COM S 474/574 Spring 2022
Introduction to Machine Learning

Homework 4

1 Directions:

- **Due: Thursday February 24, 2022 at 9pm.** Late submissions will be accepted for 24 hours with a 15% penalty. (the enforcement is strict, beginning at 9:01pm, except for extreme situations; having a poor wifi connection or minor computer problems is not sufficient for the penalty to be waived.)
- Upload the homework to Canvas as a single pdf file.
- If the graders cannot easily read your submission (writing is illegible, image is too dark, or if the contrast is too low) then you might receive a zero or only partial credit.
- Any non-administrative questions must be asked in office hours or (if a brief response is sufficient) Piazza.

2 Problems

Problem 1. [24 points total (5,5,3,3,4,4)]

Suppose you use lasso to fit a linear model for a data set. Let $\beta^*(\lambda)$ denote the lasso solution for a specific λ (i.e. the coefficient vector you get for that λ).

Provide explanations for your answers to the following questions.

(a). Describe how the training MSE changes as a function of λ, including $\lambda = 0$ and as $\lambda \to \infty$.
$\lambda = 0$ means the penalty is removed and we have O.L.S. problem, which finds the linear model that achieves the smallest training MSE. As λ increases, it increases monotonically with λ. This happens since we are increasingly penalized for using large coefficients and will consequently use smaller (in L_1 norm) coefficient vectors, resulting in worse training MSE.

(b). Describe how the hold-out MSE changes as a function of λ, including $\lambda = 0$ and as $\lambda \to \infty$.
$\lambda = 0$ means the penalty is removed and we have O.L.S. problem, which finds the linear model that achieves the smallest training MSE, but likely over-fits. $\lambda = 0$ likely has high hold-out MSE.
As $\lambda \to \infty$, the penalty is high enough all coefficients (except β_0) to 0, resulting in the best constant model, likely under fitting, thus high hold-out MSE.
In between, there will typically be a "U" shape for the hold-out MSE plot, with a λ that performs better on hold-out data than the best constant or O.L.S. models.

(c). Describe $\beta^*(0)$.

$\lambda = 0$ means the penalty is removed and we have O.L.S. problem, so $\beta^*(0)$ will be the O.L.S. solution.

(d). Describe what happens to $\beta^*(\lambda)$ as λ grows.

As described above, for $\lambda \to \infty$, the $\beta^*(\lambda)_j \to 0$ for each feature X_j; for sufficiently large λ, $\beta^*(\lambda)_j = 0$ for each feature X_j and $\beta^*(\lambda)_0$ is equal to the best intercept (average values of Y's).

(e). If you used ridge regression instead of lasso, explain how your answers to (a).-(d). would differ.

Only (d). would change; as the coefficients would shrink towards 0 as λ grows but never equal 0.

(that is, the specific coefficients and location of λ^* would be different, but the general trends described above would be the same)

(f). We discussed the "constrained form" of lasso, with a constraint of the form

$$\sum_{i=1}^{p} |\beta_j| \leq t.$$

Which value (or limiting value) of t corresponds to $\lambda = 0$ and which corresponds to $\lambda \to \infty$.

$t = 0$ means all the coefficients (except β_0) are 0, which corresponds to $\lambda \to \infty$. As $t \to \infty$, eventually the O.L.S. solution will be feasible, which corresponds to $\lambda = 0$.

Problem 2. [15 points]

You have already seen formulas for the best intercept in linear models when there are no features $p = 0$ and a single feature $p = 1$. You will now look at what happens with p features when we center the data.

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 $\{\beta_1^*, \ldots, \beta_p^*\}$, just use the first derivative test $0 = \frac{\partial}{\partial \beta_0}$ MSE)

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

The following question is only for 574 students and should be submitted separately on canvas.

Problem 3. [15 points total — just for 574 students]

Recall the general ridge regression problem,

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

Suppose features X_1 and X_2 are perfectly (linearly) correlated, with $X_1(i) = aX_2(i)$ for all samples $i \in \{1, \ldots, n\}$, where a is some constant. (For example, $a = 1$ corresponds to the two features being exactly the same.)

Using calculus, identify how their coefficients β_1^* and β_2^* are related. (You only need to look at first (partial) derivatives with respect to β_1 and β_2; no second derivative tests are needed for this problem; you do not need to solve for β_1^* and β_2^*, just show how they are related.)

The optimal solution satisfies

$$\begin{aligned} 0 &= \frac{\partial}{\partial \beta_1} \left[\frac{1}{n} \sum_{i=1}^{n} \left(Y(i) - \beta_0 - \sum_{j=1}^{p} \beta_j X_j(i) \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j|^2 \right] \\ &= \frac{1}{n} \sum_{i=1}^{n} 2 \left(Y(i) - \beta_0 - \sum_{j=1}^{p} \beta_j X_j(i) \right) \Big(-X_1(i) \Big) + 2\lambda\beta_1 \end{aligned} \tag{1}$$

Repeating with $\frac{\partial}{\partial \beta_2}$

$$\begin{aligned} 0 &= \frac{\partial}{\partial \beta_2} \left[\frac{1}{n} \sum_{i=1}^{n} \left(Y(i) - \beta_0 - \sum_{j=1}^{p} \beta_j X_j(i) \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j|^2 \right] \\ &= \frac{1}{n} \sum_{i=1}^{n} 2 \left(Y(i) - \beta_0 - \sum_{j=1}^{p} \beta_j X_j(i) \right) \Big(-X_2(i) \Big) + 2\lambda\beta_2 \end{aligned}$$

Using $X_1(i) = aX_2(i)$, we can express (1) as

$$\begin{aligned} 0 &= \frac{1}{n} \sum_{i=1}^{n} 2 \left(Y(i) - \beta_0 - \sum_{j=1}^{p} \beta_j X_j(i) \right) \Big(-X_1(i) \Big) + 2\lambda\beta_1 \\ &= \frac{1}{n} \sum_{i=1}^{n} 2 \left(Y(i) - \beta_0 - \sum_{j=1}^{p} \beta_j X_j(i) \right) \Big(-(aX_2(i)) \Big) + 2\lambda\beta_1 \\ &= a \left[\frac{1}{n} \sum_{i=1}^{n} 2 \left(Y(i) - \beta_0 - \sum_{j=1}^{p} \beta_j X_j(i) \right) \Big(-X_2(i) \Big) \right] + 2\lambda\beta_1 \end{aligned} \tag{2}$$

To simplify notation, set

$$\gamma = \left[\frac{1}{n} \sum_{i=1}^{n} 2 \left(Y(i) - \beta_0 - \sum_{j=1}^{p} \beta_j X_j(i) \right) \Big(-X_2(i) \Big) \right].$$

Then we have the system of equations

$$0 = \gamma + 2\lambda\beta_2 \tag{3}$$

$$0 = a\gamma + 2\lambda\beta_1. \tag{4}$$

Isolating γ in both equations and combining them,

$$2\lambda\beta_2 = \frac{2\lambda}{a}\beta_1$$

so $\beta_1^* = a\beta_2^*$.