e) Circle all of the following that would represent a qualitative variable.

Age      Year      Dog\_breed      Country\_of\_origin

Student\_(True/False)      Weight      Speed      MPG

(f) What equation would you use to evaluate the result of a regression model?

(g) What equation would you use to evaluate the result of a classification model?



2. (15 Points) I'm building a model to predict amount of a given brand of dog food eaten by a collection of dogs. I have 100 dogs eat this dog food, and I collect information on their height, weight, breed, and whether they live with another dog in the house.
- (a) List all input variables and specify whether they are quantitative or qualitative.
- (b) Say our dog breeds sampled are Huskies, Terriers, and Spaniels. Write down your prediction model.
- (c) We found that weight is a key predictor, and its relationship with the response is beyond linear. What extension can you come up with? Write down your updated model.



E932FAF1-ABB5-474C-893B-16A6D18E5133

midterm1-b9ace

#53 Page 4 of 10

3. (15 points)

(a) What is bias-variance tradeoff? Explain the meaning of each term in the formula.

(b) Provide a sketch of typical (squared) bias, variance, training error, test error, and Bayes (or irreducible) error curves, on a single plot, as we go from less flexible statistical learning methods towards more flexible approaches. The  $x$ -axis should represent the amount of flexibility in the method, and the  $y$ -axis should represent the values for each curve. There should be five curves. Make sure to label each one.



(c) Explain why the (i) training error and (ii) testing error lines in your drawing have the shape displayed.



4. (10 points) The data for this example come from a study by Stamey et al. (1989). The data comes from a number of clinical measures in men who were about to receive a radical prostatectomy. The goal is to predict the log of PSA (**lpsa**) (a prostate-specific antigen measured in nanograms of PSA per milliliter of blood (ng/mL)) from a number of measurements. The variables are log cancer volume (**lcavol**), log prostate weight (**lweight**), age, log of the amount of benign prostatic hyperplasia (**lbph**), seminal vesicle invasion (**svi**), log of capsular penetration (**lcp**), Gleason score (**gleason**), and percent of Gleason scores 4 or 5 (**pgg45**).

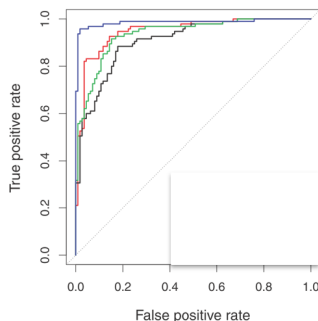
(a) Is this a case of regression or classification? Why?

(b) A linear model is fit to the data set, and the following table was returned.

Term	Coefficient	Std. Error	t-Score	p-value
Intercept	2.46	0.09	27.6	<0.00001
lcavol	0.68	0.13	5.37	<0.00001
lweight	0.26	0.1	2.75	0.00596
age	-0.14	0.1	-1.4	0.16153
lbph	0.21	0.1	2.06	0.039399
svi	0.31	0.12	2.47	0.013511
lcp	-0.29	0.15	-1.87	0.061484
gleason	-0.02	0.15	-0.15	0.880765
pgg45	0.27	0.15	1.74	0.081859

Are all the predictors useful? If yes, explain why. If not, explain what you would try to do next with the data.

5. (5 pts) We are trying to predict whether a patient will have a heart attack within one year. We have tried four different methods on a training dataset and have the following ROC curves. Which curve has the best performance on this training set? Justify your answer.





6. (15 Points) I get way too much email, so I decide to build a logistic regression model to predict whether a new incoming message is spam or not. I decide to just use a few variables:

$$\begin{aligned}X_1 &= \text{Number\_of\_sentences} \\X_2 &= \text{Number\_of\_references\_to\_a\_Nigerian\_Prince}\end{aligned}$$

and am training a logistic model to predict

$$Pr(Y = \text{spam} \mid X_1, X_2)$$

- (a) Write down the equation for the model you would train, using our standard notation with  $\beta_i$ 's.
- (b) If my trained model used  $\beta_0 = -13.1$ ,  $\beta_1 = 1.9$ , and  $\beta_2 = 6.1$ , what is the probability that a 5 sentence email with one reference to a Nigerian prince is spam? .
- (c) We know that if we miss an important email, it may cost a lot, How can you modify your model to avoid this?



7. (15 Points) Sparty generated 100 pair  $x$  and  $y$  via the following relationship:  $Y = 0.3 + 2x + x^2 + \epsilon$  but didn't tell you. You will try to find a good model to fit these data and more importantly to predict future response  $y$  when Sparty gives you another  $x$ . You will start with two models 1)  $Y = \beta_0 + \beta_1x$ ; 2)  $Y = \alpha_0 + \alpha_1x + \alpha_2x^2$ ; and 3)  $Y = \gamma_0 + \gamma_1x + \gamma_2x^2 + \gamma_3x^3$ .
- (a) If you use all 100 data to fit these three model and use the same 100 data to calculate MSE, which model will have the smallest MSE? Justify your answer
- (b) Since we focus on the ability to predict unseen data, you decide use validation set to evaluate the performance of the three models. There are two way to split the data: 1) 80 testing points and 20 training points or 2) 20 testing points and 80 training points. Which one will you choose? Justify your choice
- (c) From the in-class lab, we know validation set is not a good choice to choose model. Explain the drawback of validation set and what procedure will you use to choose the best model?





8. (10 Points) Suppose we have a data set with five predictors,  $X_1 = \text{GPA}$ ,  $X_2 = \text{IQ}$ ,  $X_3 = \text{Level}$  (1 for College and 0 for High School),  $X_4 = \text{Interaction between GPA and IQ}$ , and  $X_5 = \text{Interaction between GPA and Level}$ . The response is starting salary after graduation (in thousands of dollars). Suppose we use least squares to fit the model, and get  $\beta_0 = 50$ ,  $\beta_1 = 20$ ,  $\beta_2 = 0.07$ ,  $\beta_3 = 35$ ,  $\beta_4 = 0.01$ ,  $\beta_5 = -10$ .

(a) Predict the salary of a college graduate with IQ of 110 and a GPA of 4.0.

- (b) True or false: Since the coefficient for the GPA/IQ interaction term is very small, there is very little evidence of an interaction effect. Justify your answer.

15EBF5A2-9B8A-4BEF-9699-3606A53FF267

midterm1-b9ace

#53 Page 9 of 10



## Scrap Paper



E4EDEEC1-A8FB-4CF7-8F5D-666F1459E32F

midterm1-b9ace

#53 Page 10 of 10



Michigan State University  
CMSE381 - Data Science

Feb 10, 2023  
Dr. Xie

## CMSE381 - Midterm #1

First name (please write as legibly as possible within the boxes)

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

Last name

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

NetID

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

*I will adhere to the Spartan Code of Honor in completing this assignment.*

Signed: \_\_\_\_\_

1. Do not open this test booklet until you are directed to do so.
2. You will have class time (2:40-4:00pm) to complete the exam.
3. This exam is closed book, but you can use the cheatsheet provided by the instructor.
4. Throughout the test, show your work so that your reasoning is clear. Otherwise no credit will be given. **BOX** your answers. Partial credit will be given where warranted.
5. Do not spend too much time on any one problem. Read them all through first and attack them in the order that allows you to make the most progress. Good luck :P



21F55F90-5DA8-48A3-91F4-AE878078EFCA

midterm1-b9ace

#54 Page 2 of 10

1. (15 points)

(a) Logistic regression is used for regression.

