Michigan State University
CMSE381 - Data Science

CMSE381 - Midterm 2 Sample

1. Do not open this test booklet until you are directed to do so.
2. You will have 2 hours to complete the exam. If you finish early go back and check your work.
3. This exam is open book. But 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: _____

Stuff to cover:

- Ch 5 Resampling / CV+bootstrap
- Ch 6 Subset selection
- Ch 6 Ridge and Lasso
- Ch 6 Dimension reduction / PCA (Really PCA + Regression = PCR)
- Ch 6 Partial Least Squares (meh?)
- Ch 7 Polynomial and step functions
- Ch 7 Basis functions
- Ch 7 Regression splines
- Ch 8 Decision trees
- Ch 8 Random Forests

1. (a) Which of the following are an approximation of the testing error? Circle all that apply.

- Bootstrap estimate
- AIC
- R^2
- LOOCV estimate
- BIC
- Total sum of squares
- Residual sum of squares
- k-fold CV estimate

(b) There are times where we can report the training error for a given model as a reasonable evaluation of the model quality.

`True` `False`

(c) Boostrapping requires that I can repeatedly simulate new data sets to evaluate the variability of a statistical estimator.

`True` `False`

(d) PCR is a form of subset selection since models learned involve only a subset of the original input variables.

`True` `False`

(e) What is the definition of scale equivariant?

If we multiply a X_j by $c \to cX_j$,
then it only results in β_j being divided by $c \to \beta_j/c$

(f) Which of the following are scale equivariant?

`Least squares` `Ridge regression` `The lasso`

2. (a) Which of the following equations are we using for optimization when we are using the lasso?

$$RSS \qquad RSS + \lambda \sum_{i=1}^{p} \beta_i^2 \qquad RSS + \lambda \sum_{i=1}^{p} |\beta_i|$$

(b) What model are we learning if $\lambda = 0$?

min RSS → least sol
square

(c) What model are we learning if $\lambda \to \infty$?

$\beta_j = 0, \forall j$ Null model (no predictors are used)

(d) The following two models were learning on a data set predicting salaries of baseball players. One was learned using ridge regression and one using the lasso. Which is which? Why?

Model A:

(Intercept)	AtBat	Hits	HmRun	Runs
1.27	-0.05	2.18	0.00	0.00
RBI	**Walks**	**Years**	**CAtBat**	**CHits**
0.00	2.29	-0.34	0.00	0.00
CHmRun	**CRuns**	**CRBI**	**CWalks**	**LeagueN**
0.03	0.22	0.42	0.00	20.29
DivisionW	**PutOuts**	**Assists**	**Errors**	**NewLeagueN**
-116.17	0.24	0.00	-0.86	0.00

Lasso

Model B:

(Intercept)	AtBat	Hits	HmRun	Runs
48.766	-0.358	1.969	-1.278	1.146
RBI	**Walks**	**Years**	**CAtBat**	**CHits**
0.804	2.716	-6.218	0.005	0.106
CHmRun	**CRuns**	**CRBI**	**CWalks**	**LeagueN**
0.624	0.221	0.219	-0.150	45.926
DivisionW	**PutOuts**	**Assists**	**Errors**	**NewLeagueN**
-118.201	0.250	0.122	-3.279	-9.497

Ridge

3. 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 training and testing error computed for the model learned using each possible subset of variables.

	Training MSE (x10^7)	k-fold CV Testing Error
Null model	8.76	10.08
X1	8.63	9.98
X2	7.42	8.01
X3	8.16	8.3
X4	8.33	9.06
X1,X2	4.33	7.47
X1,X3	5.82	5.22
X1,X4	3.17	4.23
X2,X3	4.07	3.78
X2,X4	3.31	4.01
X3,X4	3.06	4.16
X1,X2,X3	3.08	5.49
X1,X2,X4	3.55	4.02
X1,X3,X4	2.97	4.23
X2,X3,X4	2.98	3.17
X1,X2,X3,X4	2.16	4.39

(a) What subset of variables is found for each of the sets $\mathcal{M}_0, \mathcal{M}_1, \mathcal{M}_2, \mathcal{M}_3, \mathcal{M}_4$ when using best subset selection?

(b) What subset of variables is returned using best subset selection?

(c) What subset of variables is found for each of the sets $\mathcal{M}_0, \mathcal{M}_1, \mathcal{M}_2, \mathcal{M}_3, \mathcal{M}_4$ when using forward selection?

(d) What subset of variables is returned using forward subset selection?

4. We perform best subset, forward stepwise, and backward stepwise selection on a single data set. For each approach, we obtain $p+1$ models $\mathcal{M}_0, \cdots, \mathcal{M}_p$ where $\mathcal{M}_k$ contains k predictors.

   (a) Which of the three models (best/forward/backward) with k predictors has the smallest training RSS? Why?

   (b) The predictors in $\mathcal{M}_k$ identified by forward stepwise are *always* a subset of the predictors in the $\mathcal{M}_{k+1}$ identified by forward stepwise selection.

