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)

Nicolai

Last name

Baird

NetID

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

Signed: Nicolai 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. $\boxed{\text{BOX}}$ 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 irreduceable 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?

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2$$

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.

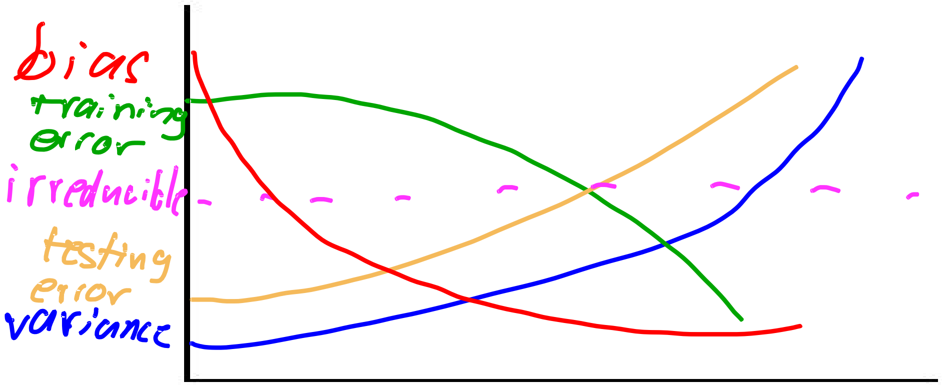
E(test mse at x_i) = Var(x_i) + bias$(x_i)^2$ + var(irreducible error)

var(x_i): variability of model at x_i

bias$(x_i)^2$: squared bias of model at x_i

var(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.

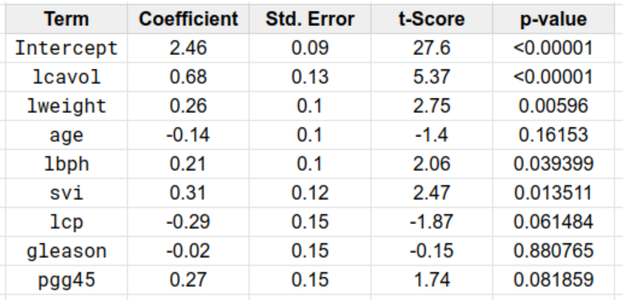
Training error decreases as the amount of flexibility increases because training error is mostly a result of overfitting, meaning that training error is a product of model bias. Testing error increases with flexibility because a less precise model cannot be trusted as much to be tested accurately.

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 (`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?

   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.

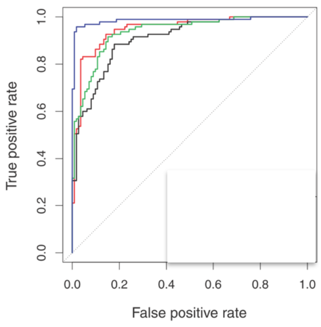


Term	Coefficient	Std. Error	t-Score	p-value
Intercept	2.46	0.09	27.6	<0.00001
lcavol	0.68	0.13	5.37	<0.00001
lweight	0.26	0.1	2.75	0.00596
age	-0.14	0.1	-1.4	0.16153
lbph	0.21	0.1	2.06	0.039399
svi	0.31	0.12	2.47	0.013511
lcp	-0.29	0.15	-1.87	0.061484
gleason	-0.02	0.15	-0.15	0.880765
pgg45	0.27	0.15	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 R^2 value for each combination of the predictors as well as model degree

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.



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.

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:

$$X_1 = \texttt{Number\_of\_sentences}$$
$$X_2 = \texttt{Number\_of\_references\_to\_a\_Nigerian\_Prince}$$

and am training a logistic model to predict

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

(a) Write down the equation for the model you would train, using our standard notation with β_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 de 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 job.

8. (10 Points) Suppose we have a data set with five predictors, X_1 = GPA, X_2 = IQ, X_3 = Level (1 for College and 0 for High School), X_4 = Interaction between GPA and IQ, and X_5 = 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 β_0 = 50, β_1 = 20, β_2 = 0.07, β_3 = 35, β_4 = 0.01, β_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 predictor 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 relavence with or without.

Scrap Paper

Michigan State University | **Feb 10, 2023**
CMSE381 - Data Science | **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 irreduceable 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.

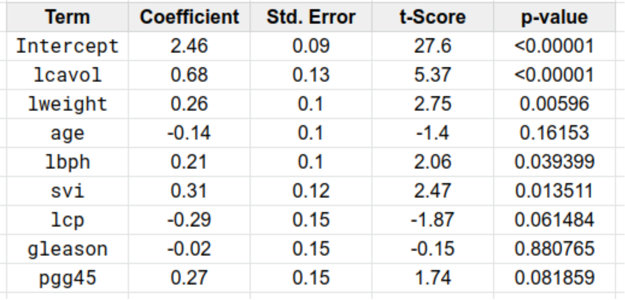
   (b) Provide a sketch of typical (squared) bias, variance, training error, test error, and Bayes (or irreducible) error curves, on a single plot, as we go from less flexible statistical learning methods towards more flexible approaches. The x-axis should represent the amount of flexibility in the method, 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?

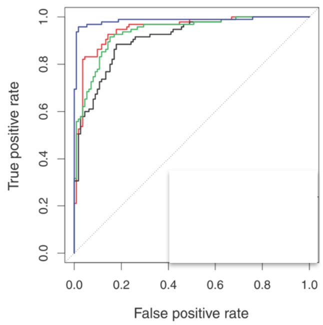
   (b) A linear model is fit to the data set, and the following table was returned.



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 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.



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:

$$X_1 = \texttt{Number\_of\_sentences}$$
$$X_2 = \texttt{Number\_of\_references\_to\_a\_Nigerian\_Prince}$$

and am training a logistic model to predict

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

(a) Write down the equation for the model you would train, using our standard notation with β_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 = GPA, X_2 = IQ, X_3 = Level (1 for College and 0 for High School), X_4 = Interaction between GPA and IQ, and X_5 = 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 $\beta_0 = 50$, $\beta_1 = 20$, $\beta_2 = 0.07$, $\beta_3 = 35$, $\beta_4 = 0.01$, β_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.

Scrap Paper

Michigan State University | **Feb 10, 2023**
CMSE381 - Data Science | **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 irreduceable 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.

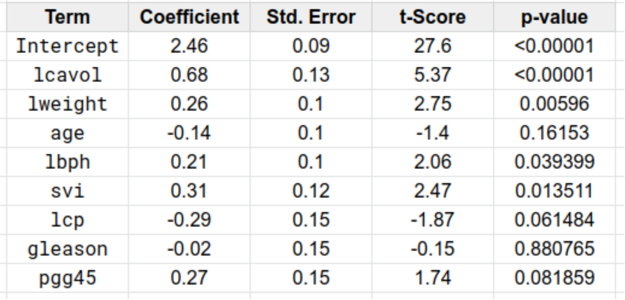
(b) Provide a sketch of typical (squared) bias, variance, training error, test error, and Bayes (or irreducible) error curves, on a single plot, as we go from less flexible statistical learning methods towards more flexible approaches. The x-axis should represent the amount of flexibility in the method, 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?

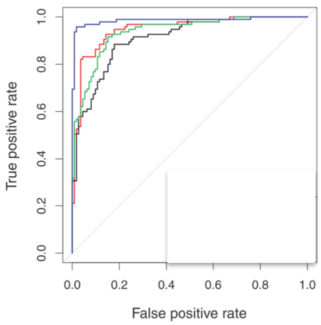
   (b) A linear model is fit to the data set, and the following table was returned.



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 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.



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:

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

and am training a logistic model to predict

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

(a) Write down the equation for the model you would train, using our standard notation with β_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 = GPA, X_2 = IQ, X_3 = Level (1 for College and 0 for High School), X_4 = Interaction between GPA and IQ, and X_5 = 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.

Scrap Paper

Michigan State University **Feb 10, 2023**
CMSE381 - Data Science **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 irreduceable 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.

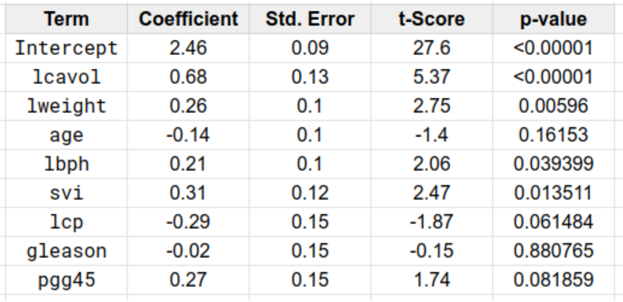
   (b) Provide a sketch of typical (squared) bias, variance, training error, test error, and Bayes (or irreducible) error curves, on a single plot, as we go from less flexible statistical learning methods towards more flexible approaches. The x-axis should represent the amount of flexibility in the method, 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?

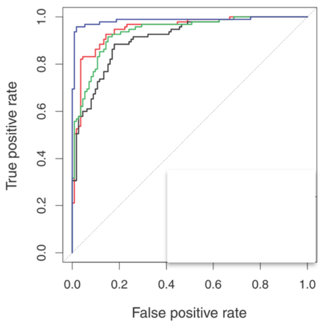
   (b) A linear model is fit to the data set, and the following table was returned.



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 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.



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:

$$X_1 = \texttt{Number\_of\_sentences}$$
$$X_2 = \texttt{Number\_of\_references\_to\_a\_Nigerian\_Prince}$$

and am training a logistic model to predict

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

(a) Write down the equation for the model you would train, using our standard notation with β_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 = GPA, X_2 = IQ, X_3 = Level (1 for College and 0 for High School), X_4 = Interaction between GPA and IQ, and X_5 = 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 $\beta_0 = 50$, $\beta_1 = 20$, $\beta_2 = 0.07$, $\beta_3 = 35$, $\beta_4 = 0.01$, β_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.

Scrap Paper

Michigan State University **Feb 10, 2023**
CMSE381 - Data Science **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. $\boxed{\text{BOX}}$ 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 irreduceable 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.

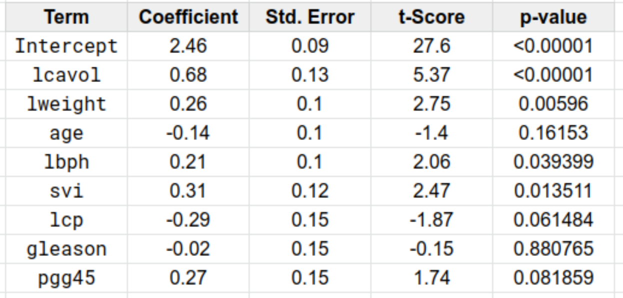
   (b) Provide a sketch of typical (squared) bias, variance, training error, test error, and Bayes (or irreducible) error curves, on a single plot, as we go from less flexible statistical learning methods towards more flexible approaches. The x-axis should represent the amount of flexibility in the method, 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?

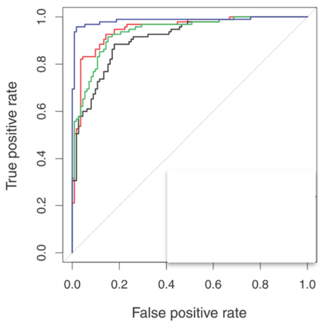
   (b) A linear model is fit to the data set, and the following table was returned.



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 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.



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:

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

and am training a logistic model to predict

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

(a) Write down the equation for the model you would train, using our standard notation with β_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 = GPA, X_2 = IQ, X_3 = Level (1 for College and 0 for High School), X_4 = Interaction between GPA and IQ, and X_5 = 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.

Scrap Paper

Michigan State University **Feb 10, 2023**
CMSE381 - Data Science **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. $\boxed{\text{BOX}}$ 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 irreduceable 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.

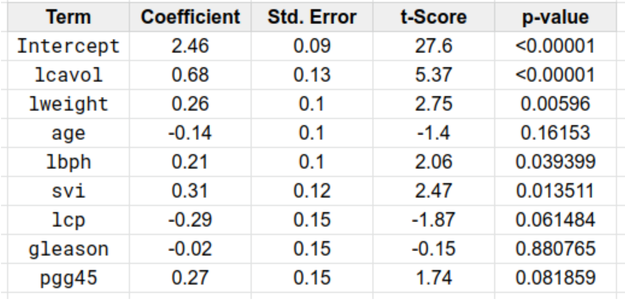
(b) Provide a sketch of typical (squared) bias, variance, training error, test error, and Bayes (or irreducible) error curves, on a single plot, as we go from less flexible statistical learning methods towards more flexible approaches. The x-axis should represent the amount of flexibility in the method, 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?

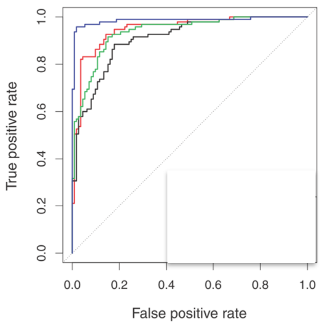
   (b) A linear model is fit to the data set, and the following table was returned.



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 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.



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:

$$X_1 = \texttt{Number\_of\_sentences}$$
$$X_2 = \texttt{Number\_of\_references\_to\_a\_Nigerian\_Prince}$$

and am training a logistic model to predict

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

(a) Write down the equation for the model you would train, using our standard notation with β_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 = GPA, X_2 = IQ, X_3 = Level (1 for College and 0 for High School), X_4 = Interaction between GPA and IQ, and X_5 = 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.

Scrap Paper

Michigan State University **Feb 10, 2023**
CMSE381 - Data Science **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 irreduceable 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.

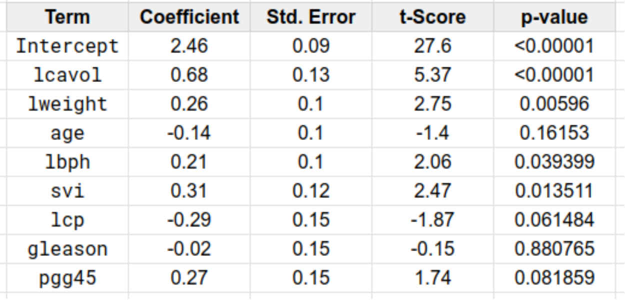
(b) Provide a sketch of typical (squared) bias, variance, training error, test error, and Bayes (or irreducible) error curves, on a single plot, as we go from less flexible statistical learning methods towards more flexible approaches. The x-axis should represent the amount of flexibility in the method, 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?