TRUE      FALSE

(b) Any model will never have training error below the irreducible error.

TRUE      FALSE

(c) Increasing your model flexibility always results in a better model.

TRUE      FALSE

(d) A logistic regression model is set up so that the odds are linear.

TRUE      FALSE

(e) Circle all of the following that would represent a qualitative variable.

Age      Year      Dog\_breed      Country\_of\_origin

Student\_(True/False)      Weight      Speed      MPG

(f) What equation would you use to evaluate the result of a regression model?

(g) What equation would you use to evaluate the result of a classification model?



2. (15 Points) I'm building a model to predict amount of a given brand of dog food eaten by a collection of dogs. I have 100 dogs eat this dog food, and I collect information on their height, weight, breed, and whether they live with another dog in the house.
- (a) List all input variables and specify whether they are quantitative or qualitative.
  - (b) Say our dog breeds sampled are Huskies, Terriers, and Spaniels. Write down your prediction model.
  - (c) We found that weight is a key predictor, and its relationship with the response is beyond linear. What extension can you come up with? Write down your updated model.



3FC31853-36D3-4A6B-8277-6C30E9DD170A

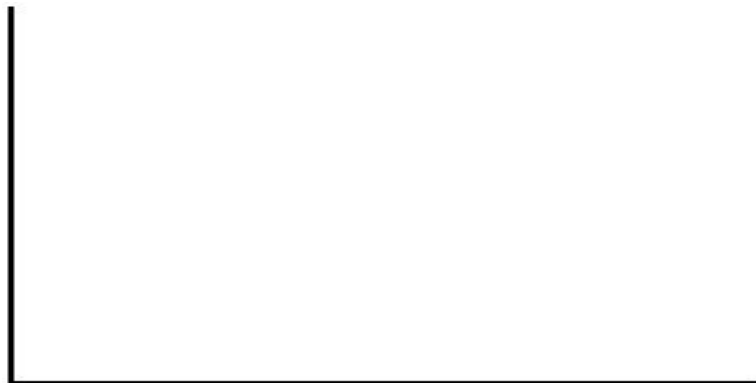
midterm1-b9ace

#54 Page 4 of 10

3. (15 points)

(a) What is bias-variance tradeoff? Explain the meaning of each term in the formula.

(b) Provide a sketch of typical (squared) bias, variance, training error, test error, and Bayes (or irreducible) error curves, on a single plot, as we go from less flexible statistical learning methods towards more flexible approaches. The  $x$ -axis should represent the amount of flexibility in the method, and the  $y$ -axis should represent the values for each curve. There should be five curves. Make sure to label each one.



(c) Explain why the (i) training error and (ii) testing error lines in your drawing have the shape displayed.



4. (10 points) The data for this example come from a study by Stamey et al. (1989). The data comes from a number of clinical measures in men who were about to receive a radical prostatectomy. The goal is to predict the log of PSA (**lpsa**) (a prostate-specific antigen measured in nanograms of PSA per milliliter of blood (ng/mL)) from a number of measurements. The variables are log cancer volume (**lcavol**), log prostate weight (**lweight**), age, log of the amount of benign prostatic hyperplasia (**lbph**), seminal vesicle invasion (**svi**), log of capsular penetration (**lcp**), Gleason score (**gleason**), and percent of Gleason scores 4 or 5 (**pgg45**).

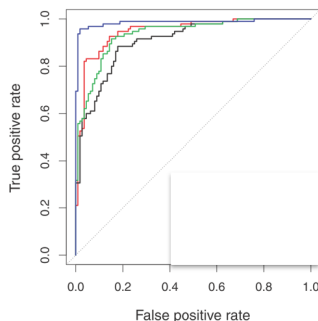
(a) Is this a case of regression or classification? Why?

(b) A linear model is fit to the data set, and the following table was returned.

Term	Coefficient	Std. Error	t-Score	p-value
Intercept	2.46	0.09	27.6	<0.00001
lcavol	0.68	0.13	5.37	<0.00001
lweight	0.26	0.1	2.75	0.00596
age	-0.14	0.1	-1.4	0.16153
lbph	0.21	0.1	2.06	0.039399
svi	0.31	0.12	2.47	0.013511
lcp	-0.29	0.15	-1.87	0.061484
gleason	-0.02	0.15	-0.15	0.880765
pgg45	0.27	0.15	1.74	0.081859

Are all the predictors useful? If yes, explain why. If not, explain what you would try to do next with the data.

5. (5 pts) We are trying to predict whether a patient will have a heart attack within one year. We have tried four different methods on a training dataset and have the following ROC curves. Which curve has the best performance on this training set? Justify your answer.







6. (15 Points) I get way too much email, so I decide to build a logistic regression model to predict whether a new incoming message is spam or not. I decide to just use a few variables:

$$\begin{aligned}X_1 &= \text{Number\_of\_sentences} \\X_2 &= \text{Number\_of\_references\_to\_a\_Nigerian\_Prince}\end{aligned}$$

and am training a logistic model to predict

$$Pr(Y = \text{spam} \mid X_1, X_2)$$

- (a) Write down the equation for the model you would train, using our standard notation with  $\beta_i$ 's.
- (b) If my trained model used  $\beta_0 = -13.1$ ,  $\beta_1 = 1.9$ , and  $\beta_2 = 6.1$ , what is the probability that a 5 sentence email with one reference to a Nigerian prince is spam? .
- (c) We know that if we miss an important email, it may cost a lot, How can you modify your model to avoid this?



7. (15 Points) Sparty generated 100 pair  $x$  and  $y$  via the following relationship:  $Y = 0.3 + 2x + x^2 + \epsilon$  but didn't tell you. You will try to find a good model to fit these data and more importantly to predict future response  $y$  when Sparty gives you another  $x$ . You will start with two models 1)  $Y = \beta_0 + \beta_1x$ ; 2)  $Y = \alpha_0 + \alpha_1x + \alpha_2x^2$ ; and 3)  $Y = \gamma_0 + \gamma_1x + \gamma_2x^2 + \gamma_3x^3$ .
- (a) If you use all 100 data to fit these three model and use the same 100 data to calculate MSE, which model will have the smallest MSE? Justify your answer
- (b) Since we focus on the ability to predict unseen data, you decide use validation set to evaluate the performance of the three models. There are two way to split the data: 1) 80 testing points and 20 training points or 2) 20 testing points and 80 training points. Which one will you choose? Justify your choice
- (c) From the in-class lab, we know validation set is not a good choice to choose model. Explain the drawback of validation set and what procedure will you use to choose the best model?



