Michigan State University
CMSE381 - Data Science

CMSE381 - Midterm 2

1. Do not open this test booklet until you are directed to do so.
2. You will have 80 mins 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: _____

1. (27 points, 3pts each)

   (a) While you can always standardize predictors X_i prior to learning your model, which of the following require standardization. Circle all that apply.

   - Least Squares Regression
   - Lasso
   - Ridge
   - PCR

   (b) Which of the following methods require shuffling the data.Circle all that apply.

   - LOOCV
   - bootstrapping
   - K-folder
   - bagging

   (c) In Lasso, as λ, which is the coefficient in front of the L1 penalty, increases from 0 to infinity, which of the following would happen. Circle all that apply.

   - The training MSE will decrease
   - The testing MSE will first in decrease then increase
   - The variance will decrease
   - The bias will increase

   (d) In PCR, assume we found the first PC as $Z_1 = \alpha_1 X_1 + \alpha_2 X_2$, which of the following is true about the coefficient vector $[\alpha_1, \alpha_2]$. Circle all that apply.

   - It is a unit-norm vector
   - It is called the first principal component
   - It is called a principal component score
   - It is pointing towards the maximal variance direction of the training data

   (e) Each bootstrapping dataset contains the same number of samples as the original training dataset.

   True False

   (f) The main difference between PCR and PLS is how the linear regression step is carried out.

   True False

   (g) In simple linear regression $y \sim X$, the residual $y - \hat{y}$ is the part of information in y that cannot be explained by X.

   True False

   (h) Indicator functions are discontinuous, so we cannot use them to approximate continuous functions.

   True False

(i) In Lasso regression, the coefficient β_0 is rarely shrunk to 0.

`True` `False`

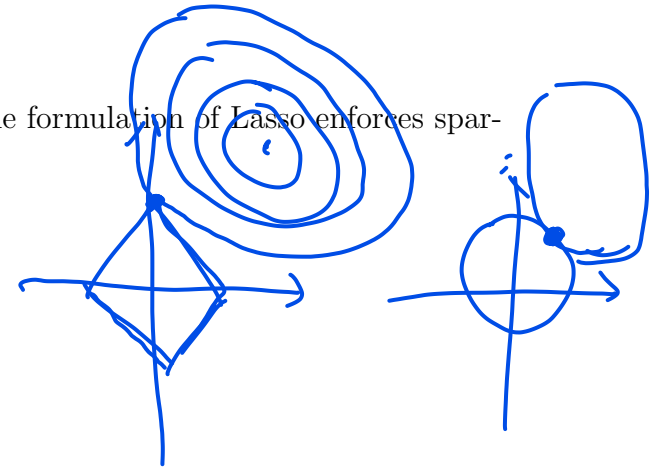
2. (10 points)

(a) (4 points) Which of the following objective can enforce sparsity of the coefficients?

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

(b) (6 points) Provide an explanation of why the formulation of Lasso enforces sparsity in the coefficients.

$$\hat{\beta}_i = \begin{cases} y_i - \lambda/2 & \text{if } y_i > \lambda/2 \\ y_i + \lambda/2 & \text{if } y_i < -\lambda/2 \\ 0 & |y_i| \leq \lambda/2 \end{cases}$$



3. (10 points) Assume my fitted model for the credit card dataset looks like this $y = 1 + 2b_1 + b_2 - 3b_3$ where $b_1 = -x1_{\{1 \leq x \leq 3\}}$, $b_2 = 1_{\{2 \leq x \leq 3\}}$, $b_3 = 1_{\{-1 \leq x \leq 1\}}$.

   (a) (4 points) sketch this function on the interval $[-2, 4]$.