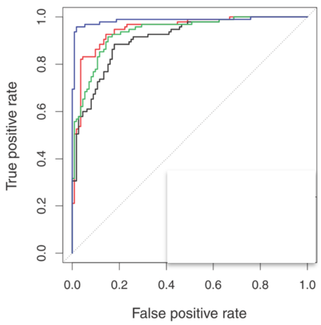
   (b) A linear model is fit to the data set, and the following table was returned.



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 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.



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:

$$X_1 = \texttt{Number\_of\_sentences}$$
$$X_2 = \texttt{Number\_of\_references\_to\_a\_Nigerian\_Prince}$$

and am training a logistic model to predict

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

(a) Write down the equation for the model you would train, using our standard notation with β_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 = GPA, X_2 = IQ, X_3 = Level (1 for College and 0 for High School), X_4 = Interaction between GPA and IQ, and X_5 = 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.

Scrap Paper

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. $\boxed{\text{BOX}}$ 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 irreduceable 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.

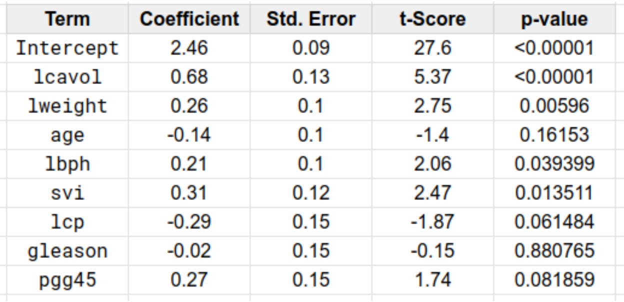
   (b) Provide a sketch of typical (squared) bias, variance, training error, test error, and Bayes (or irreducible) error curves, on a single plot, as we go from less flexible statistical learning methods towards more flexible approaches. The x-axis should represent the amount of flexibility in the method, 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?

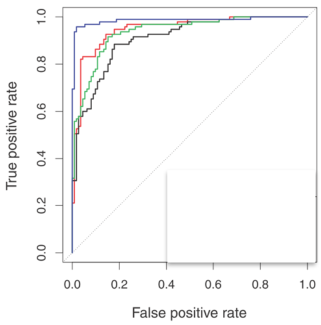
   (b) A linear model is fit to the data set, and the following table was returned.



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 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.



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:

$$X_1 = \texttt{Number\_of\_sentences}$$
$$X_2 = \texttt{Number\_of\_references\_to\_a\_Nigerian\_Prince}$$

and am training a logistic model to predict

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

(a) Write down the equation for the model you would train, using our standard notation with β_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 = GPA, X_2 = IQ, X_3 = Level (1 for College and 0 for High School), X_4 = Interaction between GPA and IQ, and X_5 = 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.

Scrap Paper

Michigan State University **Feb 10, 2023**
CMSE381 - Data Science **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. $\boxed{\text{BOX}}$ 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 irreduceable 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.

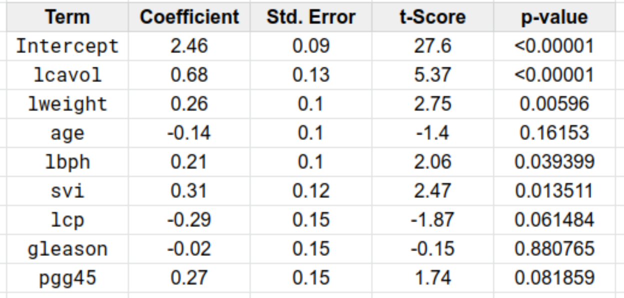
   (b) Provide a sketch of typical (squared) bias, variance, training error, test error, and Bayes (or irreducible) error curves, on a single plot, as we go from less flexible statistical learning methods towards more flexible approaches. The x-axis should represent the amount of flexibility in the method, 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?

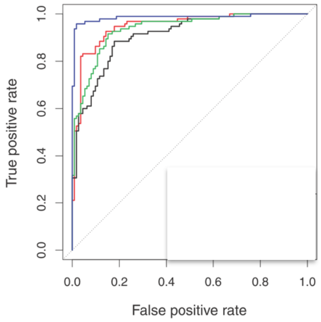
   (b) A linear model is fit to the data set, and the following table was returned.



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 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.



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:

$$X_1 = \texttt{Number\_of\_sentences}$$
$$X_2 = \texttt{Number\_of\_references\_to\_a\_Nigerian\_Prince}$$

and am training a logistic model to predict

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

(a) Write down the equation for the model you would train, using our standard notation with β_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 = GPA, X_2 = IQ, X_3 = Level (1 for College and 0 for High School), X_4 = Interaction between GPA and IQ, and X_5 = 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.

Scrap Paper

Michigan State University **Feb 10, 2023**
CMSE381 - Data Science **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. $\boxed{\text{BOX}}$ 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 irreduceable 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.

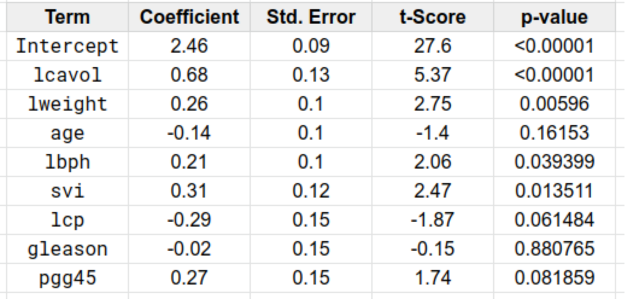
   (b) Provide a sketch of typical (squared) bias, variance, training error, test error, and Bayes (or irreducible) error curves, on a single plot, as we go from less flexible statistical learning methods towards more flexible approaches. The x-axis should represent the amount of flexibility in the method, 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?

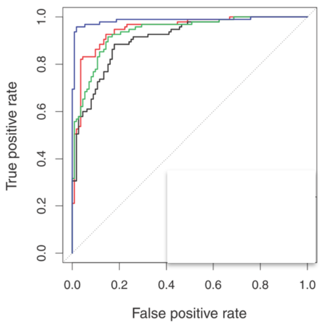
   (b) A linear model is fit to the data set, and the following table was returned.



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 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.



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:

$$X_1 = \texttt{Number\_of\_sentences}$$
$$X_2 = \texttt{Number\_of\_references\_to\_a\_Nigerian\_Prince}$$

and am training a logistic model to predict

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

(a) Write down the equation for the model you would train, using our standard notation with β_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 = GPA, X_2 = IQ, X_3 = Level (1 for College and 0 for High School), X_4 = Interaction between GPA and IQ, and X_5 = 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.

Scrap Paper

Michigan State University **Feb 10, 2023**
CMSE381 - Data Science **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. $\boxed{\text{BOX}}$ 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 irreduceable 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.

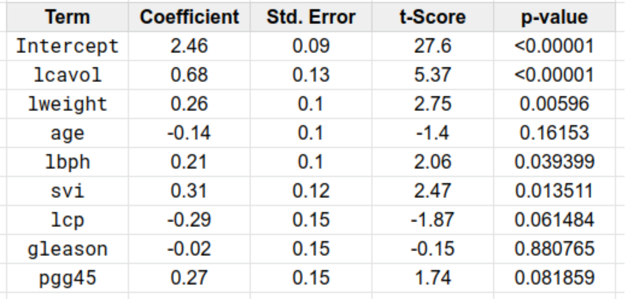
(b) Provide a sketch of typical (squared) bias, variance, training error, test error, and Bayes (or irreducible) error curves, on a single plot, as we go from less flexible statistical learning methods towards more flexible approaches. The x-axis should represent the amount of flexibility in the method, 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?

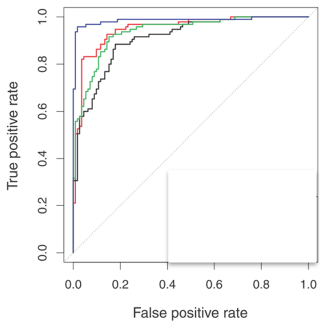
   (b) A linear model is fit to the data set, and the following table was returned.



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 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.



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:

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

and am training a logistic model to predict

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

(a) Write down the equation for the model you would train, using our standard notation with β_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 = GPA, X_2 = IQ, X_3 = Level (1 for College and 0 for High School), X_4 = Interaction between GPA and IQ, and X_5 = 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 $\beta_0 = 50$, $\beta_1 = 20$, $\beta_2 = 0.07$, $\beta_3 = 35$, $\beta_4 = 0.01$, β_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.

Scrap Paper

Michigan State University **Feb 10, 2023**
CMSE381 - Data Science **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. $\boxed{\text{BOX}}$ 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 irreduceable 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.

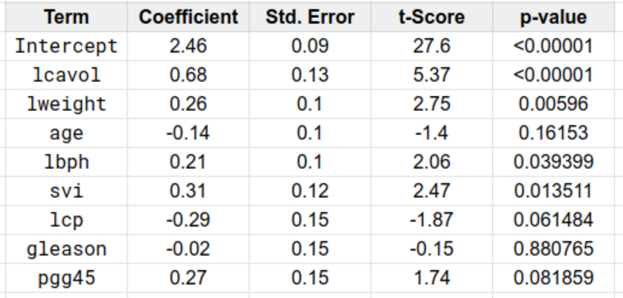
(b) Provide a sketch of typical (squared) bias, variance, training error, test error, and Bayes (or irreducible) error curves, on a single plot, as we go from less flexible statistical learning methods towards more flexible approaches. The x-axis should represent the amount of flexibility in the method, 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?

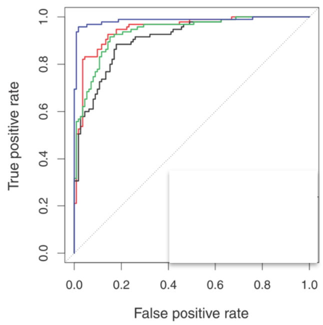
   (b) A linear model is fit to the data set, and the following table was returned.



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 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.



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:

$$X_1 = \texttt{Number\_of\_sentences}$$
$$X_2 = \texttt{Number\_of\_references\_to\_a\_Nigerian\_Prince}$$

and am training a logistic model to predict

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

(a) Write down the equation for the model you would train, using our standard notation with β_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 = GPA, X_2 = IQ, X_3 = Level (1 for College and 0 for High School), X_4 = Interaction between GPA and IQ, and X_5 = 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.

Scrap Paper

Michigan State University **Feb 10, 2023**
CMSE381 - Data Science **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. $\boxed{\text{BOX}}$ 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 irreduceable 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.

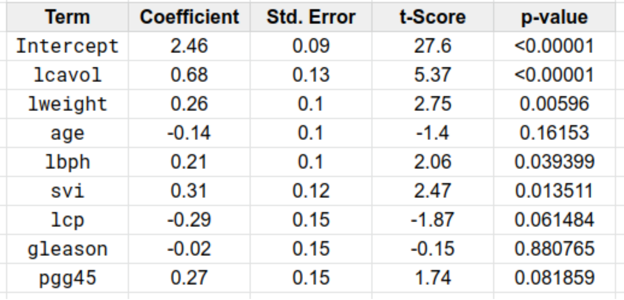
(b) Provide a sketch of typical (squared) bias, variance, training error, test error, and Bayes (or irreducible) error curves, on a single plot, as we go from less flexible statistical learning methods towards more flexible approaches. The x-axis should represent the amount of flexibility in the method, 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?

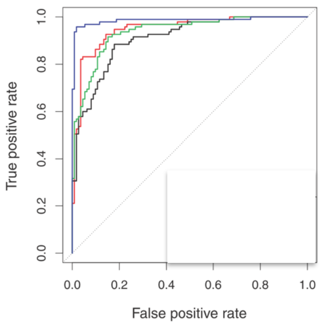
   (b) A linear model is fit to the data set, and the following table was returned.



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 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.



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:

$$X_1 = \texttt{Number\_of\_sentences}$$
$$X_2 = \texttt{Number\_of\_references\_to\_a\_Nigerian\_Prince}$$

and am training a logistic model to predict

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

(a) Write down the equation for the model you would train, using our standard notation with β_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 = GPA, X_2 = IQ, X_3 = Level (1 for College and 0 for High School), X_4 = Interaction between GPA and IQ, and X_5 = 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.

Scrap Paper

Michigan State University **Feb 10, 2023**
CMSE381 - Data Science **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. $\boxed{\text{BOX}}$ 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 irreduceable 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.

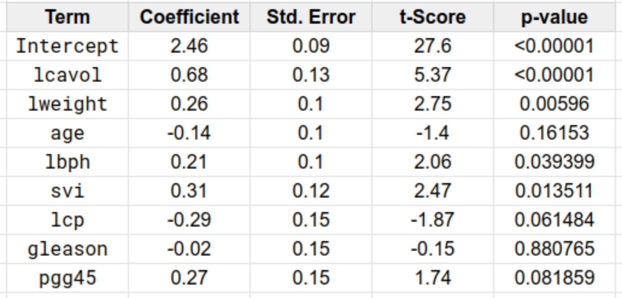
   (b) Provide a sketch of typical (squared) bias, variance, training error, test error, and Bayes (or irreducible) error curves, on a single plot, as we go from less flexible statistical learning methods towards more flexible approaches. The x-axis should represent the amount of flexibility in the method, 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?

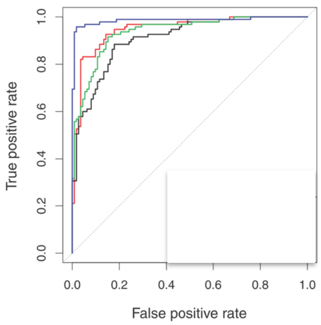
   (b) A linear model is fit to the data set, and the following table was returned.



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 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.



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:

$$X_1 = \texttt{Number\_of\_sentences}$$
$$X_2 = \texttt{Number\_of\_references\_to\_a\_Nigerian\_Prince}$$

and am training a logistic model to predict

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

(a) Write down the equation for the model you would train, using our standard notation with β_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 = GPA, X_2 = IQ, X_3 = Level (1 for College and 0 for High School), X_4 = Interaction between GPA and IQ, and X_5 = 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.