   True False

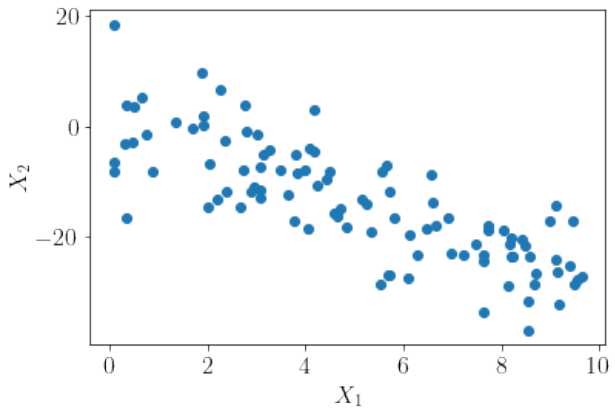
   (c) The predictors in $\mathcal{M}_k$ identified by best subset are *always* a subset of the predictors in the $\mathcal{M}_{k+1}$ identified by best subset selection.

   True False

5. (a) In doing dimension reduction such as PCA, we have input variables $X_1, \cdots, X_p$ and we want to construct new predictors $Z_1, \cdots, Z_M$ that are linear combinations of the X_i's. What are we aiming for in terms of M and p?

A. $p < M$ B. $p = M$ C. $M < p$

(b) Sketch the lines that would be found for the (i) first PC and (ii) second PC using PCA given the following data set where the axes are both input variables, X_1 and X_2.



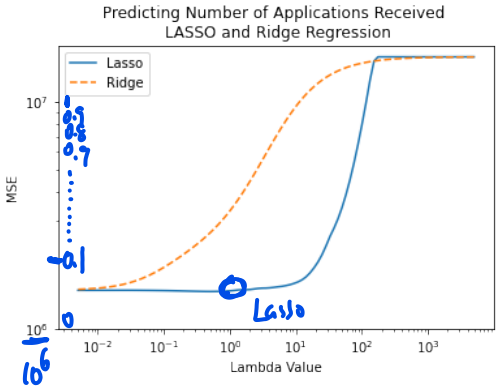
(c) Assume we have three input variables X_1, X_2, X_3 and two PCs,

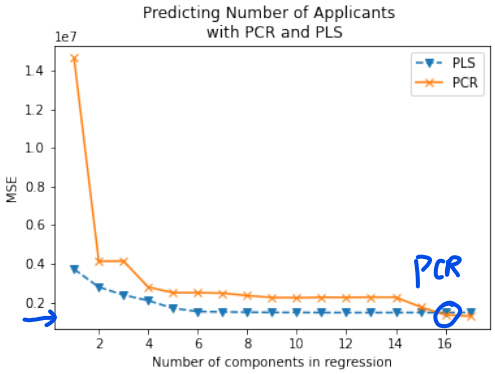
$$Z_1 = 0.266 \cdot X_1 - 0.077 \cdot X_2 + 0.961 \cdot X_3$$
$$Z_2 = 0.968 \cdot X_1 + 0.136 \cdot X_2 + -0.254 \cdot X_3.$$

Using linear regression, we learn the model $Y = -1 + 5Z_1 - 6Z_2$. What are the coefficients for the model learned in terms of the X_i's?

$$Y = -1 + (5 \cdot 0.266 - 6 \cdot 0.968) X_1 + (5 \cdot 1 - 0.077 - 6 \cdot 0.136) X_2 + (5 \cdot 0.961 - 6 \cdot (-0.256)) X_3$$

6. You've just spent a while working with the `College` data set predicting the number of applications recieved using the remaining variables in the data set. You've set up four different regression models, and ran K-fold cross validation to see how the choice of parameters in each affect the results of the analysis. The figures you generated are below.





Which of the four models (LASSO, Ridge, PCR, and PLS) would you choose? What parameters would you use? Why? Be sure to justify your choices.

7. Consider the equation

$$f(x) = \begin{cases} \beta_{01} + \beta_{11}x + \beta_{21}x^2 + \beta_{31}x^3 & \text{if } x < c \\ \beta_{02} + \beta_{12}x + \beta_{22}x^2 + \beta_{32}x^3 & \text{if } x \geq c \end{cases}$$

(a) What needs to be true for $f(x)$ to be a cubic spline? Be sure to list all requirements.

