Michigan State University **April, 2024**
CMSE381 - Data Science **Dr. Wang**

CMSE381 - Final

1. Do not open this test booklet until you are directed to do so.
2. You will have 2 hours (5:45-7:45pm) to complete the exam. If you finish early go back and check your work.
3. This exam is open book. The use of generative AI is not allowed.
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

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

Signed: _____ Print Name: _____

1. (10 points)

(a) We know the true test error in the case of simulated data, but not for real data.

TRUE FALSE

each problem: 2pts

(b) Logistic regression is used for regression.

TRUE FALSE

(c) When doing regression using a categorical predictor variable taking values `N,S,E,W`, it is enough to associate each letter to a number and then proceed to do regression as usual.

TRUE FALSE

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

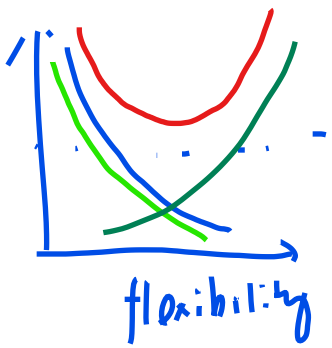
TRUE FALSE

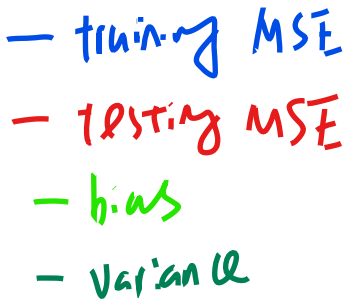
(e) Which of the following is associated with or can be explained by the variance-bias tradeoff? Circle all that apply

i. The difference between LOOCV and k-fold cross-validation
ii. The extension from decision tree to random forests
iii. The extension from LDA to QDA
iv. the U-shape curve for test error in the over-fitting plot

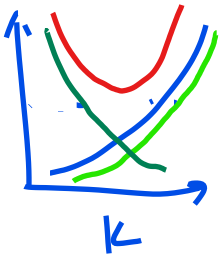
2. (9 points)

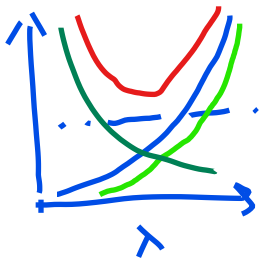
3pts (a) Provide a sketch of typical (squared) bias, variance, training error, test error, and Bayes (or irreducible) error curves, on a single plot, as we go from less flexible statistical learning methods towards more flexible approaches. The x-axis should represent the amount of flexibility (or complexity) 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.





3pts (b) Repeat the above plot for the kNN and Lasso methods. More specifically, for kNN, please use the number of nearest neighbours k as the x-axis, and plot the tendency of the 5 curves as k increases. For Lasso, please use the parameter lambda as the x-axis, and plot the tendency of the 5 curves as λ increases.





3pts (c) For Lasso, explain why the (i) training error and (ii) testing error lines in your drawing have the shape displayed.

3. (8 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?

3pts

5pts (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.

No, some predictor has too large of p values, remove those and try linear regression again

2pts for saying no, 3 pts for the correct explanation

4. (10 points) Suppose we collect data for a group of students in a statistics class with variables $X_1 =$ hours studied, $X_2 =$ undergrad GPA, and $Y =$ received an A. We fit a logistic regression and produce estimated coefficients, $\hat{\beta}_0 = -5$, $\hat{\beta}_1 = 0.06$ and $\hat{\beta}_2 = 1.2$

(a) Estimate the probability that a student who studies for 15 hours and has an undergrad GPA of 3.5 gets an A in the class.

4pts

$$-5 + 0.06 \cdot 15 + 1.2 \cdot 3.5 = 0.1$$

$$\frac{e^{0.1}}{1 + e^{0.1}}$$

(b) How many hours would the student in part (a) need to study to have a 75% chance of getting an A in the class?