Scrap Paper

Michigan State University **Feb 10, 2023**
CMSE381 - Data Science **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. $\boxed{\text{BOX}}$ 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 irreduceable 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.

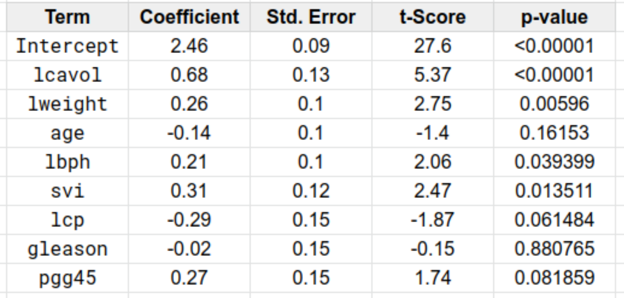
   (b) Provide a sketch of typical (squared) bias, variance, training error, test error, and Bayes (or irreducible) error curves, on a single plot, as we go from less flexible statistical learning methods towards more flexible approaches. The x-axis should represent the amount of flexibility in the method, 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?

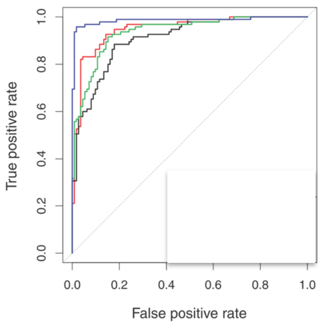
   (b) A linear model is fit to the data set, and the following table was returned.



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 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.



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:

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

and am training a logistic model to predict

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

(a) Write down the equation for the model you would train, using our standard notation with β_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 = GPA, X_2 = IQ, X_3 = Level (1 for College and 0 for High School), X_4 = Interaction between GPA and IQ, and X_5 = 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.

Scrap Paper

Michigan State University **Feb 10, 2023**
CMSE381 - Data Science **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. $\boxed{\text{BOX}}$ 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 irreduceable 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.

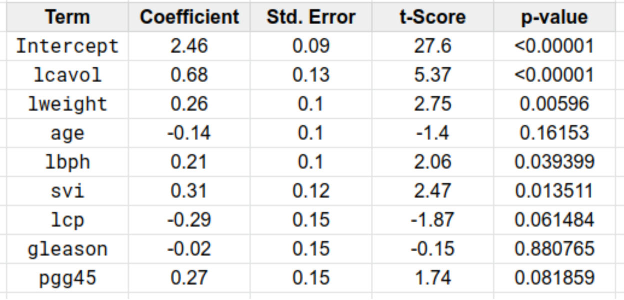
(b) Provide a sketch of typical (squared) bias, variance, training error, test error, and Bayes (or irreducible) error curves, on a single plot, as we go from less flexible statistical learning methods towards more flexible approaches. The x-axis should represent the amount of flexibility in the method, 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?

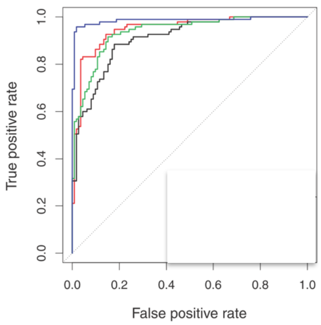
   (b) A linear model is fit to the data set, and the following table was returned.



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 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.



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:

$$X_1 = \texttt{Number\_of\_sentences}$$
$$X_2 = \texttt{Number\_of\_references\_to\_a\_Nigerian\_Prince}$$

and am training a logistic model to predict

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

(a) Write down the equation for the model you would train, using our standard notation with β_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 = GPA, X_2 = IQ, X_3 = Level (1 for College and 0 for High School), X_4 = Interaction between GPA and IQ, and X_5 = 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.

Scrap Paper

Michigan State University | **Feb 10, 2023**
CMSE381 - Data Science | **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. $\boxed{\text{BOX}}$ 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 irreduceable 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.

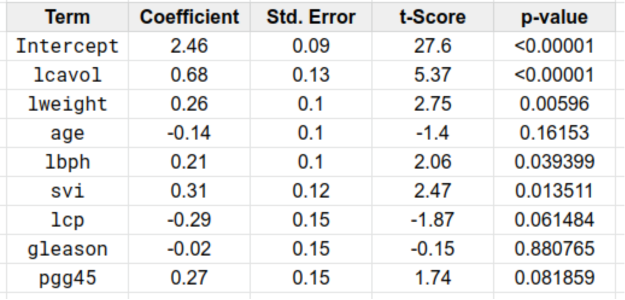
(b) Provide a sketch of typical (squared) bias, variance, training error, test error, and Bayes (or irreducible) error curves, on a single plot, as we go from less flexible statistical learning methods towards more flexible approaches. The x-axis should represent the amount of flexibility in the method, 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?

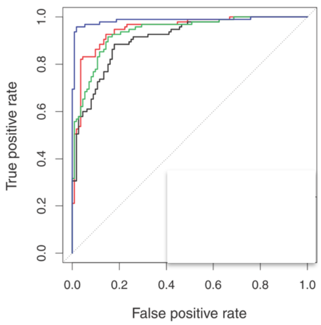
   (b) A linear model is fit to the data set, and the following table was returned.



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 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.



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:

$$X_1 = \texttt{Number\_of\_sentences}$$
$$X_2 = \texttt{Number\_of\_references\_to\_a\_Nigerian\_Prince}$$

and am training a logistic model to predict

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

(a) Write down the equation for the model you would train, using our standard notation with β_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 = GPA, X_2 = IQ, X_3 = Level (1 for College and 0 for High School), X_4 = Interaction between GPA and IQ, and X_5 = 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.

Scrap Paper

Michigan State University | **Feb 10, 2023**
CMSE381 - Data Science | **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 irreduceable 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.

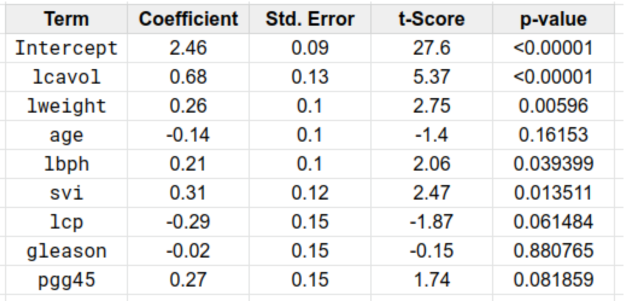
   (b) Provide a sketch of typical (squared) bias, variance, training error, test error, and Bayes (or irreducible) error curves, on a single plot, as we go from less flexible statistical learning methods towards more flexible approaches. The x-axis should represent the amount of flexibility in the method, 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?

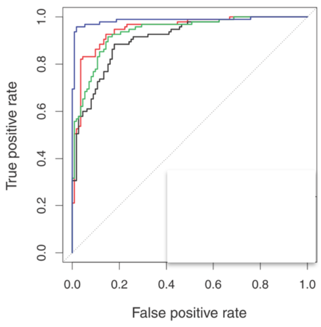
   (b) A linear model is fit to the data set, and the following table was returned.



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 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.



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:

$$X_1 = \texttt{Number\_of\_sentences}$$
$$X_2 = \texttt{Number\_of\_references\_to\_a\_Nigerian\_Prince}$$

and am training a logistic model to predict

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

(a) Write down the equation for the model you would train, using our standard notation with β_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 = GPA, X_2 = IQ, X_3 = Level (1 for College and 0 for High School), X_4 = Interaction between GPA and IQ, and X_5 = 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.

Scrap Paper

Michigan State University — **Feb 10, 2023**
CMSE381 - Data Science — **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. $\boxed{\text{BOX}}$ 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 irreduceable 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.

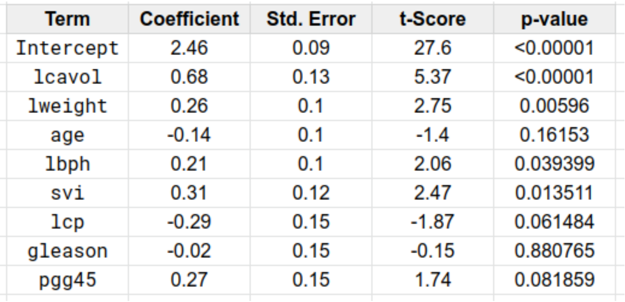
   (b) Provide a sketch of typical (squared) bias, variance, training error, test error, and Bayes (or irreducible) error curves, on a single plot, as we go from less flexible statistical learning methods towards more flexible approaches. The x-axis should represent the amount of flexibility in the method, 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?

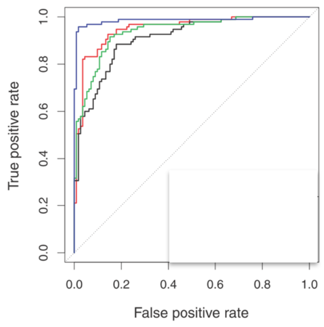
   (b) A linear model is fit to the data set, and the following table was returned.



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 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.



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:

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

and am training a logistic model to predict

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

(a) Write down the equation for the model you would train, using our standard notation with β_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 = GPA, X_2 = IQ, X_3 = Level (1 for College and 0 for High School), X_4 = Interaction between GPA and IQ, and X_5 = 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.

Scrap Paper

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. $\boxed{\text{BOX}}$ 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 irreduceable 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.

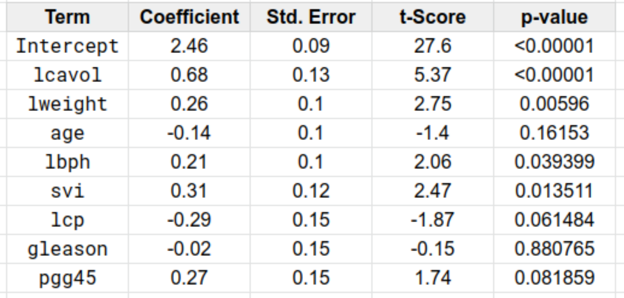
(b) Provide a sketch of typical (squared) bias, variance, training error, test error, and Bayes (or irreducible) error curves, on a single plot, as we go from less flexible statistical learning methods towards more flexible approaches. The x-axis should represent the amount of flexibility in the method, 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?

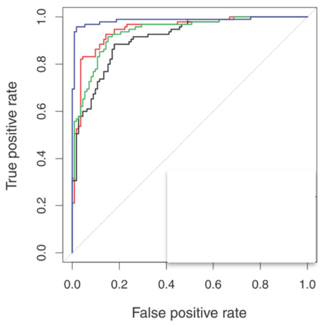
   (b) A linear model is fit to the data set, and the following table was returned.



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 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.



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:

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

and am training a logistic model to predict

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

(a) Write down the equation for the model you would train, using our standard notation with β_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 = GPA, X_2 = IQ, X_3 = Level (1 for College and 0 for High School), X_4 = Interaction between GPA and IQ, and X_5 = 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.

Scrap Paper

Michigan State University | **Feb 10, 2023**
CMSE381 - Data Science | **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 irreduceable 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.

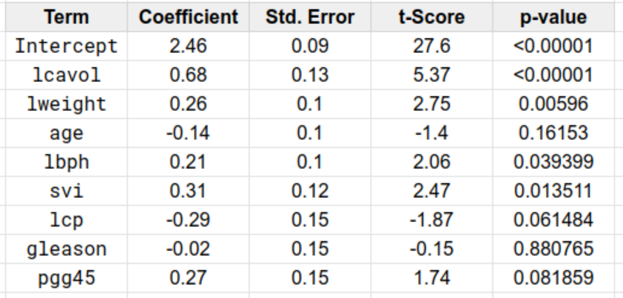
   (b) Provide a sketch of typical (squared) bias, variance, training error, test error, and Bayes (or irreducible) error curves, on a single plot, as we go from less flexible statistical learning methods towards more flexible approaches. The x-axis should represent the amount of flexibility in the method, 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?

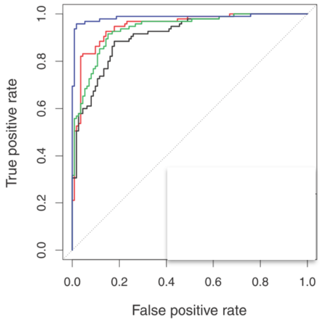
   (b) A linear model is fit to the data set, and the following table was returned.



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 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.



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:

$$X_1 = \texttt{Number\_of\_sentences}$$
$$X_2 = \texttt{Number\_of\_references\_to\_a\_Nigerian\_Prince}$$

and am training a logistic model to predict

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

(a) Write down the equation for the model you would train, using our standard notation with β_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 = GPA, X_2 = IQ, X_3 = Level (1 for College and 0 for High School), X_4 = Interaction between GPA and IQ, and X_5 = 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.

Scrap Paper

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. $\boxed{\text{BOX}}$ 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 irreduceable 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.

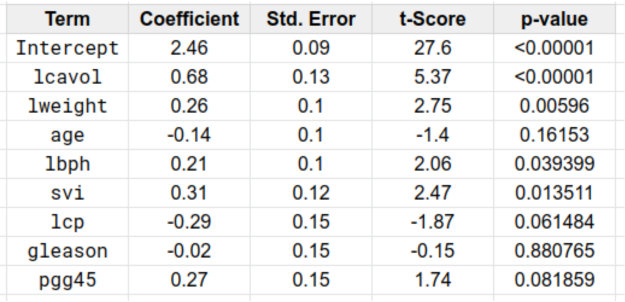
   (b) Provide a sketch of typical (squared) bias, variance, training error, test error, and Bayes (or irreducible) error curves, on a single plot, as we go from less flexible statistical learning methods towards more flexible approaches. The x-axis should represent the amount of flexibility in the method, 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?

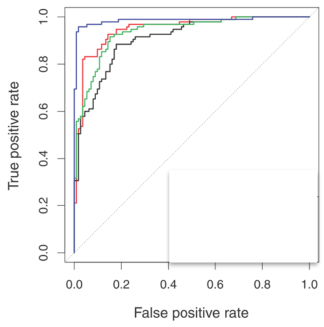
   (b) A linear model is fit to the data set, and the following table was returned.



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 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.



