Delivery 1

Daniel Salgado Rojo

Mathematics for Big Data. Course 2017/2018.

3) Suppose we estimate the regression coefficients in a linear regression model by minimizing

$$RSS = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \text{ subject to } \sum_{j=1}^{p} |\beta_j| \leq s \tag{1}$$

for a particular value of s. For parts (a) through (e), indicate which of i. through v. is correct. Justify your answer.

Comment: Note that (1) is equivalent to solve a Lasso regression problem for some $\lambda(s)$, i.e., minimizing

$$J := RSS + \lambda(s) \sum_{i=1}^{p} |\beta_i| \tag{2}$$

where λ decreases as s increases and vice-versa.

(a) As we increase s from 0, the *training RSS* will:

i. Increase initially, and the eventually start decreasing in a n inverted U shape.
ii. Decrease initially, and then eventually start increasing in a U shape.
iii. Steadily increase.
iv. Steadily decrease. ✓
v. Remain constant.

Solution. As one can deduce from the left-hand panel in Figure 1, as s increases, the training RSS under the restriction $\sum_{j=1}^{p} |\beta_j| \leq s$ will decrease since RSS is (negatively) quadratic and the feasible region of solutions for β approaches $\hat{\beta}$ (the least-squares solution) which we know that minimize the training RSS. Once $\hat{\beta}$ is inside the starts feasible region, the training RSS becomes constant and equal to the least-squares RSS [1].

[1] When we say training RSS we refer to the RSS obtained by solving (1) and the training data.

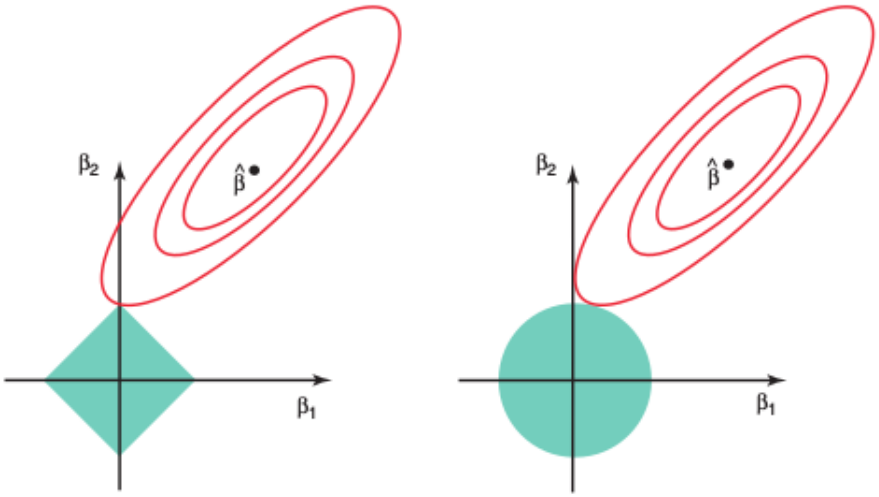

FIGURE 6.7. *Contours of the error and constraint functions for the lasso* (left) *and ridge regression* (right). *The solid blue areas are the constraint regions,* $|\beta_1| + |\beta_2| \leq s$ *and* $\beta_1^2 + \beta_2^2 \leq s$, *while the red ellipses are the contours of the RSS.*

Figure 1: From [1]

◀

[!] Although we do not give strict mathematical proofs for (b), (c) and (d), we try just to give intuitive justifications. In the notes [2], proofs for the same answers of (b), (c) and (d) but for the ridge regression estimator $\vec{\hat{\beta}}$.

(b) As we increase s from 0, the *test RSS* will:

i. Increase initially, and the eventually start decreasing in a n inverted U shape.
ii. Decrease initially, and then eventually start increasing in a U shape. ✓
iii. Steadily increase.
iv. Steadily decrease.
v. Remain constant.

Solution. For $s = 0$ all coefficients except from β_0 have to be zero so that for the test data $RSS = RSS + 0 = J = \sum_{i \in \text{test}} (y_i - \beta_0)^2 \geq 0$.

For s sufficiently small, it would hold that

$$\sum_{i \in \text{test}} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \leq \sum_{i \in \text{test}} (y_i - \beta_0)^2$$

and at some point, for s sufficiently large, the β coefficients will be approaching the optimal least squares solution and hence overfitting the training data, which would increase the test RSS,

$$\sum_{i \in \text{test}} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \geq \sum_{i \in \text{test}} (y_i - \beta_0)^2$$

Thus, we may expect a U shape for the test RSS when increasing s from 0.

◀

(c) As we increase s from 0, the *variance* will:

i. Increase initially, and the eventually start decreasing in an inverted U shape.
ii. Decrease initially, and then eventually start increasing in a U shape.
iii. Steadily increase. ✓
iv. Steadily decrease.
v. Remain constant.

