Name:

COM S 474/574 Fall 2021
Introduction to Machine Learning

Exam I

- You have 75 minutes to complete this exam. If you open the test before stated or do not turn it in on time you will lose 20 points.
- There are four questions, with a total of 100 points.
- This exam is closed book and notes. You will receive a zero for this exam for using books, notes, or any electronic devices. You will also be reported and appropriate disciplinary action will be taken.
- Your answers must be legible. Circle, underline, or leave sufficient white-space to distinguish your answers from intermediate work.
- Show all your work.
- Write your initials on each sheet used. If you use extra sheets for work, write your name on those.
- Do not write along the edge of the paper. There are blank pages you can use for scratch work or overflow.

Grade:

1. [16] _____
2. [40] _____
3. [24] _____
4. [20] _____

Total: _____

Problem 1. [16 points]

Your friend Ariana G. needs your help. She has a data set with three features $\{A, B, C\}$ she's using to predict the popularity Y of her songs. She ran best subset (aka 'exhaustive') search, forward search, and backward search. She saved the features that each search method selected for each cardinality, but cannot remember which search method produced which results.

Cardinality	Search 1	Search 2	Search 3
0	$\emptyset$	$\emptyset$	$\emptyset$
1	B	C	C
2	A, B	B, C	A, B
3	A, B, C	A, B, C	A, B, C

Identify which search methods produced which results and briefly explain how you can tell.

(extra space)

Problem 2. [40 points]

A. Your friend Vin D. tells you that he wants to predict muscle size Y using push-up counts X. He has a data set with $n = 10$ samples and is thinking of fitting a 7th order polynomial of X. Briefly explain why that is a bad idea.

B. A friend in medical school wants to predict blood pressure Y using features collected about people's diets (consumption of sugar, salt, trans fats, etc.). Your friend wants to fit a model using a large data set that was collected during 1910-1920. Briefly explain why that is a bad idea.

C. Your old friend Jerome P. told you that he fit a linear model of many economic features to predict gross domestic product (GDP). His data set had $n = 100$ samples. He fit the model using all $n = 100$ samples. He tells you that the model did such a great job (such low MSE for those $n = 100$ samples) that he will use the model for important decisions. You tell him he should have left some data out for evaluation. He asks you "why shouldn't I use all my data to fit the model?" Briefly explain why.

D. An acquaintance, Jeff B., asks you to predict how much money he will make Y from a rocket ship flight based on the number of people who go up in the rocket X_1 and the rocket's weight X_2 (measured in grams). He tells you that he thinks both X_1 and X_2 are important features. He asks you to use lasso to fit the model. Briefly explain why it is a bad idea to fit a model using lasso directly to that data set.

Problem 3. [24 points]

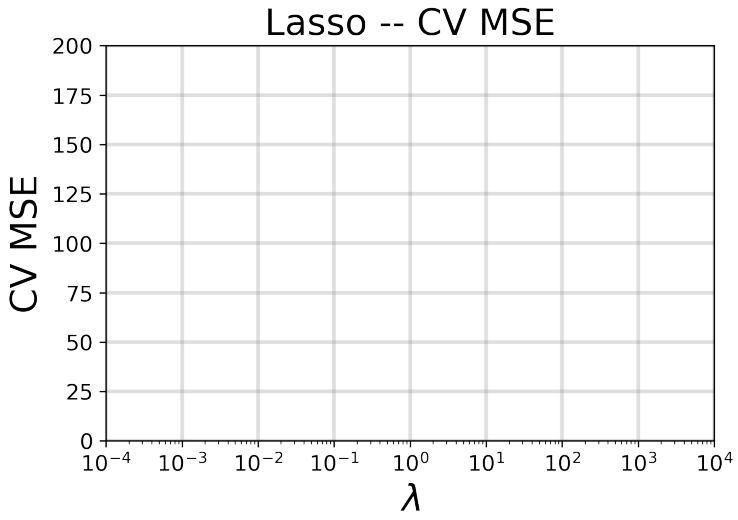
You fit a model to a data set using lasso. You forgot to print out plots for your COMS 474/574 homework, so you decide to make a sketch.

You wrote down the following information:

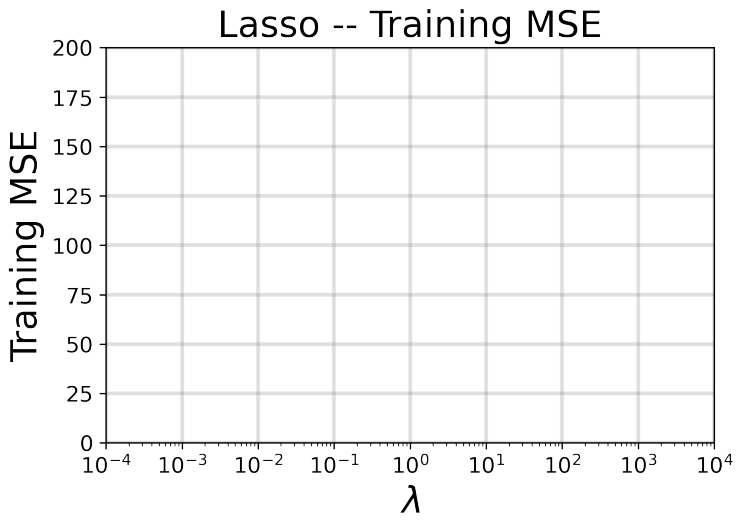
- The OLS model (i.e. all features) had a training MSE of 25 and a test MSE of 150.
- The best intercept (i.e. no features) had a training MSE of 100 and a test MSE of 125.
- The empirically best parameter was $\lambda = 10$. The final lasso model using $\lambda = 10$ had a training MSE of 50 and a test MSE of 75.

Draw a plausible curve for lasso's cross-validated MSE and for training MSE.

Briefly explain why you drew the curves the way you did.



Brief explanation:



Brief explanation:

Problem 4. [20 points]

Recall that "centering" a feature means subtracting its mean. For example, if the sample values for feature X_4 are $\begin{bmatrix} 5 \\ 0 \\ 1 \end{bmatrix}$, which has a mean of 2, we could replace it with $\begin{bmatrix} 5-2 \\ 0-2 \\ 1-2 \end{bmatrix} = \begin{bmatrix} 3 \\ -2 \\ -1 \end{bmatrix}$ which has a mean of 0. Thus, if feature X_4 is centered, then $\sum_{i=1}^{n} X_4(i) = 0$.

What is the value of the intercept β_0^* in the ordinary least squares solution, i.e.

$$(\beta_0^*, \beta_1^*, \ldots, \beta_p^*) = \underset{(\beta_0,\beta_1,\ldots,\beta_p)\in\mathbb{R}^{p+1}}{\arg\min} \quad \frac{1}{n}\sum_{i=1}^{n}\left(Y(i) - \beta_0 - \sum_{j=1}^{p}\beta_j X_j(i)\right)^2$$

when the features $\{X_1, \ldots, X_p\}$ are all centered? (e.g. $\sum_{i=1}^{n} X_j(i) = 0$ for $j = 1, \ldots, p$.) (You do not need to use a second derivative test or solve for $\{\beta_1^*, \ldots, \beta_p^*\}$).

(extra space)

(extra space)

(extra space)