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:

$$X_1 = \texttt{Number\_of\_sentences}$$
$$X_2 = \texttt{Number\_of\_references\_to\_a\_Nigerian\_Prince}$$

and am training a logistic model to predict

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

(a) Write down the equation for the model you would train, using our standard notation with β_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 = GPA, X_2 = IQ, X_3 = Level (1 for College and 0 for High School), X_4 = Interaction between GPA and IQ, and X_5 = 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.

Scrap Paper

Michigan State University **Feb 10, 2023**
CMSE381 - Data Science **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. $\boxed{\text{BOX}}$ 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 irreduceable 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.

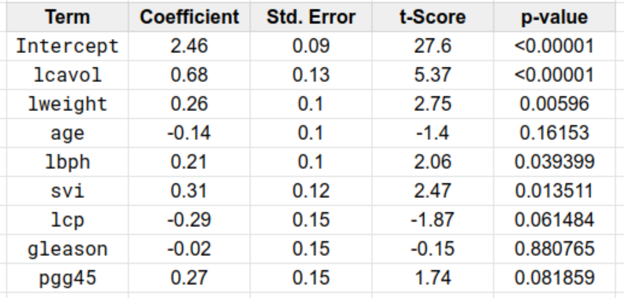
   (b) Provide a sketch of typical (squared) bias, variance, training error, test error, and Bayes (or irreducible) error curves, on a single plot, as we go from less flexible statistical learning methods towards more flexible approaches. The x-axis should represent the amount of flexibility in the method, 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?

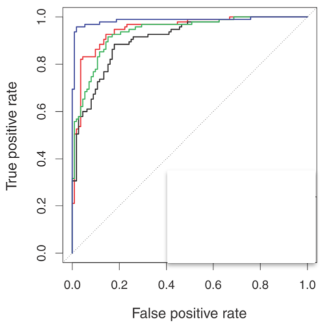
   (b) A linear model is fit to the data set, and the following table was returned.



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 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.



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:

$$X_1 = \texttt{Number\_of\_sentences}$$
$$X_2 = \texttt{Number\_of\_references\_to\_a\_Nigerian\_Prince}$$

and am training a logistic model to predict

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

(a) Write down the equation for the model you would train, using our standard notation with β_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 = GPA, X_2 = IQ, X_3 = Level (1 for College and 0 for High School), X_4 = Interaction between GPA and IQ, and X_5 = 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.

Scrap Paper

Michigan State University **Feb 10, 2023**
CMSE381 - Data Science **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. $\boxed{\text{BOX}}$ 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 irreduceable 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.

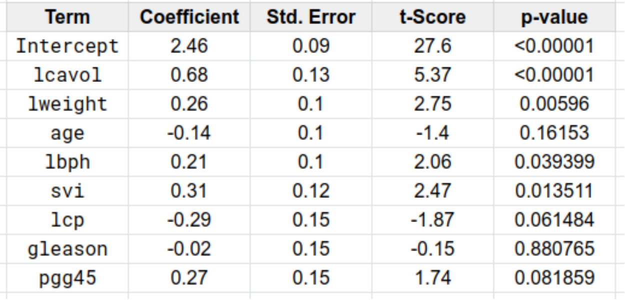
   (b) Provide a sketch of typical (squared) bias, variance, training error, test error, and Bayes (or irreducible) error curves, on a single plot, as we go from less flexible statistical learning methods towards more flexible approaches. The x-axis should represent the amount of flexibility in the method, 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?

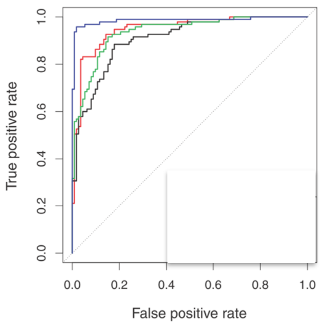
   (b) A linear model is fit to the data set, and the following table was returned.



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 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.



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:

$$X_1 = \texttt{Number\_of\_sentences}$$
$$X_2 = \texttt{Number\_of\_references\_to\_a\_Nigerian\_Prince}$$

and am training a logistic model to predict

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

(a) Write down the equation for the model you would train, using our standard notation with β_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 = GPA, X_2 = IQ, X_3 = Level (1 for College and 0 for High School), X_4 = Interaction between GPA and IQ, and X_5 = 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.

Scrap Paper

Michigan State University **Feb 10, 2023**
CMSE381 - Data Science **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. $\boxed{\text{BOX}}$ 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 irreduceable 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.

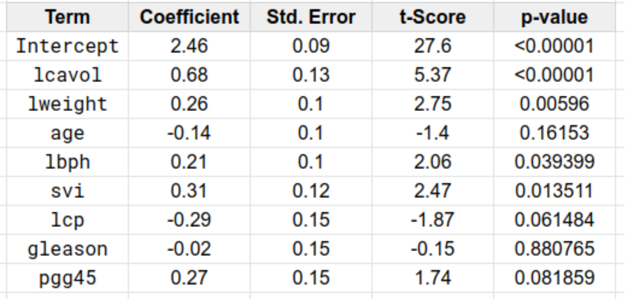
(b) Provide a sketch of typical (squared) bias, variance, training error, test error, and Bayes (or irreducible) error curves, on a single plot, as we go from less flexible statistical learning methods towards more flexible approaches. The x-axis should represent the amount of flexibility in the method, 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?

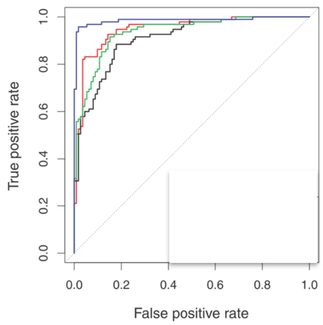
   (b) A linear model is fit to the data set, and the following table was returned.



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 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.



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:

$$X_1 = \texttt{Number\_of\_sentences}$$
$$X_2 = \texttt{Number\_of\_references\_to\_a\_Nigerian\_Prince}$$

and am training a logistic model to predict

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

(a) Write down the equation for the model you would train, using our standard notation with β_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 = GPA, X_2 = IQ, X_3 = Level (1 for College and 0 for High School), X_4 = Interaction between GPA and IQ, and X_5 = 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.

Scrap Paper

Michigan State University **Feb 10, 2023**
CMSE381 - Data Science **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. $\boxed{\text{BOX}}$ 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 irreduceable 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.

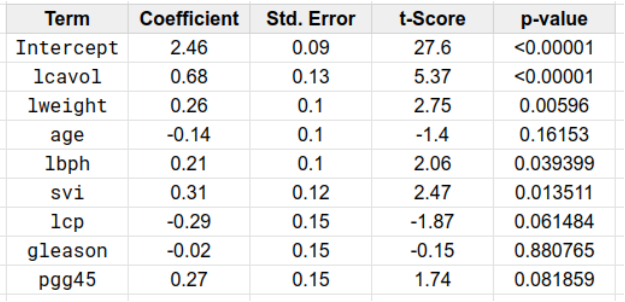
   (b) Provide a sketch of typical (squared) bias, variance, training error, test error, and Bayes (or irreducible) error curves, on a single plot, as we go from less flexible statistical learning methods towards more flexible approaches. The x-axis should represent the amount of flexibility in the method, 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?

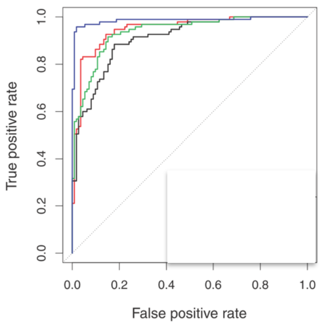
   (b) A linear model is fit to the data set, and the following table was returned.



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 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.



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:

$$X_1 = \texttt{Number\_of\_sentences}$$
$$X_2 = \texttt{Number\_of\_references\_to\_a\_Nigerian\_Prince}$$

and am training a logistic model to predict

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

(a) Write down the equation for the model you would train, using our standard notation with β_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 = GPA, X_2 = IQ, X_3 = Level (1 for College and 0 for High School), X_4 = Interaction between GPA and IQ, and X_5 = 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.

Scrap Paper

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. $\boxed{\text{BOX}}$ 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 irreduceable 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.

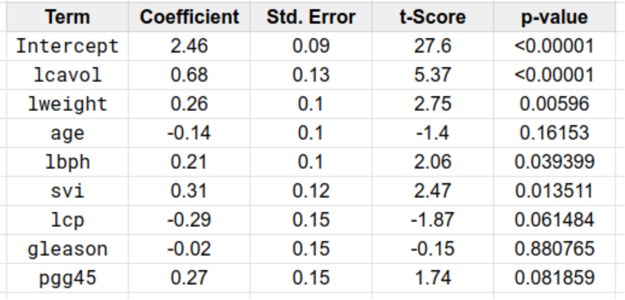
   (b) Provide a sketch of typical (squared) bias, variance, training error, test error, and Bayes (or irreducible) error curves, on a single plot, as we go from less flexible statistical learning methods towards more flexible approaches. The x-axis should represent the amount of flexibility in the method, 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?

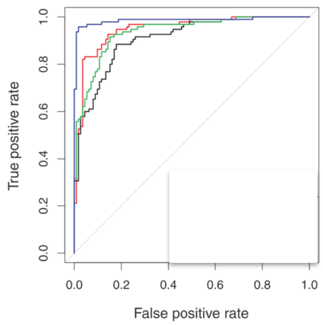
   (b) A linear model is fit to the data set, and the following table was returned.



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 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.



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:

$$X_1 = \texttt{Number\_of\_sentences}$$
$$X_2 = \texttt{Number\_of\_references\_to\_a\_Nigerian\_Prince}$$

and am training a logistic model to predict

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

(a) Write down the equation for the model you would train, using our standard notation with β_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 = GPA, X_2 = IQ, X_3 = Level (1 for College and 0 for High School), X_4 = Interaction between GPA and IQ, and X_5 = 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 $\beta_0 = 50$, $\beta_1 = 20$, $\beta_2 = 0.07$, $\beta_3 = 35$, $\beta_4 = 0.01$, β_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.

Scrap Paper

Michigan State University **Feb 10, 2023**
CMSE381 - Data Science **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 irreduceable 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.

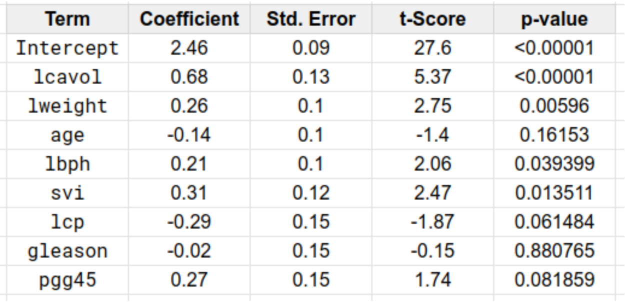
   (b) Provide a sketch of typical (squared) bias, variance, training error, test error, and Bayes (or irreducible) error curves, on a single plot, as we go from less flexible statistical learning methods towards more flexible approaches. The x-axis should represent the amount of flexibility in the method, 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?

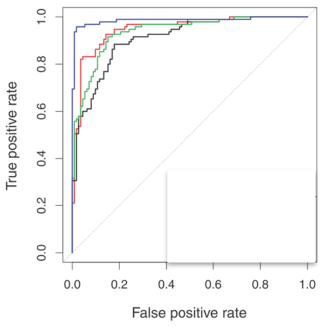
   (b) A linear model is fit to the data set, and the following table was returned.



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 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.



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:

$$X_1 = \texttt{Number\_of\_sentences}$$
$$X_2 = \texttt{Number\_of\_references\_to\_a\_Nigerian\_Prince}$$

and am training a logistic model to predict

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

(a) Write down the equation for the model you would train, using our standard notation with β_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 = GPA, X_2 = IQ, X_3 = Level (1 for College and 0 for High School), X_4 = Interaction between GPA and IQ, and X_5 = 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.

Scrap Paper

Michigan State University | **Feb 10, 2023**
CMSE381 - Data Science | **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. $\boxed{\text{BOX}}$ 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 irreduceable 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.

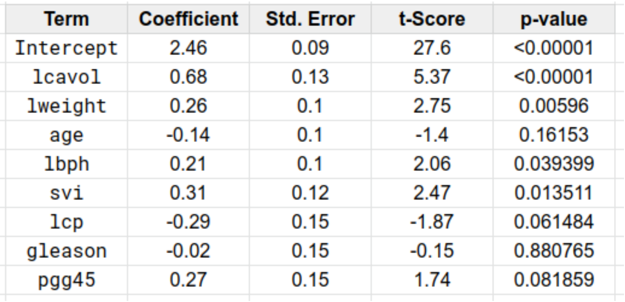
   (b) Provide a sketch of typical (squared) bias, variance, training error, test error, and Bayes (or irreducible) error curves, on a single plot, as we go from less flexible statistical learning methods towards more flexible approaches. The x-axis should represent the amount of flexibility in the method, 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?

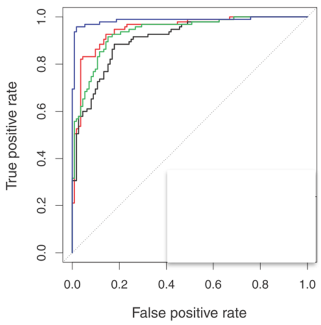
   (b) A linear model is fit to the data set, and the following table was returned.



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 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.



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:

$$X_1 = \texttt{Number\_of\_sentences}$$
$$X_2 = \texttt{Number\_of\_references\_to\_a\_Nigerian\_Prince}$$

and am training a logistic model to predict

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

(a) Write down the equation for the model you would train, using our standard notation with β_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 = GPA, X_2 = IQ, X_3 = Level (1 for College and 0 for High School), X_4 = Interaction between GPA and IQ, and X_5 = 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.

Scrap Paper

Michigan State University **Feb 10, 2023**
CMSE381 - Data Science **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. $\boxed{\text{BOX}}$ 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 irreduceable 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.

