COMS 474

Homework 4

Haadi Majeed

Spring 2022

Problem 1

[24 points total (5,5,3,3,4,4)]
Suppose you use lasso fit to a linear model for a data set. Let $\beta^*(\lambda)$ denote the lasso solution for a specific λ (i.e. the coeffcient vector you get for that λ).

Provide explanations for your answers to the following questions.

Part A

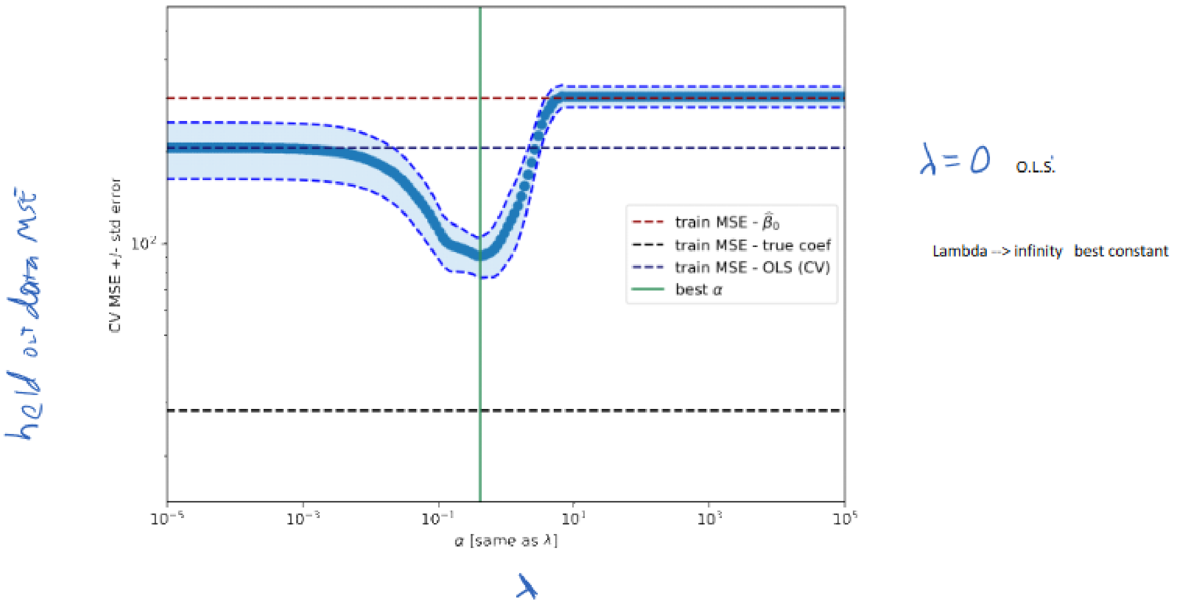
Describe how the training MSE changes as a function of λ, including $\lambda = 0$ and as $\lambda \to \infty$

The training MSE of the function $\beta^*(\lambda)$ would return the least squares fit when $\lambda = 0$. The value output does change though as λ grows in size, with the upper bound of ∞ being the most ideal constant that could be supplied to the equation, this is because at *some* point, λ will hit the absolute minimium of the function, which would be within the range of $0 \to \infty$.

Part B

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

The hold-out MSE *should* look similar to the training MSE and by applying a similar set of logic, the models should look relatively the same as it is from the same overall set of data. The resulting model should have an absolute minimum which would indicate what the most ideal λ would be. This value should be, if not exactly, then very close to where the training set's absolute minimium should be at. The image I have attached below is from the notes and displays what the models could look like as an example.



Part C

Describe $\beta^*(0)$.

When initially at $\beta^*(0)$, the function's value that it would return would simply be the L.S.F (least squares fit) value.

Part D

Describe what happens to $\beta^*(0)$ as λ grows.

The model $\beta^*(0)$ would cover every possible value between $0 \to \infty$ and would eventually hit the functions aboslute minimium in a concave region. This is where it deviates off the null model and no longer just returns zero, and insteal will indicate what the absolute minimium, or the most ideal λ, would be.