41125C24-12E9-4119-84DA-BE95DD4DB7FA

midterm1-b9ace

#54 Page 8 of 10

8. (10 Points) Suppose we have a data set with five predictors,  $X_1 = \text{GPA}$ ,  $X_2 = \text{IQ}$ ,  $X_3 = \text{Level}$  (1 for College and 0 for High School),  $X_4 = \text{Interaction between GPA and IQ}$ , and  $X_5 = \text{Interaction between GPA and Level}$ . The response is starting salary after graduation (in thousands of dollars). Suppose we use least squares to fit the model, and get  $\beta_0 = 50$ ,  $\beta_1 = 20$ ,  $\beta_2 = 0.07$ ,  $\beta_3 = 35$ ,  $\beta_4 = 0.01$ ,  $\beta_5 = -10$ .

(a) Predict the salary of a college graduate with IQ of 110 and a GPA of 4.0.

(b) True or false: Since the coefficient for the GPA/IQ interaction term is very small, there is very little evidence of an interaction effect. Justify your answer.

ECEABCEC-EFA4-462E-8396-9CC7A9E42D36

midterm1-b9ace

#54 Page 9 of 10



## Scrap Paper



52B46D15-97BA-4CAC-AD56-2941BF95B85B

midterm1-b9ace

#54 Page 10 of 10



Michigan State University  
CMSE381 - Data Science

Feb 10, 2023  
Dr. Xie

## CMSE381 - Midterm #1

First name (please write as legibly as possible within the boxes)

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

Last name

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

NetID

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

*I will adhere to the Spartan Code of Honor in completing this assignment.*

Signed: \_\_\_\_\_

1. Do not open this test booklet until you are directed to do so.
2. You will have class time (2:40-4:00pm) to complete the exam.
3. This exam is closed book, but you can use the cheatsheet provided by the instructor.
4. Throughout the test, show your work so that your reasoning is clear. Otherwise no credit will be given. **BOX** your answers. Partial credit will be given where warranted.
5. Do not spend too much time on any one problem. Read them all through first and attack them in the order that allows you to make the most progress. Good luck :P



56D1F996-6AC2-4B69-B377-5B72CF1C3800

midterm1-b9ace

#55 Page 2 of 10

1. (15 points)

(a) Logistic regression is used for regression.

TRUE      FALSE

(b) Any model will never have training error below the irreducible error.

TRUE      FALSE

(c) Increasing your model flexibility always results in a better model.

TRUE      FALSE

(d) A logistic regression model is set up so that the odds are linear.

TRUE      FALSE

(e) Circle all of the following that would represent a qualitative variable.

Age      Year      Dog\_breed      Country\_of\_origin

Student\_(True/False)      Weight      Speed      MPG

(f) What equation would you use to evaluate the result of a regression model?

(g) What equation would you use to evaluate the result of a classification model?



2. (15 Points) I'm building a model to predict amount of a given brand of dog food eaten by a collection of dogs. I have 100 dogs eat this dog food, and I collect information on their height, weight, breed, and whether they live with another dog in the house.
- (a) List all input variables and specify whether they are quantitative or qualitative.
- (b) Say our dog breeds sampled are Huskies, Terriers, and Spaniels. Write down your prediction model.
- (c) We found that weight is a key predictor, and its relationship with the response is beyond linear. What extension can you come up with? Write down your updated model.





6534BD45-A624-44E2-8E7F-CA636D7A19E7

midterm1-b9ace

#55 Page 4 of 10

3. (15 points)

(a) What is bias-variance tradeoff? Explain the meaning of each term in the formula.

(b) Provide a sketch of typical (squared) bias, variance, training error, test error, and Bayes (or irreducible) error curves, on a single plot, as we go from less flexible statistical learning methods towards more flexible approaches. The  $x$ -axis should represent the amount of flexibility in the method, and the  $y$ -axis should represent the values for each curve. There should be five curves. Make sure to label each one.



(c) Explain why the (i) training error and (ii) testing error lines in your drawing have the shape displayed.



4. (10 points) The data for this example come from a study by Stamey et al. (1989). The data comes from a number of clinical measures in men who were about to receive a radical prostatectomy. The goal is to predict the log of PSA (**lpsa**) (a prostate-specific antigen measured in nanograms of PSA per milliliter of blood (ng/mL)) from a number of measurements. The variables are log cancer volume (**lcavol**), log prostate weight (**lweight**), age, log of the amount of benign prostatic hyperplasia (**lbph**), seminal vesicle invasion (**svi**), log of capsular penetration (**lcp**), Gleason score (**gleason**), and percent of Gleason scores 4 or 5 (**pgg45**).

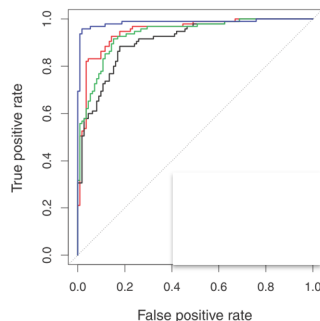
(a) Is this a case of regression or classification? Why?

(b) A linear model is fit to the data set, and the following table was returned.

Term	Coefficient	Std. Error	t-Score	p-value
Intercept	2.46	0.09	27.6	<0.00001
lcavol	0.68	0.13	5.37	<0.00001
lweight	0.26	0.1	2.75	0.00596
age	-0.14	0.1	-1.4	0.16153
lbph	0.21	0.1	2.06	0.039399
svi	0.31	0.12	2.47	0.013511
lcp	-0.29	0.15	-1.87	0.061484
gleason	-0.02	0.15	-0.15	0.880765
pgg45	0.27	0.15	1.74	0.081859

Are all the predictors useful? If yes, explain why. If not, explain what you would try to do next with the data.

5. (5 pts) We are trying to predict whether a patient will have a heart attack within one year. We have tried four different methods on a training dataset and have the following ROC curves. Which curve has the best performance on this training set? Justify your answer.





6A2ED242-7C12-4898-996F-156D99C53CD5

midterm1-b9ace

#55 Page 6 of 10

6. (15 Points) I get way too much email, so I decide to build a logistic regression model to predict whether a new incoming message is spam or not. I decide to just use a few variables:

$$\begin{aligned} X_1 &= \text{Number\_of\_sentences} \\ X_2 &= \text{Number\_of\_references\_to\_a\_Nigerian\_Prince} \end{aligned}$$

and am training a logistic model to predict

$$Pr(Y = \text{spam} \mid X_1, X_2)$$

- (a) Write down the equation for the model you would train, using our standard notation with  $\beta_i$ 's.
- (b) If my trained model used  $\beta_0 = -13.1$ ,  $\beta_1 = 1.9$ , and  $\beta_2 = 6.1$ , what is the probability that a 5 sentence email with one reference to a Nigerian prince is spam? .
- (c) We know that if we miss an important email, it may cost a lot, How can you modify your model to avoid this?