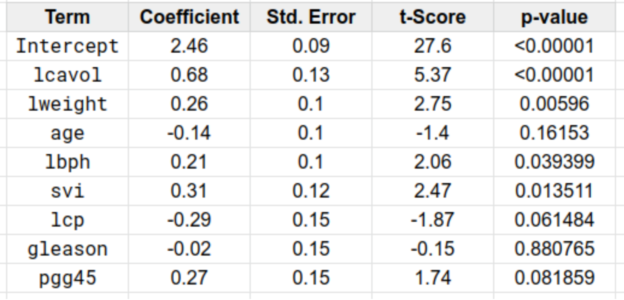
   (b) Provide a sketch of typical (squared) bias, variance, training error, test error, and Bayes (or irreducible) error curves, on a single plot, as we go from less flexible statistical learning methods towards more flexible approaches. The x-axis should represent the amount of flexibility in the method, 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?

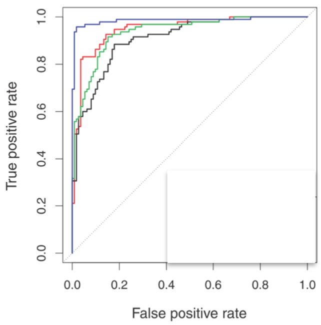
   (b) A linear model is fit to the data set, and the following table was returned.



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 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.



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:

$$X_1 = \texttt{Number\_of\_sentences}$$
$$X_2 = \texttt{Number\_of\_references\_to\_a\_Nigerian\_Prince}$$

and am training a logistic model to predict

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

(a) Write down the equation for the model you would train, using our standard notation with β_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 = GPA, X_2 = IQ, X_3 = Level (1 for College and 0 for High School), X_4 = Interaction between GPA and IQ, and X_5 = 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 $\beta_0 = 50$, $\beta_1 = 20$, $\beta_2 = 0.07$, $\beta_3 = 35$, $\beta_4 = 0.01$, β_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.

Scrap Paper

Michigan State University **Feb 10, 2023**
CMSE381 - Data Science **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 irreduceable 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.

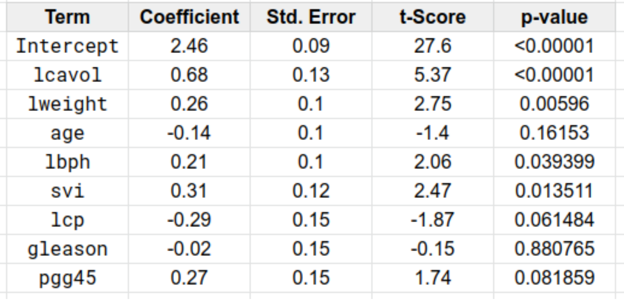
(b) Provide a sketch of typical (squared) bias, variance, training error, test error, and Bayes (or irreducible) error curves, on a single plot, as we go from less flexible statistical learning methods towards more flexible approaches. The x-axis should represent the amount of flexibility in the method, 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?

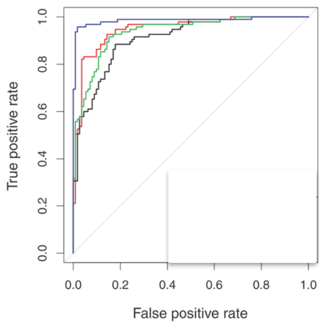
   (b) A linear model is fit to the data set, and the following table was returned.



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 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.



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:

$$X_1 = \texttt{Number\_of\_sentences}$$
$$X_2 = \texttt{Number\_of\_references\_to\_a\_Nigerian\_Prince}$$

and am training a logistic model to predict

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

(a) Write down the equation for the model you would train, using our standard notation with β_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 = GPA, X_2 = IQ, X_3 = Level (1 for College and 0 for High School), X_4 = Interaction between GPA and IQ, and X_5 = 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.

Scrap Paper

Michigan State University | **Feb 10, 2023**
CMSE381 - Data Science | **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. $\boxed{\text{BOX}}$ 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 irreduceable 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.

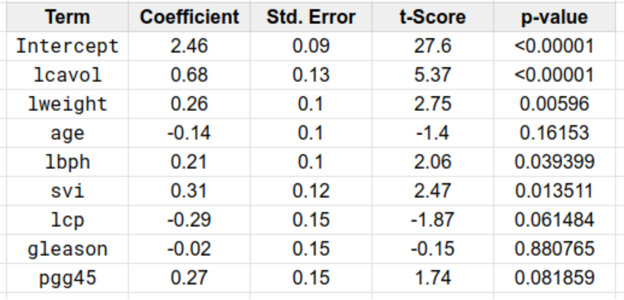
   (b) Provide a sketch of typical (squared) bias, variance, training error, test error, and Bayes (or irreducible) error curves, on a single plot, as we go from less flexible statistical learning methods towards more flexible approaches. The x-axis should represent the amount of flexibility in the method, 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?

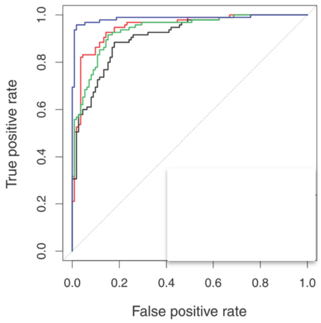
   (b) A linear model is fit to the data set, and the following table was returned.



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 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.



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:

$$X_1 = \texttt{Number\_of\_sentences}$$
$$X_2 = \texttt{Number\_of\_references\_to\_a\_Nigerian\_Prince}$$

and am training a logistic model to predict

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

(a) Write down the equation for the model you would train, using our standard notation with β_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 = GPA, X_2 = IQ, X_3 = Level (1 for College and 0 for High School), X_4 = Interaction between GPA and IQ, and X_5 = 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.

Scrap Paper

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. $\boxed{\text{BOX}}$ 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 irreduceable 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.

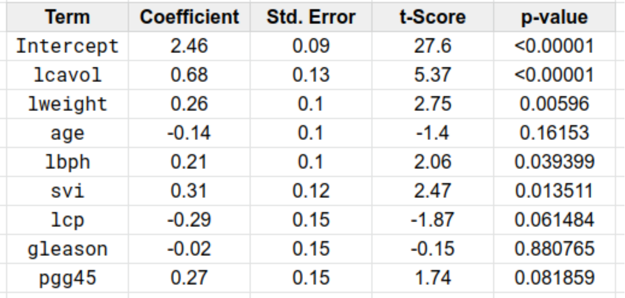
   (b) Provide a sketch of typical (squared) bias, variance, training error, test error, and Bayes (or irreducible) error curves, on a single plot, as we go from less flexible statistical learning methods towards more flexible approaches. The x-axis should represent the amount of flexibility in the method, 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?

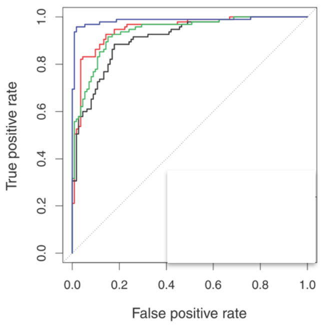
   (b) A linear model is fit to the data set, and the following table was returned.



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 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.



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:

$$X_1 = \texttt{Number\_of\_sentences}$$
$$X_2 = \texttt{Number\_of\_references\_to\_a\_Nigerian\_Prince}$$

and am training a logistic model to predict

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

(a) Write down the equation for the model you would train, using our standard notation with β_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 = GPA, X_2 = IQ, X_3 = Level (1 for College and 0 for High School), X_4 = Interaction between GPA and IQ, and X_5 = 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.

Scrap Paper

Michigan State University **Feb 10, 2023**
CMSE381 - Data Science **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. $\boxed{\text{BOX}}$ 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 irreduceable 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.

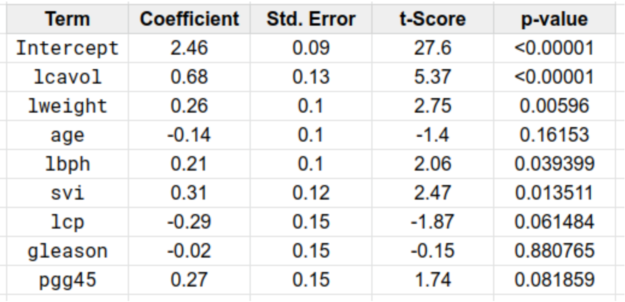
(b) Provide a sketch of typical (squared) bias, variance, training error, test error, and Bayes (or irreducible) error curves, on a single plot, as we go from less flexible statistical learning methods towards more flexible approaches. The x-axis should represent the amount of flexibility in the method, 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?

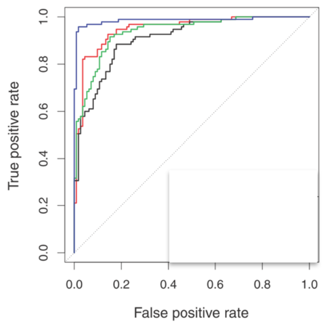
   (b) A linear model is fit to the data set, and the following table was returned.



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 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.



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:

$$X_1 = \texttt{Number\_of\_sentences}$$
$$X_2 = \texttt{Number\_of\_references\_to\_a\_Nigerian\_Prince}$$

and am training a logistic model to predict

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

(a) Write down the equation for the model you would train, using our standard notation with β_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 = GPA, X_2 = IQ, X_3 = Level (1 for College and 0 for High School), X_4 = Interaction between GPA and IQ, and X_5 = 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.

Scrap Paper

Michigan State University **Feb 10, 2023**
CMSE381 - Data Science **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 irreduceable 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.

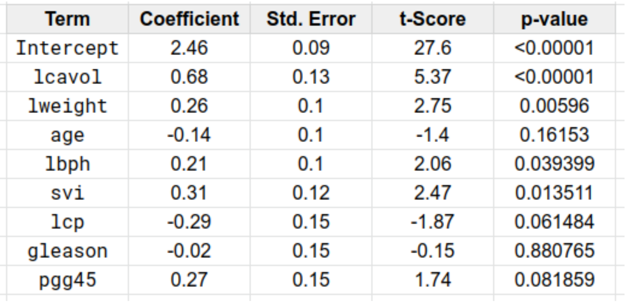
   (b) Provide a sketch of typical (squared) bias, variance, training error, test error, and Bayes (or irreducible) error curves, on a single plot, as we go from less flexible statistical learning methods towards more flexible approaches. The x-axis should represent the amount of flexibility in the method, 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?

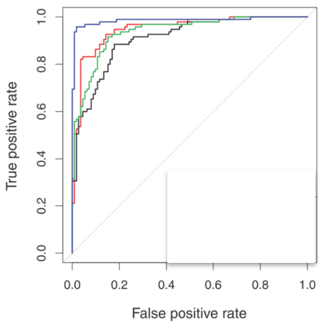
   (b) A linear model is fit to the data set, and the following table was returned.



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 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.



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:

$$X_1 = \texttt{Number\_of\_sentences}$$
$$X_2 = \texttt{Number\_of\_references\_to\_a\_Nigerian\_Prince}$$

and am training a logistic model to predict

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

(a) Write down the equation for the model you would train, using our standard notation with β_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 = GPA, X_2 = IQ, X_3 = Level (1 for College and 0 for High School), X_4 = Interaction between GPA and IQ, and X_5 = 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.

Scrap Paper

Michigan State University **Feb 10, 2023**
CMSE381 - Data Science **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. $\boxed{\text{BOX}}$ 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 irreduceable 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.

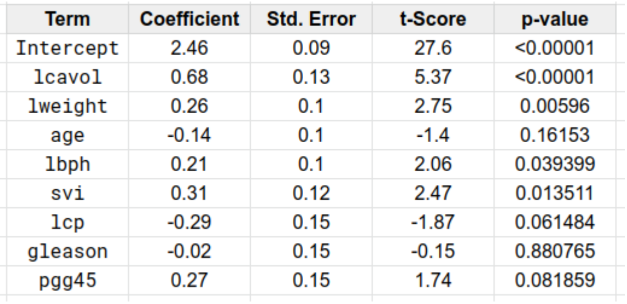
   (b) Provide a sketch of typical (squared) bias, variance, training error, test error, and Bayes (or irreducible) error curves, on a single plot, as we go from less flexible statistical learning methods towards more flexible approaches. The x-axis should represent the amount of flexibility in the method, 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?

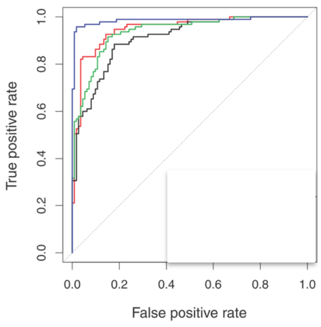
   (b) A linear model is fit to the data set, and the following table was returned.



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 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.



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:

$$X_1 = \texttt{Number\_of\_sentences}$$
$$X_2 = \texttt{Number\_of\_references\_to\_a\_Nigerian\_Prince}$$

and am training a logistic model to predict

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

(a) Write down the equation for the model you would train, using our standard notation with β_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 = GPA, X_2 = IQ, X_3 = Level (1 for College and 0 for High School), X_4 = Interaction between GPA and IQ, and X_5 = 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 β_0 = 50, β_1 = 20, β_2 = 0.07, β_3 = 35, β_4 = 0.01, β_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.

Scrap Paper

Michigan State University **Feb 10, 2023**
CMSE381 - Data Science **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. $\boxed{\text{BOX}}$ 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 irreduceable 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.

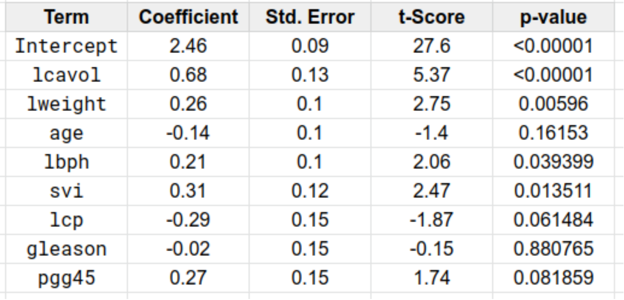
   (b) Provide a sketch of typical (squared) bias, variance, training error, test error, and Bayes (or irreducible) error curves, on a single plot, as we go from less flexible statistical learning methods towards more flexible approaches. The x-axis should represent the amount of flexibility in the method, 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?

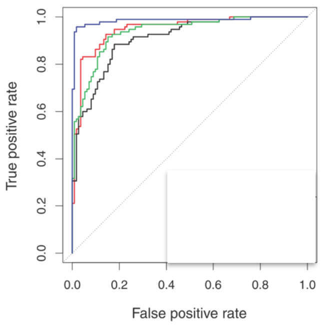
   (b) A linear model is fit to the data set, and the following table was returned.