   (b) (6 points) Suppose my friend came up with another model $y = 1 - b_1 + 3b_3$. Based on the data I have below, which model do you think is better? Why?

y	x
1	1
3	2
4	1

4. (11 points)

   (a) (8 points) In the best subset selection method, the first step is to identify the best subset $\mathcal{M}_0, \mathcal{M}_1, \mathcal{M}_2, \mathcal{M}_3, ..., \mathcal{M}_p$ among all subsets with a fixed cardinality. Dr. Wang made the following code to implement this step. In the code, auto is the name of the dataframe, y is the list holding the values of the response variable, inputvars is a list containing names of all the predictors, and p is the number of predictors. But there are two mistakes in the code, can you help to identify them?

```python
from itertools import combinations
def myscore_cv(df,X,y):
    model = LinearRegression()

    scores = cross_val_score(model, X,y,
                             scoring='neg_mean_squared_error',
                             cv=5)
    return np.average(np.absolute(scores))

Ms = []

for k in range(1,p):
    myvars = []
    myscores = []

    for Xs in combinations(inputvars,k):
        myvars.append(Xs)
        myscores.append(myscore_cv(auto,Xs,y))

    myResults = pd.DataFrame({'Vars':myvars, 'Score':myscores})


    indexmin = myResults.idxmin(numeric_only = True)
    Ms.append(myResults.Vars[indexmin].iloc[0])

```

(b) (3 points) Compared with the best subset selection method, what is an advantage and a disadvantage of the forward subset selection method?

faster

↓ not as accurate as best subset

5. (14 points) Consider the equation

$$f(x) = \begin{cases} \beta_{01} + \beta_{11}x + \beta_{21}x^2 + \beta_{31}x^3 & \text{if } x < 1 \\ \beta_{02} + \beta_{12}x + \beta_{22}x^2 + \beta_{32}x^3 & \text{if } 1 \le x \le 2 \\ \beta_{03} + \beta_{13}x + \beta_{23}x^2 + \beta_{33}x^3 & \text{if } x \ge 2 \end{cases}$$

(a) (4 points) How many knots are used in the above piecewise cubic polynomial?

2

(b) (4 points) Suppose $f(x)$ is cubic spline, how many degrees of freedom does it have? Why?

dof = No variable − No constraints
= 3·4 − 2·3
= 12 − 6 = 6

(c) (2 points) What needs to be true for $f(x)$ to be a cubic spline? Be sure to list all requirements.

$\lim_{x\to 1^-} f(x) = \lim_{x\to 1^+} f(x)$ $\Rightarrow$ $\beta_{01}+\beta_{11}+\beta_{21}+\beta_{31} = \beta_{02}+\beta_{12}+\beta_{22}+\beta_{32}$ $\Leftarrow$

$\beta_{11} + 2\beta_{21} + 3\beta_{31} = \beta_{12} + 2\beta_{22} + 3\beta_{32}$

(d) (4 points) Describe two ways to determine the placement of knots given input data $\{x_i, y_i\}_{i=1}^n$.

① equal division of $[\min\{x_i\}, \max\{x_i\}]$

② use percentile to ensure each subinterval has the same amount of training data.

6. (18 points) We want to build a regression tree using the carseat data set by predicting `Sales` using the CompPrice and Education 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) (8 points) Build a tree of two level based on the first 4 training samples. (Use pictures, codes or anything that helps. When submit, please include a photo of the code if any)

(b) (6 points) prune the tree with $\alpha = .3$.

(c) (4 points) 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.

7. (10 points)

(a) (6 points) Explain how a random forest is generated given input data points $\{x_i, y_i\}_{i=1}^n$. Use pictures, pseudocode, words, anything that helps.

Generate n bootstrapping dataset $S_1 \cdots S_n$.

for $i = 1 : n$:

randomly draw a subset of r predictors

use them and S_i to build a decision tree

return n decision trees

(b) (4 points) How does a random forest predict a value for a given data point?

Regression: $\frac{1}{n}\sum_{i=1}^{n}\hat{f}^i(x)$

Classification: majority voting among $\{\hat{f}^i(x)\}$

$\hat{f}^i$: i-th tree

x : new observation.