7. (15 Points) Sparty generated 100 pair  $x$  and  $y$  via the following relationship:  $Y = 0.3 + 2x + x^2 + \epsilon$  but didn't tell you. You will try to find a good model to fit these data and more importantly to predict future response  $y$  when Sparty gives you another  $x$ . You will start with two models 1)  $Y = \beta_0 + \beta_1x$ ; 2)  $Y = \alpha_0 + \alpha_1x + \alpha_2x^2$ ; and 3)  $Y = \gamma_0 + \gamma_1x + \gamma_2x^2 + \gamma_3x^3$ .
- (a) If you use all 100 data to fit these three model and use the same 100 data to calculate MSE, which model will have the smallest MSE? Justify your answer
- (b) Since we focus on the ability to predict unseen data, you decide use validation set to evaluate the performance of the three models. There are two way to split the data: 1) 80 testing points and 20 training points or 2) 20 testing points and 80 training points. Which one will you choose? Justify your choice
- (c) From the in-class lab, we know validation set is not a good choice to choose model. Explain the drawback of validation set and what procedure will you use to choose the best model?



8. (10 Points) Suppose we have a data set with five predictors,  $X_1 = \text{GPA}$ ,  $X_2 = \text{IQ}$ ,  $X_3 = \text{Level}$  (1 for College and 0 for High School),  $X_4 = \text{Interaction between GPA and IQ}$ , and  $X_5 = \text{Interaction between GPA and Level}$ . The response is starting salary after graduation (in thousands of dollars). Suppose we use least squares to fit the model, and get  $\beta_0 = 50$ ,  $\beta_1 = 20$ ,  $\beta_2 = 0.07$ ,  $\beta_3 = 35$ ,  $\beta_4 = 0.01$ ,  $\beta_5 = -10$ .

(a) Predict the salary of a college graduate with IQ of 110 and a GPA of 4.0.

- (b) True or false: Since the coefficient for the GPA/IQ interaction term is very small, there is very little evidence of an interaction effect. Justify your answer.

3528A76A-7AB9-4F20-BC74-56FC8D6CAC0F

midterm1-b9ace

#55 Page 9 of 10



**Scrap Paper**



771C9A29-0A5C-48A6-9444-C4AB79C5317C

midterm1-b9ace

#55 Page 10 of 10



Michigan State University  
CMSE381 - Data Science

Feb 10, 2023  
Dr. Xie

## CMSE381 - Midterm #1

First name (please write as legibly as possible within the boxes)

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

Last name

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

NetID

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

*I will adhere to the Spartan Code of Honor in completing this assignment.*

Signed: \_\_\_\_\_

1. Do not open this test booklet until you are directed to do so.
2. You will have class time (2:40-4:00pm) to complete the exam.
3. This exam is closed book, but you can use the cheatsheet provided by the instructor.
4. Throughout the test, show your work so that your reasoning is clear. Otherwise no credit will be given. **BOX** your answers. Partial credit will be given where warranted.
5. Do not spend too much time on any one problem. Read them all through first and attack them in the order that allows you to make the most progress. Good luck :P





2431449E-7DE0-4F26-ADF5-92419E235CE7

midterm1-b9ace

#56 Page 2 of 10

1. (15 points)

(a) Logistic regression is used for regression.

TRUE      FALSE

(b) Any model will never have training error below the irreducible error.

TRUE      FALSE

(c) Increasing your model flexibility always results in a better model.

TRUE      FALSE

(d) A logistic regression model is set up so that the odds are linear.

TRUE      FALSE

(e) Circle all of the following that would represent a qualitative variable.

Age      Year      Dog\_breed      Country\_of\_origin

Student\_(True/False)      Weight      Speed      MPG

(f) What equation would you use to evaluate the result of a regression model?

(g) What equation would you use to evaluate the result of a classification model?



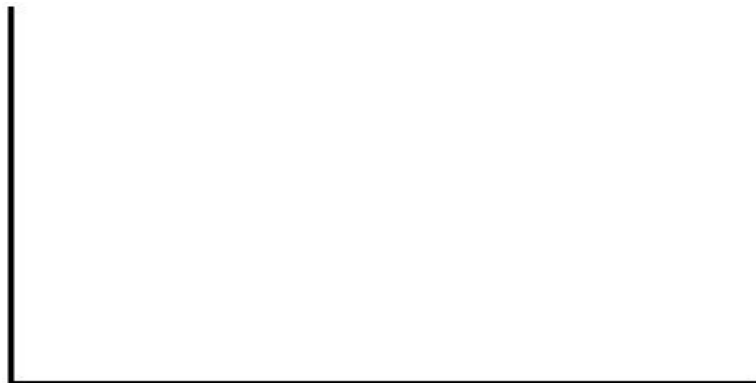
2. (15 Points) I'm building a model to predict amount of a given brand of dog food eaten by a collection of dogs. I have 100 dogs eat this dog food, and I collect information on their height, weight, breed, and whether they live with another dog in the house.
- (a) List all input variables and specify whether they are quantitative or qualitative.
- (b) Say our dog breeds sampled are Huskies, Terriers, and Spaniels. Write down your prediction model.
- (c) We found that weight is a key predictor, and its relationship with the response is beyond linear. What extension can you come up with? Write down your updated model.



3. (15 points)

(a) What is bias-variance tradeoff? Explain the meaning of each term in the formula.

(b) Provide a sketch of typical (squared) bias, variance, training error, test error, and Bayes (or irreducible) error curves, on a single plot, as we go from less flexible statistical learning methods towards more flexible approaches. The  $x$ -axis should represent the amount of flexibility in the method, and the  $y$ -axis should represent the values for each curve. There should be five curves. Make sure to label each one.



(c) Explain why the (i) training error and (ii) testing error lines in your drawing have the shape displayed.



4. (10 points) The data for this example come from a study by Stamey et al. (1989). The data comes from a number of clinical measures in men who were about to receive a radical prostatectomy. The goal is to predict the log of PSA (**lpsa**) (a prostate-specific antigen measured in nanograms of PSA per milliliter of blood (ng/mL)) from a number of measurements. The variables are log cancer volume (**lcavol**), log prostate weight (**lweight**), age, log of the amount of benign prostatic hyperplasia (**lbph**), seminal vesicle invasion (**svi**), log of capsular penetration (**lcp**), Gleason score (**gleason**), and percent of Gleason scores 4 or 5 (**pgg45**).

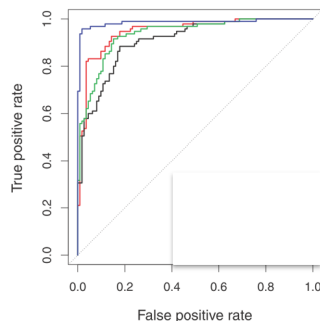
(a) Is this a case of regression or classification? Why?