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 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.



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:

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

and am training a logistic model to predict

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

(a) Write down the equation for the model you would train, using our standard notation with β_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 = GPA, X_2 = IQ, X_3 = Level (1 for College and 0 for High School), X_4 = Interaction between GPA and IQ, and X_5 = 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.

Scrap Paper

Michigan State University **Feb 10, 2023**
CMSE381 - Data Science **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. $\boxed{\text{BOX}}$ 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 irreduceable 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.

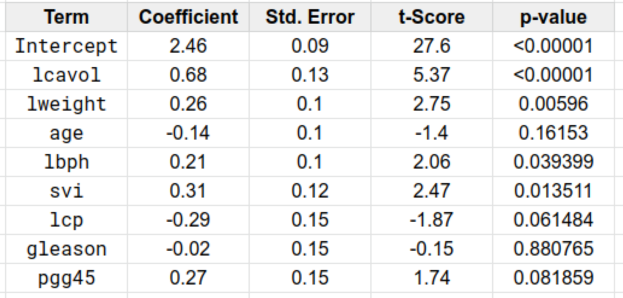
(b) Provide a sketch of typical (squared) bias, variance, training error, test error, and Bayes (or irreducible) error curves, on a single plot, as we go from less flexible statistical learning methods towards more flexible approaches. The x-axis should represent the amount of flexibility in the method, 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?

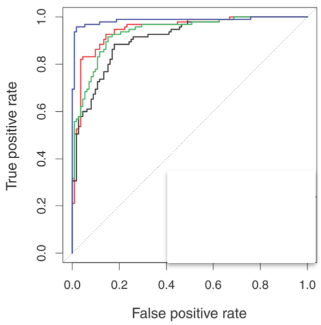
   (b) A linear model is fit to the data set, and the following table was returned.



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 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.



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:

$$X_1 = \texttt{Number\_of\_sentences}$$
$$X_2 = \texttt{Number\_of\_references\_to\_a\_Nigerian\_Prince}$$

and am training a logistic model to predict

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

(a) Write down the equation for the model you would train, using our standard notation with β_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 = GPA, X_2 = IQ, X_3 = Level (1 for College and 0 for High School), X_4 = Interaction between GPA and IQ, and X_5 = 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.

Scrap Paper

Michigan State University | **Feb 10, 2023**
CMSE381 - Data Science | **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. $\boxed{\text{BOX}}$ 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 irreduceable 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.

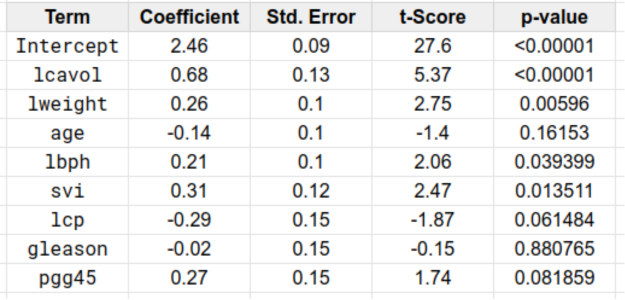
(b) Provide a sketch of typical (squared) bias, variance, training error, test error, and Bayes (or irreducible) error curves, on a single plot, as we go from less flexible statistical learning methods towards more flexible approaches. The x-axis should represent the amount of flexibility in the method, 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?

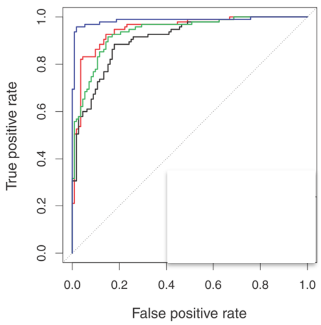
   (b) A linear model is fit to the data set, and the following table was returned.



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 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.



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:

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

and am training a logistic model to predict

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

(a) Write down the equation for the model you would train, using our standard notation with β_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 = GPA, X_2 = IQ, X_3 = Level (1 for College and 0 for High School), X_4 = Interaction between GPA and IQ, and X_5 = 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.

Scrap Paper

Michigan State University **Feb 10, 2023**
CMSE381 - Data Science **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. $\boxed{\text{BOX}}$ 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 irreduceable 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.

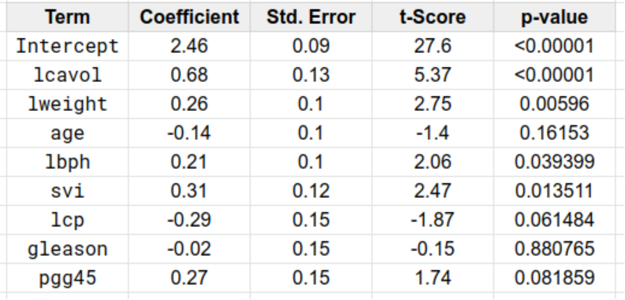
   (b) Provide a sketch of typical (squared) bias, variance, training error, test error, and Bayes (or irreducible) error curves, on a single plot, as we go from less flexible statistical learning methods towards more flexible approaches. The x-axis should represent the amount of flexibility in the method, 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?

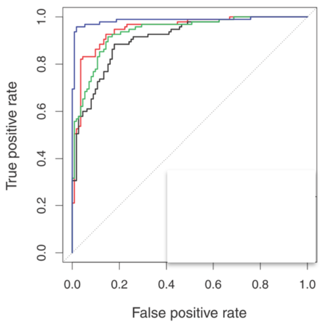
   (b) A linear model is fit to the data set, and the following table was returned.



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 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.



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:

$$X_1 = \texttt{Number\_of\_sentences}$$
$$X_2 = \texttt{Number\_of\_references\_to\_a\_Nigerian\_Prince}$$

and am training a logistic model to predict

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

(a) Write down the equation for the model you would train, using our standard notation with β_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 = GPA, X_2 = IQ, X_3 = Level (1 for College and 0 for High School), X_4 = Interaction between GPA and IQ, and X_5 = 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.

Scrap Paper

Michigan State University **Feb 10, 2023**
CMSE381 - Data Science **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. $\boxed{\text{BOX}}$ 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 irreduceable 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.

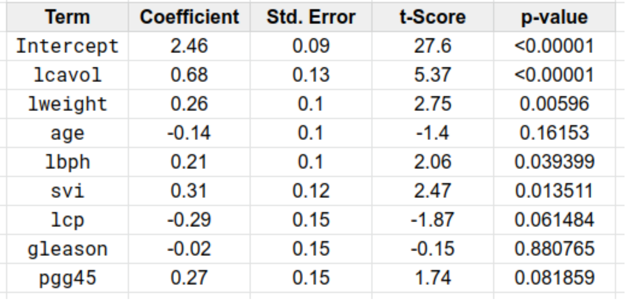
(b) Provide a sketch of typical (squared) bias, variance, training error, test error, and Bayes (or irreducible) error curves, on a single plot, as we go from less flexible statistical learning methods towards more flexible approaches. The x-axis should represent the amount of flexibility in the method, 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?

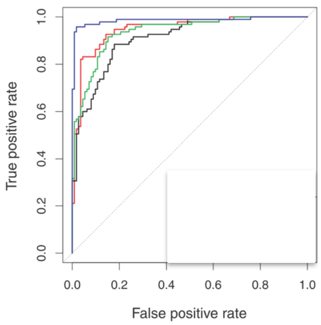
   (b) A linear model is fit to the data set, and the following table was returned.



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 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.



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:

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

and am training a logistic model to predict

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

(a) Write down the equation for the model you would train, using our standard notation with β_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 = GPA, X_2 = IQ, X_3 = Level (1 for College and 0 for High School), X_4 = Interaction between GPA and IQ, and X_5 = 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.

Scrap Paper

Michigan State University **Feb 10, 2023**
CMSE381 - Data Science **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. $\boxed{\text{BOX}}$ 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 irreduceable 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.

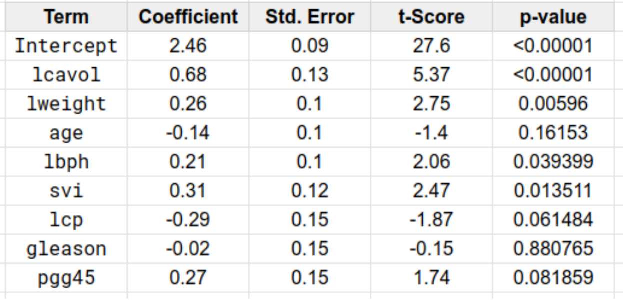
   (b) Provide a sketch of typical (squared) bias, variance, training error, test error, and Bayes (or irreducible) error curves, on a single plot, as we go from less flexible statistical learning methods towards more flexible approaches. The x-axis should represent the amount of flexibility in the method, 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?

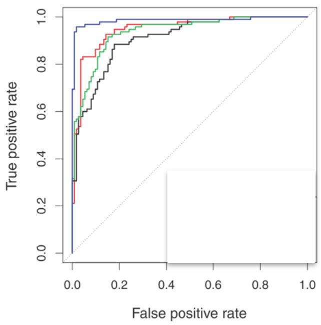
   (b) A linear model is fit to the data set, and the following table was returned.



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 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.



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:

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

and am training a logistic model to predict

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

(a) Write down the equation for the model you would train, using our standard notation with β_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 = GPA, X_2 = IQ, X_3 = Level (1 for College and 0 for High School), X_4 = Interaction between GPA and IQ, and X_5 = 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.

Scrap Paper

Michigan State University — **Feb 10, 2023**
CMSE381 - Data Science — **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. $\boxed{\text{BOX}}$ 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 irreduceable 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.

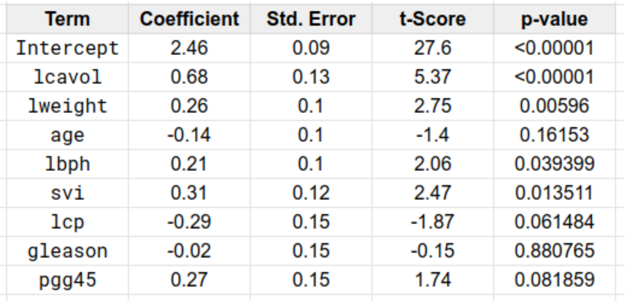
(b) Provide a sketch of typical (squared) bias, variance, training error, test error, and Bayes (or irreducible) error curves, on a single plot, as we go from less flexible statistical learning methods towards more flexible approaches. The x-axis should represent the amount of flexibility in the method, 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?

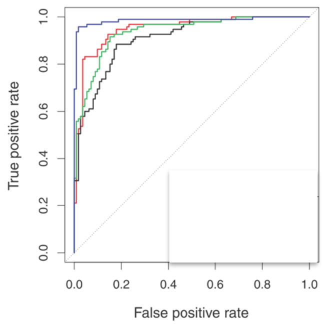
   (b) A linear model is fit to the data set, and the following table was returned.



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 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.



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:

$$X_1 = \texttt{Number\_of\_sentences}$$
$$X_2 = \texttt{Number\_of\_references\_to\_a\_Nigerian\_Prince}$$

and am training a logistic model to predict

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

(a) Write down the equation for the model you would train, using our standard notation with β_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 = GPA, X_2 = IQ, X_3 = Level (1 for College and 0 for High School), X_4 = Interaction between GPA and IQ, and X_5 = 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.

Scrap Paper

Michigan State University **Feb 10, 2023**
CMSE381 - Data Science **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. $\boxed{\text{BOX}}$ 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 irreduceable 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.

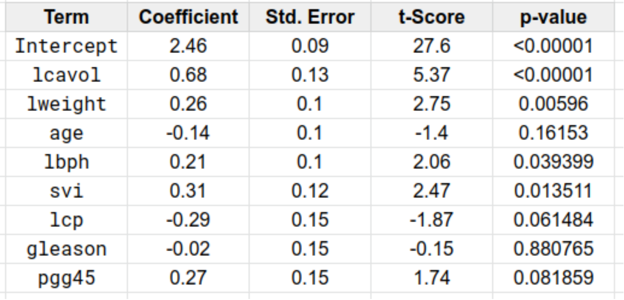
   (b) Provide a sketch of typical (squared) bias, variance, training error, test error, and Bayes (or irreducible) error curves, on a single plot, as we go from less flexible statistical learning methods towards more flexible approaches. The x-axis should represent the amount of flexibility in the method, 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?

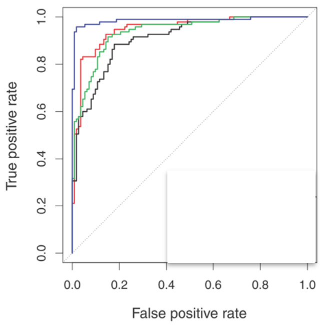
   (b) A linear model is fit to the data set, and the following table was returned.



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 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.



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:

$$X_1 = \texttt{Number\_of\_sentences}$$
$$X_2 = \texttt{Number\_of\_references\_to\_a\_Nigerian\_Prince}$$

and am training a logistic model to predict

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

(a) Write down the equation for the model you would train, using our standard notation with β_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 = GPA, X_2 = IQ, X_3 = Level (1 for College and 0 for High School), X_4 = Interaction between GPA and IQ, and X_5 = 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.

Scrap Paper

Michigan State University **Feb 10, 2023**
CMSE381 - Data Science **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 irreduceable 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.

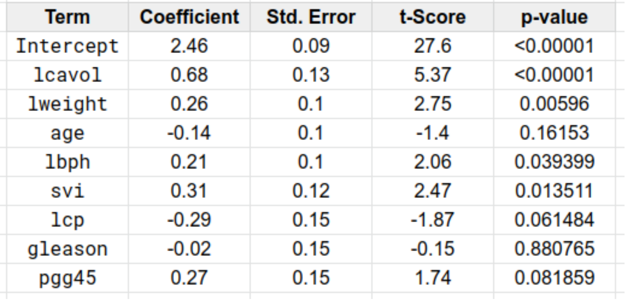
(b) Provide a sketch of typical (squared) bias, variance, training error, test error, and Bayes (or irreducible) error curves, on a single plot, as we go from less flexible statistical learning methods towards more flexible approaches. The x-axis should represent the amount of flexibility in the method, 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?

