COM S 474/574
Introduction to Machine Learning

Spring 2022

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 $\lambda \to \infty$.

The MSE increase as λ increase from 0 because it makes the model less flexible which leads to increment in training MSE.

(b). Describe how the hold-out MSE changes as a function of λ, including $\lambda = 0$ and as $\lambda \to \infty$.

The MSE initially decrease and eventually start increas as we saw in the class (2/15/2022)

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

OLS. No penalty on β's

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

We will only have β_0 to fit with and end up using hold-out data to select.

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

The ridge regression will result very similar to the results of the lasso, but it might have a bit larger MSE. But (c) and (d) will result very similar.

(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 $\lambda \to \infty$.

λ-> 0: large enough t (t->∞)
λ-> ∞: small t (t->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.

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

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)

$$\begin{aligned}
0 &= \frac{d}{d\beta_0} MSE \\
&= \frac{d}{d\beta_0} \left(\frac{1}{n} \sum_{i=1}^{n} (Y(i) - \beta_0)^2\right) \\
&= -\frac{1}{n} 2 \sum_{i=1}^{n} (Y(i) - \beta_0) \\
&= \sum_{i=1}^{n} (Y(i) - \beta_0)
\end{aligned}$$

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

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