(b) A linear model is fit to the data set, and the following table was returned.

Term	Coefficient	Std. Error	t-Score	p-value
Intercept	2.46	0.09	27.6	<0.00001
lcavol	0.68	0.13	5.37	<0.00001
lweight	0.26	0.1	2.75	0.00596
age	-0.14	0.1	-1.4	0.16153
lbph	0.21	0.1	2.06	0.039399
svi	0.31	0.12	2.47	0.013511
lcp	-0.29	0.15	-1.87	0.061484
gleason	-0.02	0.15	-0.15	0.880765
pgg45	0.27	0.15	1.74	0.081859

Are all the predictors useful? If yes, explain why. If not, explain what you would try to do next with the data.

5. (5 pts) We are trying to predict whether a patient will have a heart attack within one year. We have tried four different methods on a training dataset and have the following ROC curves. Which curve has the best performance on this training set? Justify your answer.





6. (15 Points) I get way too much email, so I decide to build a logistic regression model to predict whether a new incoming message is spam or not. I decide to just use a few variables:

$$\begin{aligned} X_1 &= \text{Number\_of\_sentences} \\ X_2 &= \text{Number\_of\_references\_to\_a\_Nigerian\_Prince} \end{aligned}$$

and am training a logistic model to predict

$$Pr(Y = \text{spam} \mid X_1, X_2)$$

- (a) Write down the equation for the model you would train, using our standard notation with  $\beta_i$ 's.
- (b) If my trained model used  $\beta_0 = -13.1$ ,  $\beta_1 = 1.9$ , and  $\beta_2 = 6.1$ , what is the probability that a 5 sentence email with one reference to a Nigerian prince is spam? .
- (c) We know that if we miss an important email, it may cost a lot, How can you modify your model to avoid this?



7. (15 Points) Sparty generated 100 pair  $x$  and  $y$  via the following relationship:  $Y = 0.3 + 2x + x^2 + \epsilon$  but didn't tell you. You will try to find a good model to fit these data and more importantly to predict future response  $y$  when Sparty gives you another  $x$ . You will start with two models 1)  $Y = \beta_0 + \beta_1x$ ; 2)  $Y = \alpha_0 + \alpha_1x + \alpha_2x^2$ ; and 3)  $Y = \gamma_0 + \gamma_1x + \gamma_2x^2 + \gamma_3x^3$ .
- (a) If you use all 100 data to fit these three model and use the same 100 data to calculate MSE, which model will have the smallest MSE? Justify your answer
- (b) Since we focus on the ability to predict unseen data, you decide use validation set to evaluate the performance of the three models. There are two way to split the data: 1) 80 testing points and 20 training points or 2) 20 testing points and 80 training points. Which one will you choose? Justify your choice
- (c) From the in-class lab, we know validation set is not a good choice to choose model. Explain the drawback of validation set and what procedure will you use to choose the best model?



8. (10 Points) Suppose we have a data set with five predictors,  $X_1 = \text{GPA}$ ,  $X_2 = \text{IQ}$ ,  $X_3 = \text{Level}$  (1 for College and 0 for High School),  $X_4 = \text{Interaction between GPA and IQ}$ , and  $X_5 = \text{Interaction between GPA and Level}$ . The response is starting salary after graduation (in thousands of dollars). Suppose we use least squares to fit the model, and get  $\beta_0 = 50$ ,  $\beta_1 = 20$ ,  $\beta_2 = 0.07$ ,  $\beta_3 = 35$ ,  $\beta_4 = 0.01$ ,  $\beta_5 = -10$ .

(a) Predict the salary of a college graduate with IQ of 110 and a GPA of 4.0.

- (b) True or false: Since the coefficient for the GPA/IQ interaction term is very small, there is very little evidence of an interaction effect. Justify your answer.

8ED1B538-00D1-4A93-AACD-6C86D2CBB9AE

midterm1-b9ace

#56 Page 9 of 10



**Scrap Paper**





13C38C6C-8784-447B-B4B0-6723ED4BC0DF

midterm1-b9ace

#56 Page 10 of 10



Michigan State University  
CMSE381 - Data Science

Feb 10, 2023  
Dr. Xie

## CMSE381 - Midterm #1

First name (please write as legibly as possible within the boxes)

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

Last name

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

NetID

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

*I will adhere to the Spartan Code of Honor in completing this assignment.*

Signed: \_\_\_\_\_

1. Do not open this test booklet until you are directed to do so.
2. You will have class time (2:40-4:00pm) to complete the exam.
3. This exam is closed book, but you can use the cheatsheet provided by the instructor.
4. Throughout the test, show your work so that your reasoning is clear. Otherwise no credit will be given. **BOX** your answers. Partial credit will be given where warranted.
5. Do not spend too much time on any one problem. Read them all through first and attack them in the order that allows you to make the most progress. Good luck :P



F1B82875-107C-4F31-A2D6-DA0825953435

midterm1-b9ace

#57 Page 2 of 10

1. (15 points)

(a) Logistic regression is used for regression.

TRUE      FALSE

(b) Any model will never have training error below the irreducible error.

TRUE      FALSE

(c) Increasing your model flexibility always results in a better model.

TRUE      FALSE

(d) A logistic regression model is set up so that the odds are linear.

TRUE      FALSE

(e) Circle all of the following that would represent a qualitative variable.

Age      Year      Dog\_breed      Country\_of\_origin

Student\_(True/False)      Weight      Speed      MPG

(f) What equation would you use to evaluate the result of a regression model?

(g) What equation would you use to evaluate the result of a classification model?



2. (15 Points) I'm building a model to predict amount of a given brand of dog food eaten by a collection of dogs. I have 100 dogs eat this dog food, and I collect information on their height, weight, breed, and whether they live with another dog in the house.
- (a) List all input variables and specify whether they are quantitative or qualitative.
  - (b) Say our dog breeds sampled are Huskies, Terriers, and Spaniels. Write down your prediction model.
  - (c) We found that weight is a key predictor, and its relationship with the response is beyond linear. What extension can you come up with? Write down your updated model.



3. (15 points)

(a) What is bias-variance tradeoff? Explain the meaning of each term in the formula.

(b) Provide a sketch of typical (squared) bias, variance, training error, test error, and Bayes (or irreducible) error curves, on a single plot, as we go from less flexible statistical learning methods towards more flexible approaches. The  $x$ -axis should represent the amount of flexibility in the method, and the  $y$ -axis should represent the values for each curve. There should be five curves. Make sure to label each one.



(c) Explain why the (i) training error and (ii) testing error lines in your drawing have the shape displayed.