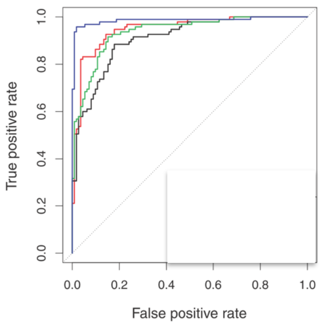
   (b) A linear model is fit to the data set, and the following table was returned.



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 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.



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:

$$X_1 = \texttt{Number\_of\_sentences}$$
$$X_2 = \texttt{Number\_of\_references\_to\_a\_Nigerian\_Prince}$$

and am training a logistic model to predict

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

(a) Write down the equation for the model you would train, using our standard notation with β_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 = GPA, X_2 = IQ, X_3 = Level (1 for College and 0 for High School), X_4 = Interaction between GPA and IQ, and X_5 = 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.

Scrap Paper

Michigan State University **Feb 10, 2023**
CMSE381 - Data Science **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. $\boxed{\text{BOX}}$ 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 irreduceable 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.

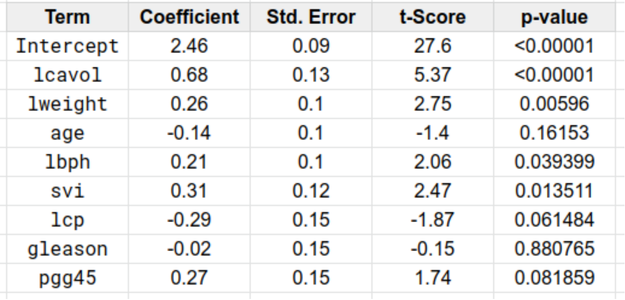
(b) Provide a sketch of typical (squared) bias, variance, training error, test error, and Bayes (or irreducible) error curves, on a single plot, as we go from less flexible statistical learning methods towards more flexible approaches. The x-axis should represent the amount of flexibility in the method, 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?

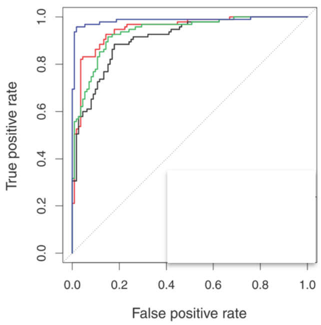
   (b) A linear model is fit to the data set, and the following table was returned.



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 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.



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:

$$X_1 = \texttt{Number\_of\_sentences}$$
$$X_2 = \texttt{Number\_of\_references\_to\_a\_Nigerian\_Prince}$$

and am training a logistic model to predict

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

(a) Write down the equation for the model you would train, using our standard notation with β_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 = GPA, X_2 = IQ, X_3 = Level (1 for College and 0 for High School), X_4 = Interaction between GPA and IQ, and X_5 = 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 $\beta_0 = 50$, $\beta_1 = 20$, $\beta_2 = 0.07$, $\beta_3 = 35$, $\beta_4 = 0.01$, β_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.

Scrap Paper

Michigan State University — **Feb 10, 2023**
CMSE381 - Data Science — **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. $\boxed{\text{BOX}}$ 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 irreduceable 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.

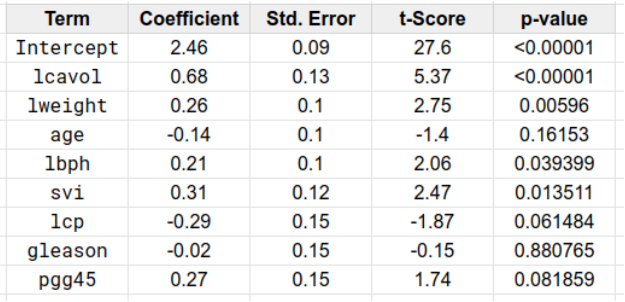
   (b) Provide a sketch of typical (squared) bias, variance, training error, test error, and Bayes (or irreducible) error curves, on a single plot, as we go from less flexible statistical learning methods towards more flexible approaches. The x-axis should represent the amount of flexibility in the method, 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?

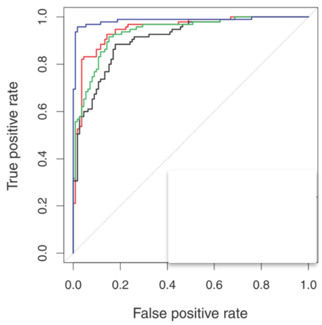
   (b) A linear model is fit to the data set, and the following table was returned.



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 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.



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:

$$X_1 = \texttt{Number\_of\_sentences}$$
$$X_2 = \texttt{Number\_of\_references\_to\_a\_Nigerian\_Prince}$$

and am training a logistic model to predict

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

(a) Write down the equation for the model you would train, using our standard notation with β_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 = GPA, X_2 = IQ, X_3 = Level (1 for College and 0 for High School), X_4 = Interaction between GPA and IQ, and X_5 = 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 $\beta_0 = 50$, $\beta_1 = 20$, $\beta_2 = 0.07$, $\beta_3 = 35$, $\beta_4 = 0.01$, $\beta_5 = \text{-}10$.

   (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.

Scrap Paper

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. $\boxed{\text{BOX}}$ 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 irreduceable 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.

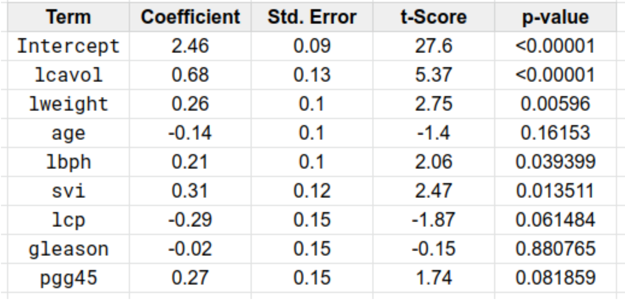
(b) Provide a sketch of typical (squared) bias, variance, training error, test error, and Bayes (or irreducible) error curves, on a single plot, as we go from less flexible statistical learning methods towards more flexible approaches. The x-axis should represent the amount of flexibility in the method, 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?

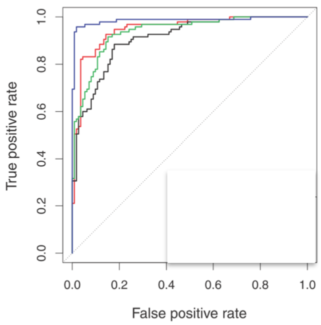
   (b) A linear model is fit to the data set, and the following table was returned.



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 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.



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:

$$X_1 = \texttt{Number\_of\_sentences}$$
$$X_2 = \texttt{Number\_of\_references\_to\_a\_Nigerian\_Prince}$$

and am training a logistic model to predict

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

(a) Write down the equation for the model you would train, using our standard notation with β_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 = GPA, X_2 = IQ, X_3 = Level (1 for College and 0 for High School), X_4 = Interaction between GPA and IQ, and X_5 = 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 β_0 = 50, β_1 = 20, β_2 = 0.07, β_3 = 35, β_4 = 0.01, β_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.

Scrap Paper

Michigan State University | **Feb 10, 2023**
CMSE381 - Data Science | **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. $\boxed{\text{BOX}}$ 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 irreduceable 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.

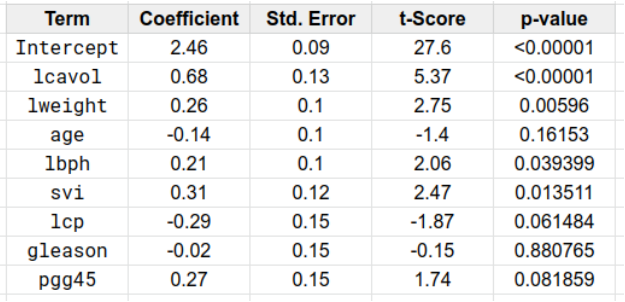
   (b) Provide a sketch of typical (squared) bias, variance, training error, test error, and Bayes (or irreducible) error curves, on a single plot, as we go from less flexible statistical learning methods towards more flexible approaches. The x-axis should represent the amount of flexibility in the method, 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?

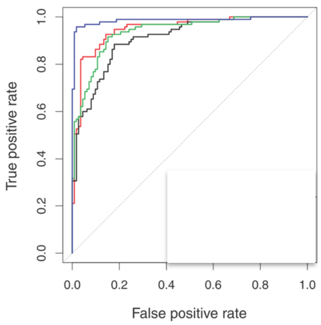
   (b) A linear model is fit to the data set, and the following table was returned.



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 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.



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:

$$X_1 = \texttt{Number\_of\_sentences}$$
$$X_2 = \texttt{Number\_of\_references\_to\_a\_Nigerian\_Prince}$$

and am training a logistic model to predict

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

(a) Write down the equation for the model you would train, using our standard notation with β_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 = GPA, X_2 = IQ, X_3 = Level (1 for College and 0 for High School), X_4 = Interaction between GPA and IQ, and X_5 = 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 $\beta_0 = 50$, $\beta_1 = 20$, $\beta_2 = 0.07$, $\beta_3 = 35$, $\beta_4 = 0.01$, β_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.

Scrap Paper

Michigan State University **Feb 10, 2023**
CMSE381 - Data Science **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. $\boxed{\text{BOX}}$ 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 irreduceable 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.

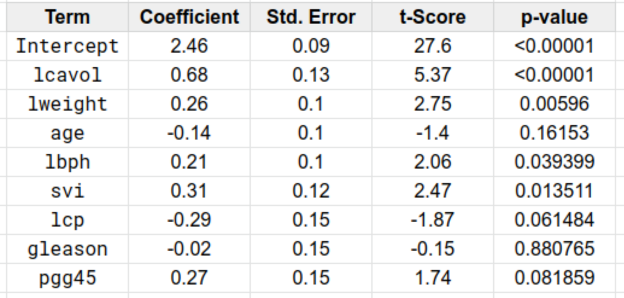
(b) Provide a sketch of typical (squared) bias, variance, training error, test error, and Bayes (or irreducible) error curves, on a single plot, as we go from less flexible statistical learning methods towards more flexible approaches. The x-axis should represent the amount of flexibility in the method, 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?

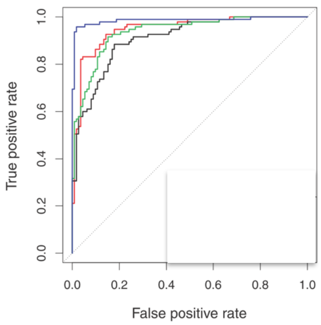
   (b) A linear model is fit to the data set, and the following table was returned.



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 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.



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:

$$X_1 = \texttt{Number\_of\_sentences}$$
$$X_2 = \texttt{Number\_of\_references\_to\_a\_Nigerian\_Prince}$$

and am training a logistic model to predict

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

(a) Write down the equation for the model you would train, using our standard notation with β_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 = GPA, X_2 = IQ, X_3 = Level (1 for College and 0 for High School), X_4 = Interaction between GPA and IQ, and X_5 = 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 β_0 = 50, β_1 = 20, β_2 = 0.07, β_3 = 35, β_4 = 0.01, β_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.

Scrap Paper

Michigan State University **Feb 10, 2023**
CMSE381 - Data Science **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. $\boxed{\text{BOX}}$ 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 irreduceable 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.

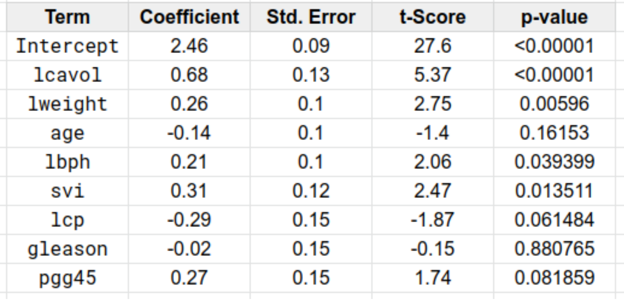
(b) Provide a sketch of typical (squared) bias, variance, training error, test error, and Bayes (or irreducible) error curves, on a single plot, as we go from less flexible statistical learning methods towards more flexible approaches. The x-axis should represent the amount of flexibility in the method, 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?

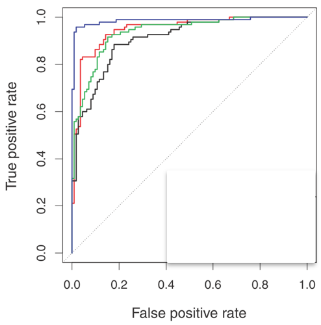
   (b) A linear model is fit to the data set, and the following table was returned.



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 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.



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:

$$X_1 = \texttt{Number\_of\_sentences}$$
$$X_2 = \texttt{Number\_of\_references\_to\_a\_Nigerian\_Prince}$$

and am training a logistic model to predict

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

(a) Write down the equation for the model you would train, using our standard notation with β_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 = GPA, X_2 = IQ, X_3 = Level (1 for College and 0 for High School), X_4 = Interaction between GPA and IQ, and X_5 = 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.

Scrap Paper

Michigan State University **Feb 10, 2023**
CMSE381 - Data Science **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. $\boxed{\text{BOX}}$ 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 irreduceable 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.

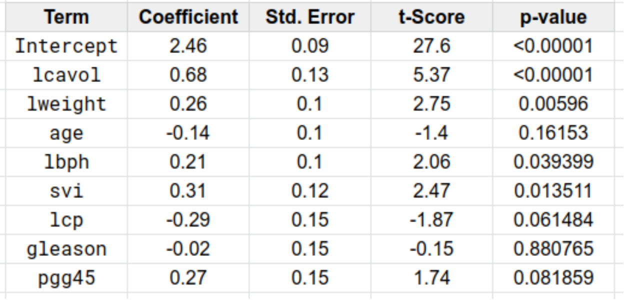
   (b) Provide a sketch of typical (squared) bias, variance, training error, test error, and Bayes (or irreducible) error curves, on a single plot, as we go from less flexible statistical learning methods towards more flexible approaches. The x-axis should represent the amount of flexibility in the method, 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?

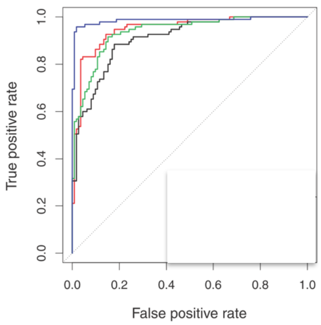
   (b) A linear model is fit to the data set, and the following table was returned.