6pts

75% chance $\Rightarrow$ odd is 3

2 points for finding odds

$$\log(3) = -5 + 0.06 \cdot X + 1.2 \cdot 3.5$$

2 pts for correctly specifying the linear relation between the odd and predictors

$$\Rightarrow X = \frac{\log 3 + 0.8}{0.06}$$

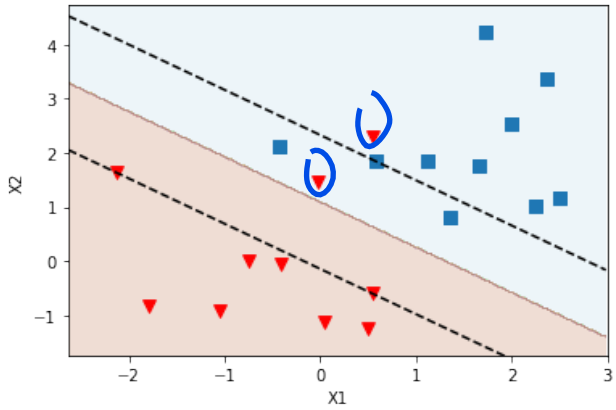
2 pt for correct final answer

Other methods are permissible, however, only giving a final answer without any derivation or explanation receives 2 pt

5. (12 points) A data set is given. We fit a support vector machine (SVC) and the resulting hyperplane and margin are drawn.

(a) Circle all the points which are misclassified by SVC

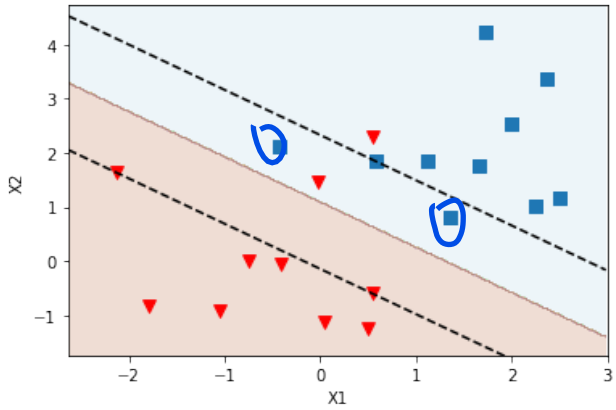
4pts



each correctly circled point is worth 2 pts

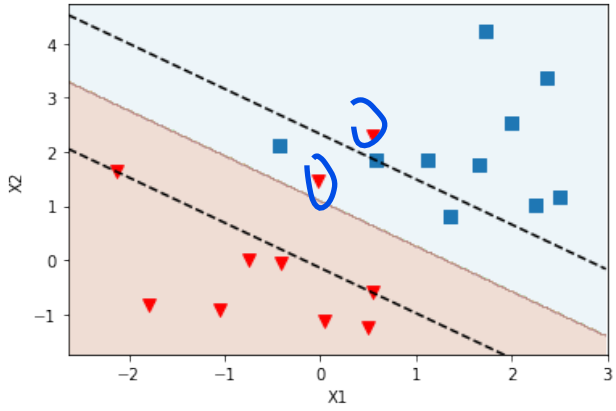
(b) Circle the points whose ε_i value is between 0 and 1. i.e. $0 < \varepsilon_i < 1$

4pts



(c) Circle all the points whose ε_i value is larger than 1,

4pts



6. (12 points) We train a model using four variables, X_1, X_2, X_3, X_4. We're interested in getting a subset of the variables to use. The following table shows the **testing mean squared error** and the **training** R^2 computed for the model learned using each possible subset of variables.

