Name: ____

COM S 474/574 Fall 2021
Introduction to Machine Learning

Exam I – Solutions

- 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.

- search 1 = backward
- search 2 = forward
- search 3 = best subset

We see that the subsets in search 3 are not nested, while they are nested for searches 1 and 2. We also know that forward and backward searches are nested; best subset is not necessarily nested. Thus, search 3 must be best subset search.

That tells us that C is the best feature individually (since best subset chose it for cardinality 1); forward search always starts with the individually best, so search 2 must be forward search (since search 1 has another feature for cardinality 1).

Lastly, search 1 must be backward. We can confirm this since for cardinality $p - 1 = 3 - 1 = 2$ it chooses the best overall subset, $\{A, B\}$, consistent with best subset's choice.

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.

Higher order polynomials are flexible models (7th order has 8 coefficients), and that is a very small data set. There is a high risk of overfitting (that the model fit will fit the underlying model plus the noise).

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.

Diets have changed a lot over the past century. Thus, the data collected in 1910-1920 will likely not be identically distributed with data collected today.

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.

Training MSE is an optimistic measure of performance, in part due to the model being able to adjust to the noise in the data. In extreme cases of overfitting, training error can go down to 0 while performance on future data is bad. The noise in held out data is independent of the data fit with, so MSE on held out data is a better estimate of performance on future data.

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.

The features are on vastly different scales. X_1 probably has a range of $0 - 10$ while X_2 might be in a range of $10^5 - 10^8$ or some huge values. In the OLS solution, coefficients will scale inversely to compensate for the feature units (so β_2 would probably be tiny in the OLS solution). That is ok for OLS, but lasso penalizes large coefficients, so it would penalize β_1 more severely simply because the X_1 values are smaller. It would be better to standardize, so both features have the same standard deviation.

Problem 3. [24 points]

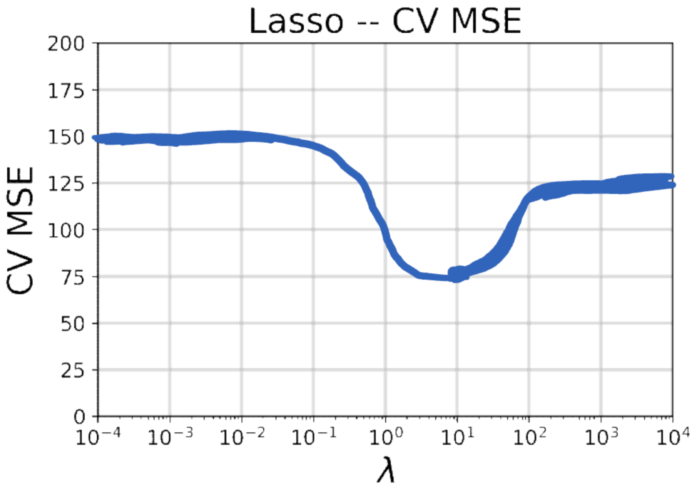
You fit a model to a data set using lasso. You forgot to print out plots for your COMS 474/574 homework, so you decide to make a sketch.

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 hold-out MSE and for the cross-validated training MSE.

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

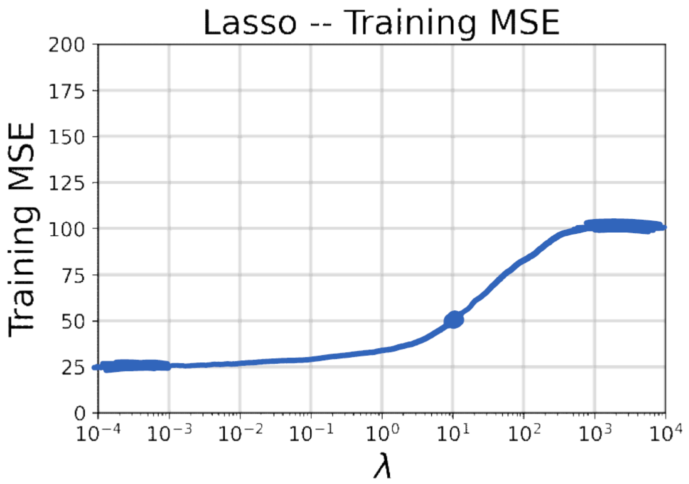


Brief explanation:

We know that as $\lambda \to 0$, the penalty is removed, so we will get the OLS solution. Cross-validated (hold-out) MSE approximates test error, so we expect the CV-MSE for $\lambda \approx 10^{-4}$ to be about 150 (the OLS test MSE)

Likewise, as $\lambda \to \infty$, the penalty dominates and will drive the model to the best intercept. Thus, we expect the CV-MSE for $\lambda \approx 10^{4}$ to be about 125 (the best intercept test MSE)

Lastly, we know the best λ, in terms of CV-MSE, is $\lambda = 10$. That CV-MSE should be approximately the test MSE of the final model (using all training data and with $\lambda = 10$), so it should be about 75.



Brief explanation:

By the same reasoning as the previous plot, we expect

- the cross-validated training MSE for $\lambda \approx 10^{-4}$ to be about 25 (the OLS training MSE)
- the cross-validated training MSE for $\lambda \approx 10^{4}$ to be about 125 (the best intercept training MSE)
- the cross-validated training MSE for $\lambda = 10$ model should be approximately the training MSE of the final model 50.

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 $\{\beta_1^*, \ldots, \beta_p^*\}$).

We can use the first derivative test for $\frac{\partial}{\partial\beta_0}$ first (and if that's insufficient, try others).

$$\begin{aligned}0 &= \frac{\partial}{\partial\beta_0}\,\frac{1}{n}\sum_{i=1}^{n}\left(Y(i) - \beta_0 - \sum_{j=1}^{p}\beta_j X_j(i)\right)^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}2\left(Y(i) - \beta_0 - \sum_{j=1}^{p}\beta_j X_j(i)\right)(-1)\end{aligned}$$

Dividing both sides by $\frac{-2}{n}$,

$$0 = \sum_{i=1}^{n}\left(Y(i) - \beta_0 - \sum_{j=1}^{p}\beta_j X_j(i)\right)$$

Breaking up the sum and rearranging,

$$\begin{aligned}0 &= \left[\sum_{i=1}^{n}Y(i)\right] - \left[\sum_{i=1}^{n}\beta_0\right] - \left[\sum_{i=1}^{n}\sum_{j=1}^{p}\beta_j X_j(i)\right] \\ &= \left[\sum_{i=1}^{n}Y(i)\right] - n\beta_0 - \left[\sum_{j=1}^{p}\beta_j\left(\sum_{i=1}^{n}X_j(i)\right)\right] \\ &= \left[\sum_{i=1}^{n}Y(i)\right] - n\beta_0 - \left[\sum_{j=1}^{p}\beta_j\left(0\right)\right] && \text{(features are centered)} \\ &= \left[\sum_{i=1}^{n}Y(i)\right] - n\beta_0\end{aligned}$$

which means $\beta_0 = \frac{1}{n}\sum_{i=1}^{n} Y(i)$.