4. (10 points) The data for this example come from a study by Stamey et al. (1989). The data comes from a number of clinical measures in men who were about to receive a radical prostatectomy. The goal is to predict the log of PSA (**lpsa**) (a prostate-specific antigen measured in nanograms of PSA per milliliter of blood (ng/mL)) from a number of measurements. The variables are log cancer volume (**lcavol**), log prostate weight (**lweight**), age, log of the amount of benign prostatic hyperplasia (**lbph**), seminal vesicle invasion (**svi**), log of capsular penetration (**lcp**), Gleason score (**gleason**), and percent of Gleason scores 4 or 5 (**pgg45**).

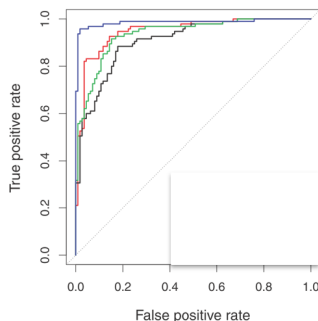
(a) Is this a case of regression or classification? Why?

(b) A linear model is fit to the data set, and the following table was returned.

Term	Coefficient	Std. Error	t-Score	p-value
Intercept	2.46	0.09	27.6	<0.00001
lcavol	0.68	0.13	5.37	<0.00001
lweight	0.26	0.1	2.75	0.00596
age	-0.14	0.1	-1.4	0.16153
lbph	0.21	0.1	2.06	0.039399
svi	0.31	0.12	2.47	0.013511
lcp	-0.29	0.15	-1.87	0.061484
gleason	-0.02	0.15	-0.15	0.880765
pgg45	0.27	0.15	1.74	0.081859

Are all the predictors useful? If yes, explain why. If not, explain what you would try to do next with the data.

5. (5 pts) We are trying to predict whether a patient will have a heart attack within one year. We have tried four different methods on a training dataset and have the following ROC curves. Which curve has the best performance on this training set? Justify your answer.





6247EC09-2D8E-4C41-9A38-4BF20E27AB9B

midterm1-b9ace

#57 Page 6 of 10

6. (15 Points) I get way too much email, so I decide to build a logistic regression model to predict whether a new incoming message is spam or not. I decide to just use a few variables:

$$\begin{aligned} X_1 &= \text{Number\_of\_sentences} \\ X_2 &= \text{Number\_of\_references\_to\_a\_Nigerian\_Prince} \end{aligned}$$

and am training a logistic model to predict

$$Pr(Y = \text{spam} \mid X_1, X_2)$$

- (a) Write down the equation for the model you would train, using our standard notation with  $\beta_i$ 's.
- (b) If my trained model used  $\beta_0 = -13.1$ ,  $\beta_1 = 1.9$ , and  $\beta_2 = 6.1$ , what is the probability that a 5 sentence email with one reference to a Nigerian prince is spam? .
- (c) We know that if we miss an important email, it may cost a lot, How can you modify your model to avoid this?



7. (15 Points) Sparty generated 100 pair  $x$  and  $y$  via the following relationship:  $Y = 0.3 + 2x + x^2 + \epsilon$  but didn't tell you. You will try to find a good model to fit these data and more importantly to predict future response  $y$  when Sparty gives you another  $x$ . You will start with two models 1)  $Y = \beta_0 + \beta_1x$ ; 2)  $Y = \alpha_0 + \alpha_1x + \alpha_2x^2$ ; and 3)  $Y = \gamma_0 + \gamma_1x + \gamma_2x^2 + \gamma_3x^3$ .
- (a) If you use all 100 data to fit these three model and use the same 100 data to calculate MSE, which model will have the smallest MSE? Justify your answer
- (b) Since we focus on the ability to predict unseen data, you decide use validation set to evaluate the performance of the three models. There are two way to split the data: 1) 80 testing points and 20 training points or 2) 20 testing points and 80 training points. Which one will you choose? Justify your choice
- (c) From the in-class lab, we know validation set is not a good choice to choose model. Explain the drawback of validation set and what procedure will you use to choose the best model?





8. (10 Points) Suppose we have a data set with five predictors,  $X_1 = \text{GPA}$ ,  $X_2 = \text{IQ}$ ,  $X_3 = \text{Level}$  (1 for College and 0 for High School),  $X_4 = \text{Interaction between GPA and IQ}$ , and  $X_5 = \text{Interaction between GPA and Level}$ . The response is starting salary after graduation (in thousands of dollars). Suppose we use least squares to fit the model, and get  $\beta_0 = 50$ ,  $\beta_1 = 20$ ,  $\beta_2 = 0.07$ ,  $\beta_3 = 35$ ,  $\beta_4 = 0.01$ ,  $\beta_5 = -10$ .

(a) Predict the salary of a college graduate with IQ of 110 and a GPA of 4.0.

(b) True or false: Since the coefficient for the GPA/IQ interaction term is very small, there is very little evidence of an interaction effect. Justify your answer.

DB5C8B69-D40E-4B86-8F74-2BA929DE862C

midterm1-b9ace

#57 Page 9 of 10



## Scrap Paper



8171A3EF-359A-4531-8BD5-BAFB053ACA26

midterm1-b9ace

#57 Page 10 of 10



Michigan State University  
CMSE381 - Data Science

Feb 10, 2023  
Dr. Xie

## CMSE381 - Midterm #1

First name (please write as legibly as possible within the boxes)

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

Last name

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

NetID

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

*I will adhere to the Spartan Code of Honor in completing this assignment.*

Signed: \_\_\_\_\_

1. Do not open this test booklet until you are directed to do so.
2. You will have class time (2:40-4:00pm) to complete the exam.
3. This exam is closed book, but you can use the cheatsheet provided by the instructor.
4. Throughout the test, show your work so that your reasoning is clear. Otherwise no credit will be given. **BOX** your answers. Partial credit will be given where warranted.
5. Do not spend too much time on any one problem. Read them all through first and attack them in the order that allows you to make the most progress. Good luck :P



4EE3EC2D-1AE0-4AE5-942B-FCE58D596163

midterm1-b9ace

#58 Page 2 of 10

1. (15 points)

(a) Logistic regression is used for regression.

TRUE      FALSE

(b) Any model will never have training error below the irreducible error.

TRUE      FALSE

(c) Increasing your model flexibility always results in a better model.

TRUE      FALSE

(d) A logistic regression model is set up so that the odds are linear.

TRUE      FALSE

(e) Circle all of the following that would represent a qualitative variable.

Age      Year      Dog\_breed      Country\_of\_origin

Student\_(True/False)      Weight      Speed      MPG

(f) What equation would you use to evaluate the result of a regression model?

(g) What equation would you use to evaluate the result of a classification model?



2. (15 Points) I'm building a model to predict amount of a given brand of dog food eaten by a collection of dogs. I have 100 dogs eat this dog food, and I collect information on their height, weight, breed, and whether they live with another dog in the house.
- (a) List all input variables and specify whether they are quantitative or qualitative.
  - (b) Say our dog breeds sampled are Huskies, Terriers, and Spaniels. Write down your prediction model.
  - (c) We found that weight is a key predictor, and its relationship with the response is beyond linear. What extension can you come up with? Write down your updated model.