Testing Training

	MSE (x10^7)	R^2
Null model	2.71	0.78
X1	5.32	0.63
X2	1.85	0.79
X3	2.15	0.81
X4	3.81	0.58
X1,X2	1.63	0.83
X1,X3	2.27	0.58
X1,X4	3.17	0.47
X2,X3	4.38	0.42
X2,X4	5.32	0.51
X3,X4	1.52	0.82
X1,X2,X3	2.08	0.61
X1,X2,X4	1.55	0.81
X1,X3,X4	1.73	0.89
X2,X3,X4	1.53	0.78
X1,X2,X3,X4	1.69	0.76

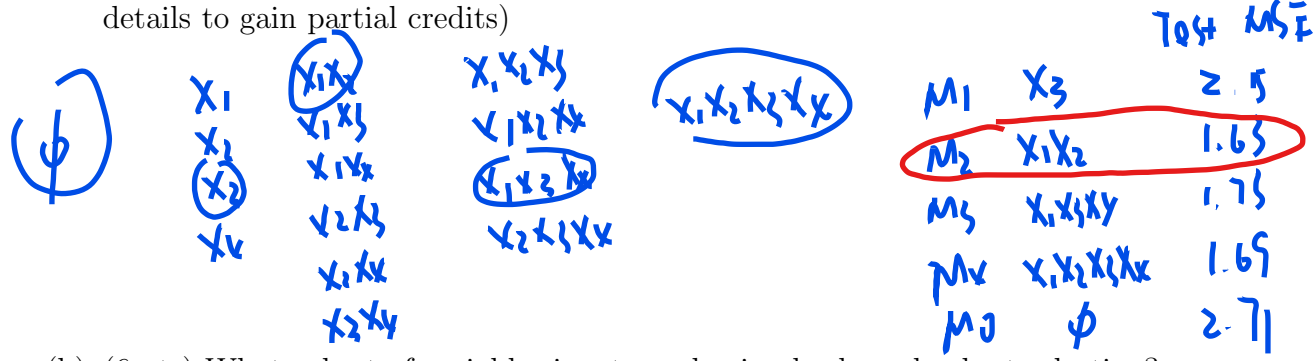
Larger R^2 is better

If Part (b) has exactly the same mistake as part (a) (with no extra mistakes), then part b only loses 1 point

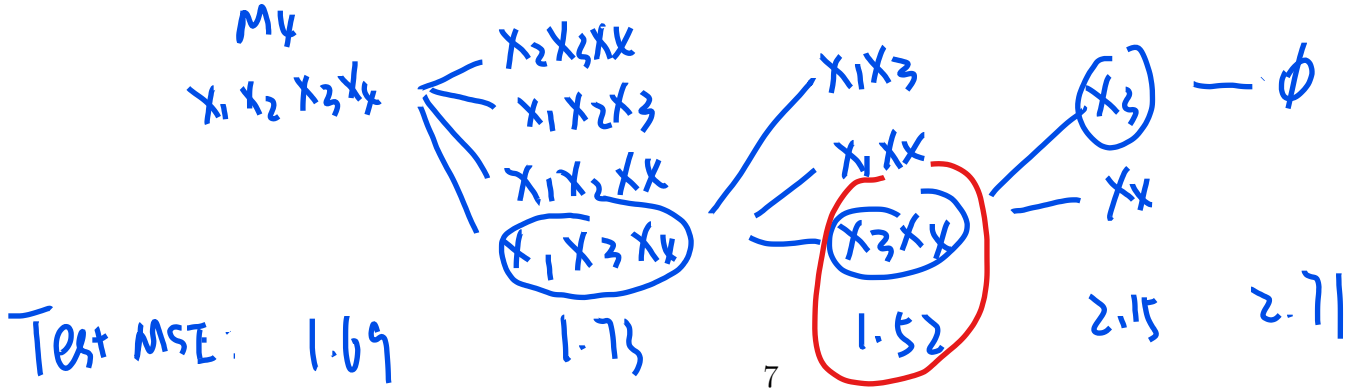
For part (a) and (b), correctly finding each M1-M4: 1pt (totally 4pts), final answer: 2pts