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 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.



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:

$$X_1 = \texttt{Number\_of\_sentences}$$
$$X_2 = \texttt{Number\_of\_references\_to\_a\_Nigerian\_Prince}$$

and am training a logistic model to predict

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

(a) Write down the equation for the model you would train, using our standard notation with β_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 = GPA, X_2 = IQ, X_3 = Level (1 for College and 0 for High School), X_4 = Interaction between GPA and IQ, and X_5 = 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.

Scrap Paper

Michigan State University **Feb 10, 2023**
CMSE381 - Data Science **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. $\boxed{\text{BOX}}$ 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 irreduceable 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.

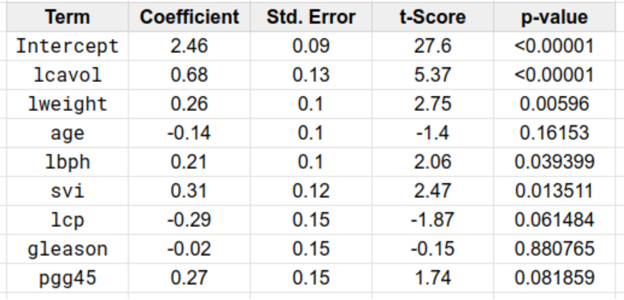
   (b) Provide a sketch of typical (squared) bias, variance, training error, test error, and Bayes (or irreducible) error curves, on a single plot, as we go from less flexible statistical learning methods towards more flexible approaches. The x-axis should represent the amount of flexibility in the method, 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?

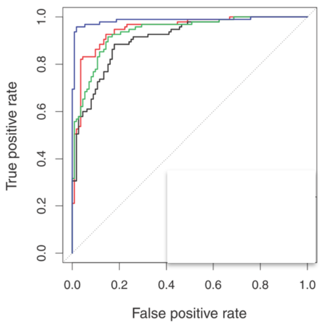
   (b) A linear model is fit to the data set, and the following table was returned.



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 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.



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:

$$X_1 = \texttt{Number\_of\_sentences}$$
$$X_2 = \texttt{Number\_of\_references\_to\_a\_Nigerian\_Prince}$$

and am training a logistic model to predict

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

(a) Write down the equation for the model you would train, using our standard notation with β_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 = GPA, X_2 = IQ, X_3 = Level (1 for College and 0 for High School), X_4 = Interaction between GPA and IQ, and X_5 = 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 β_0 = 50, β_1 = 20, β_2 = 0.07, β_3 = 35, β_4 = 0.01, β_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.

Scrap Paper

Michigan State University **Feb 10, 2023**
CMSE381 - Data Science **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 irreduceable 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.

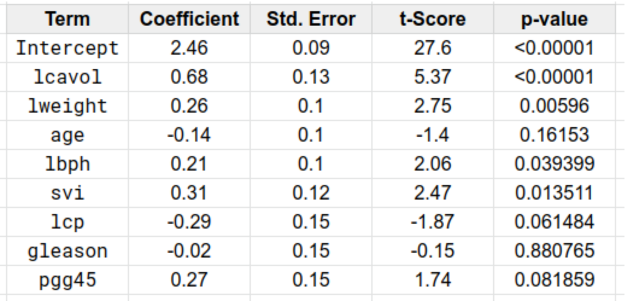
(b) Provide a sketch of typical (squared) bias, variance, training error, test error, and Bayes (or irreducible) error curves, on a single plot, as we go from less flexible statistical learning methods towards more flexible approaches. The x-axis should represent the amount of flexibility in the method, 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?

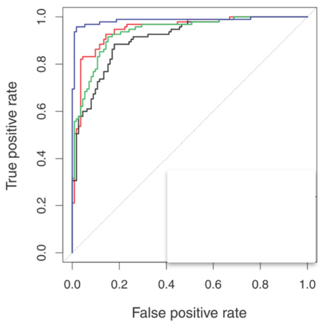
   (b) A linear model is fit to the data set, and the following table was returned.



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 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.



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:

$$X_1 = \texttt{Number\_of\_sentences}$$
$$X_2 = \texttt{Number\_of\_references\_to\_a\_Nigerian\_Prince}$$

and am training a logistic model to predict

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

(a) Write down the equation for the model you would train, using our standard notation with β_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 = GPA, X_2 = IQ, X_3 = Level (1 for College and 0 for High School), X_4 = Interaction between GPA and IQ, and X_5 = 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.

Scrap Paper

Michigan State University **Feb 10, 2023**
CMSE381 - Data Science **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. $\boxed{\text{BOX}}$ 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 irreduceable 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.

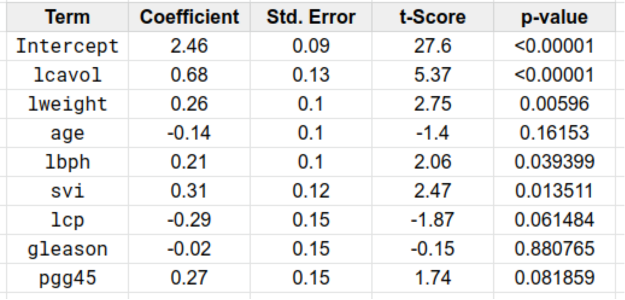
   (b) Provide a sketch of typical (squared) bias, variance, training error, test error, and Bayes (or irreducible) error curves, on a single plot, as we go from less flexible statistical learning methods towards more flexible approaches. The x-axis should represent the amount of flexibility in the method, 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?

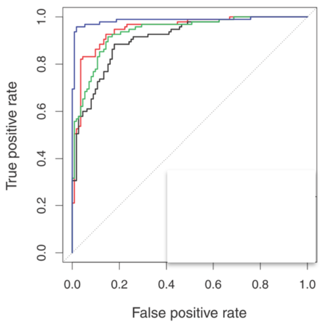
   (b) A linear model is fit to the data set, and the following table was returned.



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 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.



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:

$$X_1 = \texttt{Number\_of\_sentences}$$
$$X_2 = \texttt{Number\_of\_references\_to\_a\_Nigerian\_Prince}$$

and am training a logistic model to predict

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

(a) Write down the equation for the model you would train, using our standard notation with β_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 = GPA, X_2 = IQ, X_3 = Level (1 for College and 0 for High School), X_4 = Interaction between GPA and IQ, and X_5 = 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 $\beta_0 = 50$, $\beta_1 = 20$, $\beta_2 = 0.07$, $\beta_3 = 35$, $\beta_4 = 0.01$, $\beta_5 = \text{-}10$.

   (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.

Scrap Paper

Michigan State University **Feb 10, 2023**
CMSE381 - Data Science **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. $\boxed{\text{BOX}}$ 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 irreduceable 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.

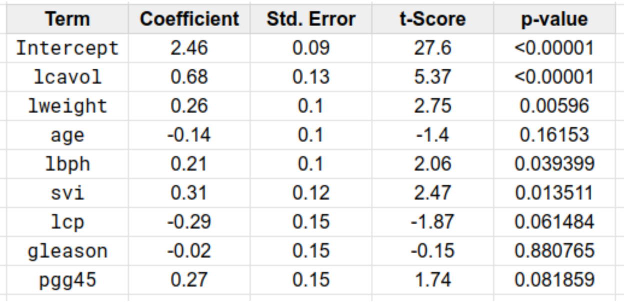
(b) Provide a sketch of typical (squared) bias, variance, training error, test error, and Bayes (or irreducible) error curves, on a single plot, as we go from less flexible statistical learning methods towards more flexible approaches. The x-axis should represent the amount of flexibility in the method, 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?

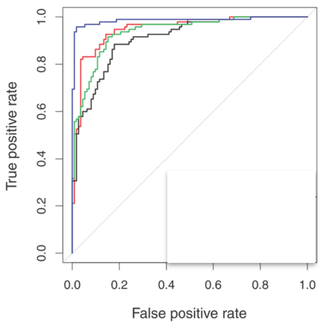
   (b) A linear model is fit to the data set, and the following table was returned.



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 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.



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:

$$X_1 = \texttt{Number\_of\_sentences}$$
$$X_2 = \texttt{Number\_of\_references\_to\_a\_Nigerian\_Prince}$$

and am training a logistic model to predict

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

(a) Write down the equation for the model you would train, using our standard notation with β_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 = GPA, X_2 = IQ, X_3 = Level (1 for College and 0 for High School), X_4 = Interaction between GPA and IQ, and X_5 = 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 β_0 = 50, β_1 = 20, β_2 = 0.07, β_3 = 35, β_4 = 0.01, β_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.

Scrap Paper

Michigan State University **Feb 10, 2023**
CMSE381 - Data Science **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. $\boxed{\text{BOX}}$ 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 irreduceable 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.

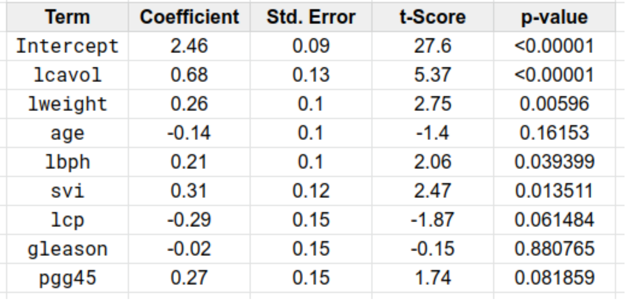
   (b) Provide a sketch of typical (squared) bias, variance, training error, test error, and Bayes (or irreducible) error curves, on a single plot, as we go from less flexible statistical learning methods towards more flexible approaches. The x-axis should represent the amount of flexibility in the method, 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?

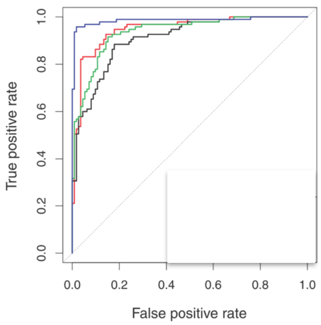
   (b) A linear model is fit to the data set, and the following table was returned.



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 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.



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:

$$X_1 = \texttt{Number\_of\_sentences}$$
$$X_2 = \texttt{Number\_of\_references\_to\_a\_Nigerian\_Prince}$$

and am training a logistic model to predict

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

(a) Write down the equation for the model you would train, using our standard notation with β_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 = GPA, X_2 = IQ, X_3 = Level (1 for College and 0 for High School), X_4 = Interaction between GPA and IQ, and X_5 = 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.

Scrap Paper

Michigan State University **Feb 10, 2023**
CMSE381 - Data Science **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. $\boxed{\text{BOX}}$ 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 irreduceable 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.

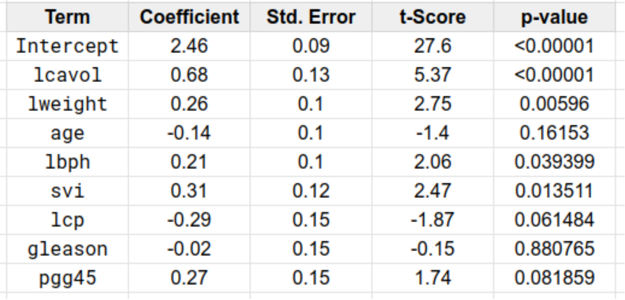
(b) Provide a sketch of typical (squared) bias, variance, training error, test error, and Bayes (or irreducible) error curves, on a single plot, as we go from less flexible statistical learning methods towards more flexible approaches. The x-axis should represent the amount of flexibility in the method, 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?

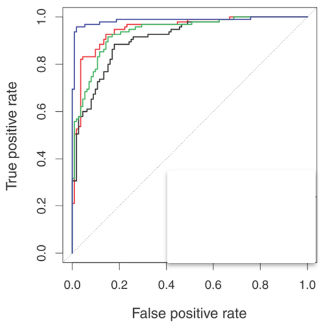
   (b) A linear model is fit to the data set, and the following table was returned.



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 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.



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:

$$X_1 = \texttt{Number\_of\_sentences}$$
$$X_2 = \texttt{Number\_of\_references\_to\_a\_Nigerian\_Prince}$$

and am training a logistic model to predict

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

(a) Write down the equation for the model you would train, using our standard notation with β_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 = GPA, X_2 = IQ, X_3 = Level (1 for College and 0 for High School), X_4 = Interaction between GPA and IQ, and X_5 = 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.

Scrap Paper

Michigan State University **Feb 10, 2023**
CMSE381 - Data Science **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 irreduceable 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.

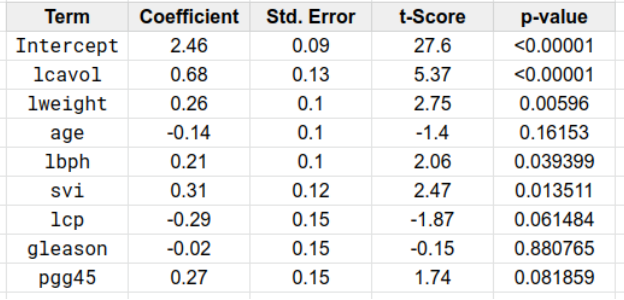
   (b) Provide a sketch of typical (squared) bias, variance, training error, test error, and Bayes (or irreducible) error curves, on a single plot, as we go from less flexible statistical learning methods towards more flexible approaches. The x-axis should represent the amount of flexibility in the method, 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?

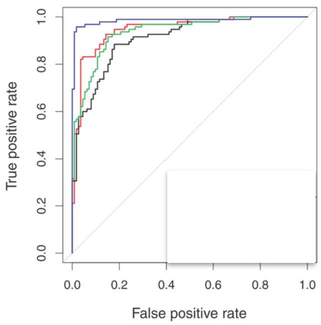
   (b) A linear model is fit to the data set, and the following table was returned.



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 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.



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:

$$X_1 = \texttt{Number\_of\_sentences}$$
$$X_2 = \texttt{Number\_of\_references\_to\_a\_Nigerian\_Prince}$$

and am training a logistic model to predict

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

(a) Write down the equation for the model you would train, using our standard notation with β_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 = GPA, X_2 = IQ, X_3 = Level (1 for College and 0 for High School), X_4 = Interaction between GPA and IQ, and X_5 = 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.

Scrap Paper