(b) Is the following possibly a cubic spline? Why or why not?

$$f(x) = \begin{cases} a - 2x + 3x^3 & \text{if } x < 1 \\ -1 - x + 5x^2 - x^3 & \text{if } x \geq 1 \end{cases}$$

$f(1^-) = f(1^+) \Rightarrow a - 2 + 3 = -1 - 1 + 5 - 1 \Rightarrow a = 1$

$\rightarrow f'(1^-) = f'(1^+) \Rightarrow -2 + 9x^2\big|_{x=1} = -1 + 10x - 3x^2\big|_{x=1}$

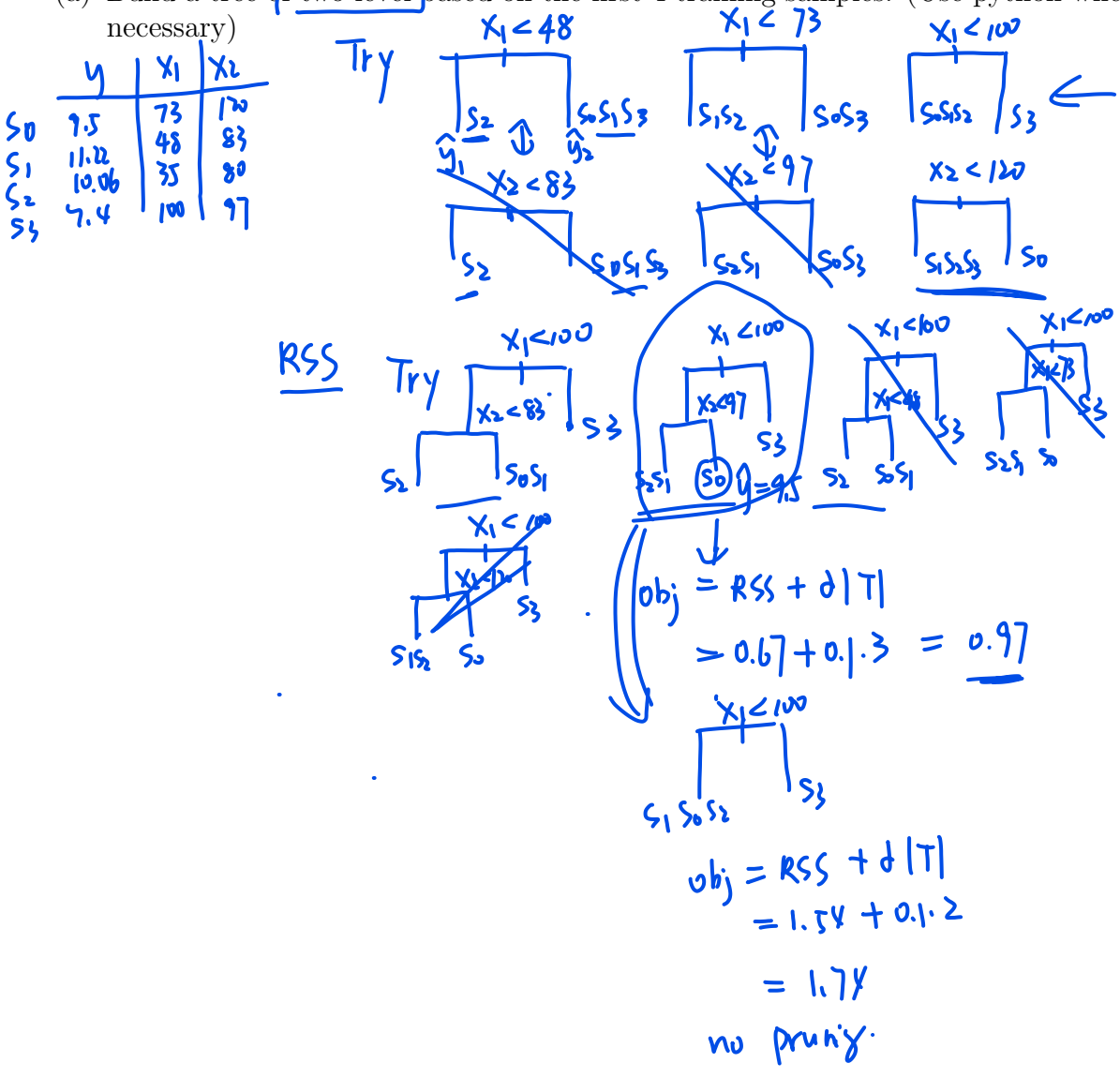
$\Leftrightarrow -2 + 9 = -1 + 10 - 3$

$\Leftrightarrow 7 = 6$ X

8. We built a regression tree using the carseat data set by predicting `Sales` using the Income and Price variables. The training data set is shown below.

	Sales	CompPrice	Income	Advertising	Population	Price	ShelveLoc	Age	Education	Urban	US
0	9.50	138	73	11	276	120	0	42	17	1	1
1	11.22	111	48	16	260	83	1	65	10	1	1
2	10.06	113	35	10	269	80	2	59	12	1	1
3	7.40	117	100	4	466	97	2	55	14	1	1
4	4.15	141	64	3	340	128	0	38	13	1	0

(a) Build a tree of two level based on the first 4 training samples. (Use python when necessary)



(b) prune the tree with $\alpha = .1$.

(c) What is the predicted sales using this tree for the last data point (labeled as row 4 above)? Be sure to explain why you got that answer.

$\hat{y} = 9.5$