Used the wrong column to find M1-M4 or to find the final answer -3pts
Used smallest R^2 instead of largest R^2, -2pts

(a) (6 pts) What subset of variables is returned using best subset selection? (Provide details to gain partial credits)

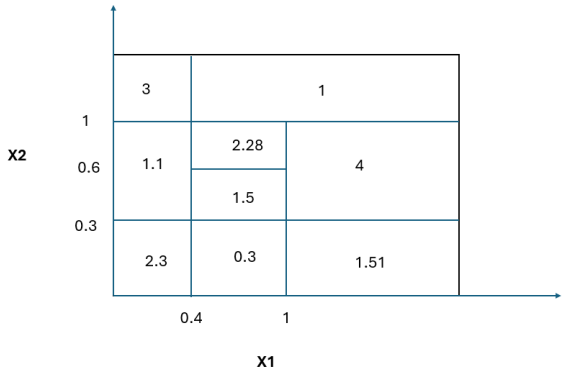


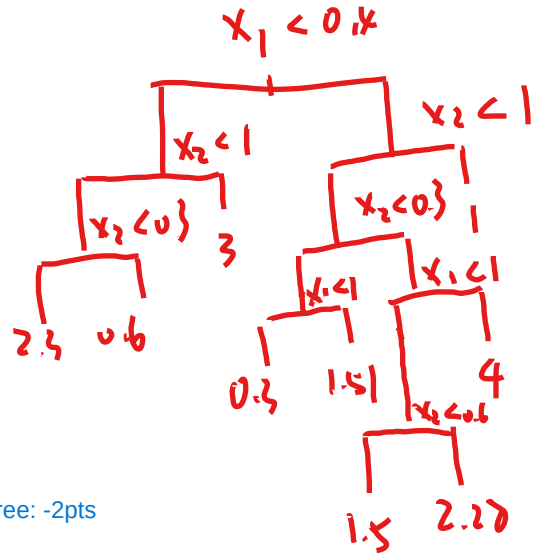
(b) (6 pts) What subset of variables is returned using backward subset selection?



7. (14 points)

(a) (7 pts) Find the decision tree corresponding to the partition of the predictor space illustrated in the figure below. The numbers inside the boxes indicate the mean of Y within each region.





Solution not unique, each mistake in the first two levels of the tree: -2pts
each mistake in the third level or up: -1pt

(b) (7 pts) Which of the following classification tree is preferred during pruning with an $\alpha = 0.1$?
Tree 1: two leaves (L_1,L_2), training samples classified to L_1 have labels $(-1, 1, 1)$, those classified to L_2 have labels $(-1, 1, 1, 1)$
Tree 2: one leaf L_1, all the above training samples go to this leaf.

Tree 1 $G_1 = \frac{2}{3}\cdot\frac{1}{3} + \frac{1}{3}\cdot\frac{2}{3} = \frac{4}{9}$ $G_2 = \frac{1}{4}\cdot\frac{3}{4}\times 2 = \frac{3}{8}$

3 pts for each tree,
Tree 1: 1pt for G1, 1pt for G2,
1pt for alpha|T|,
Tree2: 2 pts for G, 1pts for alpha|T|

1pt for final pick

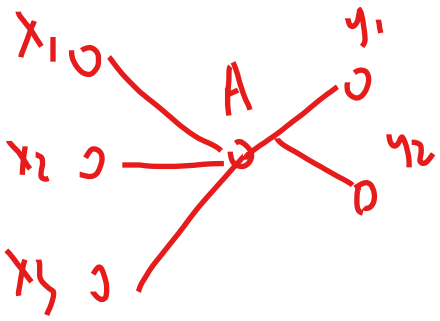
$Obj = G_1 + G_2 + \alpha|T| = \frac{4}{9} + \frac{3}{8} + 0.1\cdot 2 > 1$

Tree 2: $G = \frac{2}{7}\cdot\frac{5}{7}\cdot 2$ $Obj = G + 0.1\cdot 1 \approx 0.5$

So Tree 2 is preferred

8. (12 points)

(a) (4 pts) Draw a sketch of a neural network taking as input data points of the form (X_1, X_2, X_3) (so $p = 3$) with one hidden layer containing 1 hidden unit and with an output layer containing 2 units.