2F65D92B-1178-4F22-BE69-4776EF27E582

midterm1-b9ace

#58 Page 4 of 10

3. (15 points)

(a) What is bias-variance tradeoff? Explain the meaning of each term in the formula.

(b) Provide a sketch of typical (squared) bias, variance, training error, test error, and Bayes (or irreducible) error curves, on a single plot, as we go from less flexible statistical learning methods towards more flexible approaches. The  $x$ -axis should represent the amount of flexibility in the method, and the  $y$ -axis should represent the values for each curve. There should be five curves. Make sure to label each one.



(c) Explain why the (i) training error and (ii) testing error lines in your drawing have the shape displayed.



4. (10 points) The data for this example come from a study by Stamey et al. (1989). The data comes from a number of clinical measures in men who were about to receive a radical prostatectomy. The goal is to predict the log of PSA (**lpsa**) (a prostate-specific antigen measured in nanograms of PSA per milliliter of blood (ng/mL)) from a number of measurements. The variables are log cancer volume (**lcavol**), log prostate weight (**lweight**), age, log of the amount of benign prostatic hyperplasia (**lbph**), seminal vesicle invasion (**svi**), log of capsular penetration (**lcp**), Gleason score (**gleason**), and percent of Gleason scores 4 or 5 (**pgg45**).

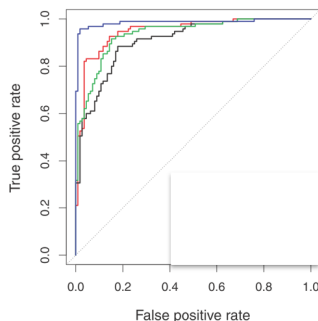
(a) Is this a case of regression or classification? Why?

(b) A linear model is fit to the data set, and the following table was returned.

Term	Coefficient	Std. Error	t-Score	p-value
Intercept	2.46	0.09	27.6	<0.00001
lcavol	0.68	0.13	5.37	<0.00001
lweight	0.26	0.1	2.75	0.00596
age	-0.14	0.1	-1.4	0.16153
lbph	0.21	0.1	2.06	0.039399
svi	0.31	0.12	2.47	0.013511
lcp	-0.29	0.15	-1.87	0.061484
gleason	-0.02	0.15	-0.15	0.880765
pgg45	0.27	0.15	1.74	0.081859

Are all the predictors useful? If yes, explain why. If not, explain what you would try to do next with the data.

5. (5 pts) We are trying to predict whether a patient will have a heart attack within one year. We have tried four different methods on a training dataset and have the following ROC curves. Which curve has the best performance on this training set? Justify your answer.







81AA178D-D8C0-4F69-96DA-FC7B0B3E99F4

midterm1-b9ace

#58 Page 6 of 10

6. (15 Points) I get way too much email, so I decide to build a logistic regression model to predict whether a new incoming message is spam or not. I decide to just use a few variables:

$$\begin{aligned} X_1 &= \text{Number\_of\_sentences} \\ X_2 &= \text{Number\_of\_references\_to\_a\_Nigerian\_Prince} \end{aligned}$$

and am training a logistic model to predict

$$Pr(Y = \text{spam} \mid X_1, X_2)$$

- (a) Write down the equation for the model you would train, using our standard notation with  $\beta_i$ 's.
- (b) If my trained model used  $\beta_0 = -13.1$ ,  $\beta_1 = 1.9$ , and  $\beta_2 = 6.1$ , what is the probability that a 5 sentence email with one reference to a Nigerian prince is spam? .
- (c) We know that if we miss an important email, it may cost a lot, How can you modify your model to avoid this?



7. (15 Points) Sparty generated 100 pair  $x$  and  $y$  via the following relationship:  $Y = 0.3 + 2x + x^2 + \epsilon$  but didn't tell you. You will try to find a good model to fit these data and more importantly to predict future response  $y$  when Sparty gives you another  $x$ . You will start with two models 1)  $Y = \beta_0 + \beta_1x$ ; 2)  $Y = \alpha_0 + \alpha_1x + \alpha_2x^2$ ; and 3)  $Y = \gamma_0 + \gamma_1x + \gamma_2x^2 + \gamma_3x^3$ .
- (a) If you use all 100 data to fit these three model and use the same 100 data to calculate MSE, which model will have the smallest MSE? Justify your answer
- (b) Since we focus on the ability to predict unseen data, you decide use validation set to evaluate the performance of the three models. There are two way to split the data: 1) 80 testing points and 20 training points or 2) 20 testing points and 80 training points. Which one will you choose? Justify your choice
- (c) From the in-class lab, we know validation set is not a good choice to choose model. Explain the drawback of validation set and what procedure will you use to choose the best model?



D5CAED81-43E1-4ED0-8EE2-5078C2FB10BC

midterm1-b9ace

#58 Page 8 of 10

8. (10 Points) Suppose we have a data set with five predictors,  $X_1 = \text{GPA}$ ,  $X_2 = \text{IQ}$ ,  $X_3 = \text{Level}$  (1 for College and 0 for High School),  $X_4 = \text{Interaction between GPA and IQ}$ , and  $X_5 = \text{Interaction between GPA and Level}$ . The response is starting salary after graduation (in thousands of dollars). Suppose we use least squares to fit the model, and get  $\beta_0 = 50$ ,  $\beta_1 = 20$ ,  $\beta_2 = 0.07$ ,  $\beta_3 = 35$ ,  $\beta_4 = 0.01$ ,  $\beta_5 = -10$ .

(a) Predict the salary of a college graduate with IQ of 110 and a GPA of 4.0.

- (b) True or false: Since the coefficient for the GPA/IQ interaction term is very small, there is very little evidence of an interaction effect. Justify your answer.

27C99B5F-55BA-4782-A072-EE60C9B056EB

midterm1-b9ace

#58 Page 9 of 10



**Scrap Paper**



72238AD8-96D6-49F9-8B3C-CB2F8E15FBFF

midterm1-b9ace

#58 Page 10 of 10



Michigan State University  
CMSE381 - Data Science

Feb 10, 2023  
Dr. Xie

## CMSE381 - Midterm #1

First name (please write as legibly as possible within the boxes)

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

Last name

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

NetID

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

*I will adhere to the Spartan Code of Honor in completing this assignment.*

Signed: \_\_\_\_\_

1. Do not open this test booklet until you are directed to do so.
2. You will have class time (2:40-4:00pm) to complete the exam.
3. This exam is closed book, but you can use the cheatsheet provided by the instructor.
4. Throughout the test, show your work so that your reasoning is clear. Otherwise no credit will be given. **BOX** your answers. Partial credit will be given where warranted.
5. Do not spend too much time on any one problem. Read them all through first and attack them in the order that allows you to make the most progress. Good luck :P



78A0DABA-EEA3-4D20-9174-2CED5FF551CD

midterm1-b9ace

#59 Page 2 of 10

1. (15 points)

(a) Logistic regression is used for regression.