Part E

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

If we were to use ridge regression, we should expect a bit of variance in the data since they approach model slightly differently. Another attribute of ridge regression is that it is not capable to have sharp points in the model, thus intersections at an axis is not common. This subsequently results in non-zero estimations. The models generated by this when comparing the training vs the hold-out data should look very similar to each other due to data distribution. Ridge regression looks at both extremes and steps through the function looking at each step from $0 \to \infty$. It also penalises $\lambda \to \infty$ signifigantly more than it would $\lambda = 0$. When$\lambda = 0$ it would return the Least Squares estimation, while when $\lambda \to \infty$ it estimates the model will approach zero and the shrinkage penalty grows.

Part 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 coresponds to $\lambda = 0$ and which corresponds to $\lambda \to \infty$

The value of t would equate to $|\beta_1|^1 + |\beta_2|^1$ since t represents λ but in a constrained environment instead. The point at which the red oval from figure 6.7 from the book (attached below)

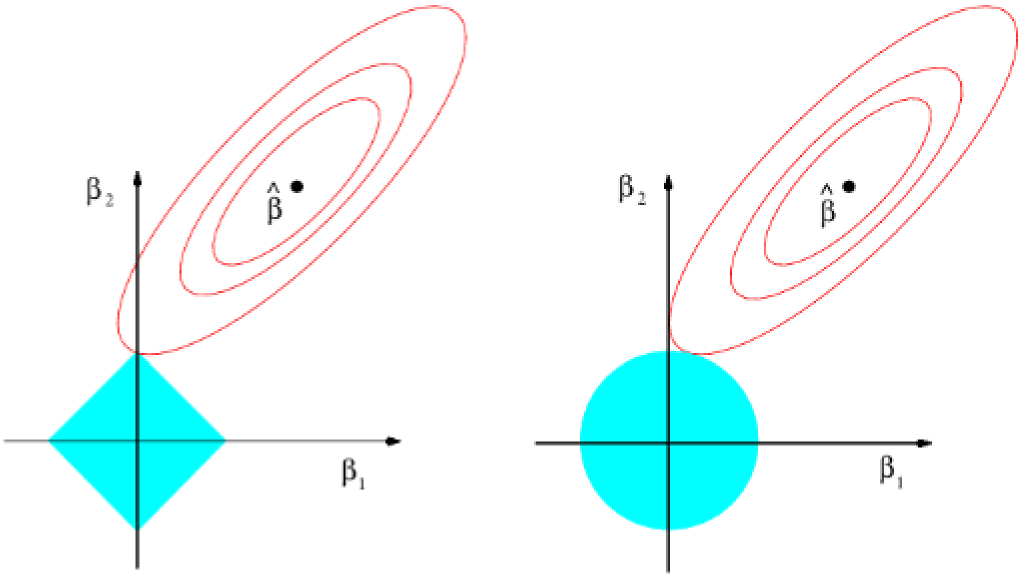


FIGURE 6.7. *Contours of the error and constraint functions for the lasso* (left) *and ridge regression* (right). *The solid blue areas are the constraint regions, $|\beta_1| + |\beta_2| \leq s$ and $\beta_1^2 + \beta_2^2 \leq s$, while the red ellipses are the contours of the RSS.*

Should $\lambda = 0$ then $t = |\beta_1|^1 + |\beta_2|^1$ since we are restricted by $\sum_{j=1}^{p} |\beta_j|$ having to be within the bound set by t and the expected output would quite small. For when $\lambda \to \infty$, then $t = 0$ since the solution of the Least Squares Fit since it would be encompassed by the range set by t.

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 centred, 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.

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

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).

Confusion on what to exactly do or if the correct thing was done, as when the first derivative is taken for the given equation we get

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

and if we go with the fact that $\sum_{i=1}^{n} X_j(i) = 0$ for $j = 1, \ldots, p$ then that would result in just $\frac{2}{n} \sum_{i=1}^{n} Y(i)$, but not sure if this is the correct steps to be taking or if completely off track.