(b) (8 pts) Assume we the trained weights β and $\beta^{(2)}$ for the first and second layer

$$\beta = \begin{pmatrix} 1 & 1 & -2 \end{pmatrix} \qquad \beta^{(2)} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$$

and the bias of the first and second layer are 1 and $(1, -2)$. The activation function is Negative ReLU

$$g(z) = (z)_- = \begin{cases} 0 & \text{if } z > 0 \\ z & \text{else.} \end{cases}$$

Predict the label of the new data $(1, -1, 1)$.

- correctly computing A1: 4pts, y1,y2 each 2 pts
- didn't or incorrect use of activation -1.5pts,
- didn't or incorrect use of bias -1.5pts

$$g\left(\beta x^{\vec{1}} + b_1\right)\beta^{(2)} + b_2$$

$$= g\left(\begin{pmatrix} 1 & 1 & -2 \end{pmatrix}\begin{pmatrix} 1 \\ -1 \\ 1 \end{pmatrix} + 1\right)\begin{pmatrix} 0 \\ 1 \end{pmatrix} + \begin{pmatrix} 1 \\ -2 \end{pmatrix}$$

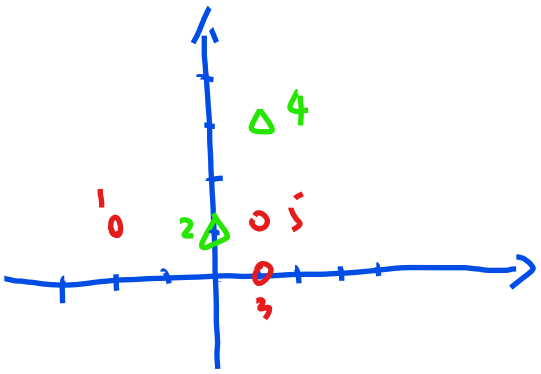
$$= g(-1)\cdot\begin{pmatrix} 0 \\ 1 \end{pmatrix} + \begin{pmatrix} 1 \\ -2 \end{pmatrix} = \begin{pmatrix} 1 \leftarrow y_1 \\ -3 \leftarrow y_2 \end{pmatrix}$$

$A_1 = -1$

9. (12 points) The table below provides a training data set containing five observations, three predictors, and one qualitative response variable.

Obs.	X_1	X_2	Y
1	-2	1	Red
2	0	1	Green
3	1	0	Red
4	1	3	Green
5	1	1	Red

(a) (4 pts) If we use k-nearest neighbors classification with $k = 3$, what is the prediction for $X_1 = X_2 = 0$?



For point (0,0), the 3 nearest neighbors are (obs 2, obs 3, obs 5)
G R R

by majority voting, the prediction is Red

1 point for the final answer, 3 pts for the correct argument/reasoning

(b) (8 pts) By hand, apply LOOCV to this dataset with $k = 1$. Provide details and compute the final LOOCV score.

LOOCV: for i = 1:5,
leave the ith sample out for test
compute $1\{y_i \neq \hat{y}_i\}$
compute the average LOOCV score $= \frac{1}{5}\sum_{i=1}^{5} 1\{y_i \neq \hat{y}_i\}$

2pts for providing this formula

1 point for each split

1 point for the final LOOCV score

Split	Obs left out	True label	3-(lowest Neighbors	predicted	
1	1	R	2,3,5	R	✓
2	2	G	1,3,5	R	✗
3	3	R	2,4,5	G	✗
4	4	G	2,3,5	R	✗
5	5	R	2,3,4	G	✗

LOOCV score = 4/5

Scrap Paper