TRUE      FALSE

(b) Any model will never have training error below the irreducible error.

TRUE      FALSE

(c) Increasing your model flexibility always results in a better model.

TRUE      FALSE

(d) A logistic regression model is set up so that the odds are linear.

TRUE      FALSE

(e) Circle all of the following that would represent a qualitative variable.

Age      Year      Dog\_breed      Country\_of\_origin

Student\_(True/False)      Weight      Speed      MPG

(f) What equation would you use to evaluate the result of a regression model?

(g) What equation would you use to evaluate the result of a classification model?



2. (15 Points) I'm building a model to predict amount of a given brand of dog food eaten by a collection of dogs. I have 100 dogs eat this dog food, and I collect information on their height, weight, breed, and whether they live with another dog in the house.
- (a) List all input variables and specify whether they are quantitative or qualitative.
  - (b) Say our dog breeds sampled are Huskies, Terriers, and Spaniels. Write down your prediction model.
  - (c) We found that weight is a key predictor, and its relationship with the response is beyond linear. What extension can you come up with? Write down your updated model.





119D7EDA-5075-42BA-92DF-432BC8B85687

midterm1-b9ace

#59 Page 4 of 10

3. (15 points)

(a) What is bias-variance tradeoff? Explain the meaning of each term in the formula.

(b) Provide a sketch of typical (squared) bias, variance, training error, test error, and Bayes (or irreducible) error curves, on a single plot, as we go from less flexible statistical learning methods towards more flexible approaches. The  $x$ -axis should represent the amount of flexibility in the method, and the  $y$ -axis should represent the values for each curve. There should be five curves. Make sure to label each one.



(c) Explain why the (i) training error and (ii) testing error lines in your drawing have the shape displayed.



4. (10 points) The data for this example come from a study by Stamey et al. (1989). The data comes from a number of clinical measures in men who were about to receive a radical prostatectomy. The goal is to predict the log of PSA (**lpsa**) (a prostate-specific antigen measured in nanograms of PSA per milliliter of blood (ng/mL)) from a number of measurements. The variables are log cancer volume (**lcavol**), log prostate weight (**lweight**), age, log of the amount of benign prostatic hyperplasia (**lbph**), seminal vesicle invasion (**svi**), log of capsular penetration (**lcp**), Gleason score (**gleason**), and percent of Gleason scores 4 or 5 (**pgg45**).

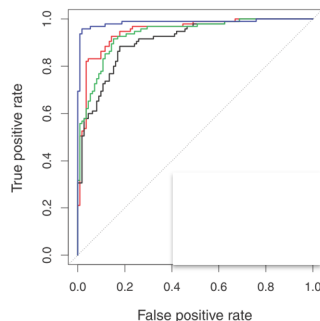
(a) Is this a case of regression or classification? Why?

(b) A linear model is fit to the data set, and the following table was returned.

Term	Coefficient	Std. Error	t-Score	p-value
Intercept	2.46	0.09	27.6	<0.00001
lcavol	0.68	0.13	5.37	<0.00001
lweight	0.26	0.1	2.75	0.00596
age	-0.14	0.1	-1.4	0.16153
lbph	0.21	0.1	2.06	0.039399
svi	0.31	0.12	2.47	0.013511
lcp	-0.29	0.15	-1.87	0.061484
gleason	-0.02	0.15	-0.15	0.880765
pgg45	0.27	0.15	1.74	0.081859

Are all the predictors useful? If yes, explain why. If not, explain what you would try to do next with the data.

5. (5 pts) We are trying to predict whether a patient will have a heart attack within one year. We have tried four different methods on a training dataset and have the following ROC curves. Which curve has the best performance on this training set? Justify your answer.





6. (15 Points) I get way too much email, so I decide to build a logistic regression model to predict whether a new incoming message is spam or not. I decide to just use a few variables:

$$\begin{aligned}X_1 &= \text{Number\_of\_sentences} \\X_2 &= \text{Number\_of\_references\_to\_a\_Nigerian\_Prince}\end{aligned}$$

and am training a logistic model to predict

$$Pr(Y = \text{spam} \mid X_1, X_2)$$

- (a) Write down the equation for the model you would train, using our standard notation with  $\beta_i$ 's.
- (b) If my trained model used  $\beta_0 = -13.1$ ,  $\beta_1 = 1.9$ , and  $\beta_2 = 6.1$ , what is the probability that a 5 sentence email with one reference to a Nigerian prince is spam? .
- (c) We know that if we miss an important email, it may cost a lot, How can you modify your model to avoid this?



7. (15 Points) Sparty generated 100 pair  $x$  and  $y$  via the following relationship:  $Y = 0.3 + 2x + x^2 + \epsilon$  but didn't tell you. You will try to find a good model to fit these data and more importantly to predict future response  $y$  when Sparty gives you another  $x$ . You will start with two models 1)  $Y = \beta_0 + \beta_1x$ ; 2)  $Y = \alpha_0 + \alpha_1x + \alpha_2x^2$ ; and 3)  $Y = \gamma_0 + \gamma_1x + \gamma_2x^2 + \gamma_3x^3$ .
- (a) If you use all 100 data to fit these three model and use the same 100 data to calculate MSE, which model will have the smallest MSE? Justify your answer
- (b) Since we focus on the ability to predict unseen data, you decide use validation set to evaluate the performance of the three models. There are two way to split the data: 1) 80 testing points and 20 training points or 2) 20 testing points and 80 training points. Which one will you choose? Justify your choice
- (c) From the in-class lab, we know validation set is not a good choice to choose model. Explain the drawback of validation set and what procedure will you use to choose the best model?



8. (10 Points) Suppose we have a data set with five predictors,  $X_1 = \text{GPA}$ ,  $X_2 = \text{IQ}$ ,  $X_3 = \text{Level}$  (1 for College and 0 for High School),  $X_4 = \text{Interaction between GPA and IQ}$ , and  $X_5 = \text{Interaction between GPA and Level}$ . The response is starting salary after graduation (in thousands of dollars). Suppose we use least squares to fit the model, and get  $\beta_0 = 50$ ,  $\beta_1 = 20$ ,  $\beta_2 = 0.07$ ,  $\beta_3 = 35$ ,  $\beta_4 = 0.01$ ,  $\beta_5 = -10$ .

(a) Predict the salary of a college graduate with IQ of 110 and a GPA of 4.0.

- (b) True or false: Since the coefficient for the GPA/IQ interaction term is very small, there is very little evidence of an interaction effect. Justify your answer.

38775884-5006-43D2-B30D-3619DDC02C6A

midterm1-b9ace

#59 Page 9 of 10



## Scrap Paper



F818A253-E249-4359-A809-7F702165C788

midterm1-b9ace

#59 Page 10 of 10