Solution. In a linear regression context we have

$$\hat{f}(x_1, ..., x_p) = \beta_0 + \sum_{j=1}^{p} \beta_j x_j$$

for some test observation $(x_1^i, ..., x_p^i), y^i$, i = 1,...,M. Where the β_j have been obtained from least-squares subject to $\sum_{j=1}^{p} |\beta_j| \leq s$ and a certain training set. If $s = 0$ we have that

$$\hat{f}(x_1, ..., x_p) = \beta_0 = \frac{\sum_{i=1}^{n} y_i}{n}$$

and

$$Var[\hat{f}] = E\left[\left(\frac{\sum_{i=1}^{n} y_i}{n}\right)^2\right] - E\left[\frac{\sum_{i=1}^{n} y_i}{n}\right]^2 = 0$$

independently of the training set used to find the β_is (things that do not depend on "X" go outside the mean).

As we increase the value of $s > 0$, condition $\sum_{j=1}^{p} |\beta_j| \leq s$ becomes less restrictive allowing more possible combinations of values for the β_j. Thus, when computing the variance of $\hat{f}$ over all possible training sets, it will increase as more and more combinations of β_j and x_j are being used. At least we can say that at the beginning variance will increase, since it is positive and for $s = 0$ it starts at the zero value.

◀

(d) As we increase s from 0, the *(squared) bias* will:

i. Increase initially, and the eventually start decreasing in a n inverted U shape.
ii. Decrease initially, and then eventually start increasing in a U shape.
iii. Steadily increase.
iv. Steadily decrease. ✓
v. Remain constant.

Solution. As in (c), when $s = 0$ we are predicting any test point with a constant value β_0, and this results in having high squared bias. As the value of s increases, some β_i become nonzero so that the predictions are expected to approach the real values (at least, predictions on train data points approach to the train data points). In this way the squared bias is expected to decrease, and if the model is able to overfit the training set, then bias will be exactly zero (for the train set).

◀

(e) As we increase s from 0, the *irreducible error* will:

i. Increase initially, and the eventually start decreasing in a n inverted U shape.
ii. Decrease initially, and then eventually start increasing in a U shape.
iii. Steadily increase.
iv. Steadily decrease.
v. Remain constant. ✓

Solution. The irreducible error is a contribution which we can attribute to randomness or natural variability in the system, and in particular it does not depend the model. Since each value of s determines a (ridge) model, irreducible error does not change when s changes.

◀

5) It is well-known that ridge regression tends to give similar coefficient values to correlated variables, whereas the lasso may give quite different coefficient values to correlated variables. We will now explore this property in a very simple setting. Suppose that $n = 2$, $p = 2$, $x_{11} = x_{12}$, $x_{21} = x_{22}$. Furthermore, suppose that $y_1 + y_2 = 0$ and $x_{11} + x_{21} = 0$ and $x_{12} + x_{22} = 0$, so that the estimate for the intercept in a least squares, ridge regression, or lasso model is zero: $\hat{\beta}_0 = 0$.

(a) Write out the ridge regression optimization problem in this setting

Solution. The ridge regression problem aims to minimize

$$J_R(\beta_0, ..., \beta_p) = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \tag{3}$$

given a train data set.

Note that using the relations above one can deduce that $x_{11} = -x_{22}$. With all the information we have, the Ridge regression optimization problem can be simplified as follows:

$$\begin{aligned}
&\sum_{i=1}^{2} \left(y_i - \beta_0 - \sum_{j=1}^{2} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{2} \beta_j^2 \\
&\quad = (y_1 - \beta_0 - \beta_1 x_{11} - \beta_2 x_{12})^2 + (y_2 - \beta_0 - \beta_1 x_{21} - \beta_2 x_{22})^2 + \lambda(\beta_1^2 + \beta_2^2) \\
&\quad = (y_1 - \beta_0 - x_{11}(\beta_1 + \beta_2))^2 + (y_2 - \beta_0 - x_{22}(\beta_1 + \beta_2))^2 + \lambda(\beta_1^2 + \beta_2^2) \\
&\quad = (y_1 - x_{11}(\beta_1 + \beta_2))^2 + (-y_1 + x_{11}(\beta_1 + \beta_2))^2 + \lambda(\beta_1^2 + \beta_2^2) \\
&\quad = 2\,(y_1 - x_{11}(\beta_1 + \beta_2))^2 + \lambda(\beta_1^2 + \beta_2^2)
\end{aligned}$$

$$\begin{aligned}
J_R(\beta_0, ..., \beta_p) &= 2\,(y_1 - x_{11}(\beta_1 + \beta_2))^2 + \lambda(\beta_1^2 + \beta_2^2) \\
&= 2(y_1^2 + x_{11}^2(\beta_1 + \beta_2)^2 - 2y_1 x_{11}(\beta_1 + \beta_2)) + \lambda(\beta_1^2 + \beta_2^2)
\end{aligned}$$

◀

(b) Argue that in this setting, the ridge coefficient estimates satisfy $\hat{\beta}_1 = \hat{\beta}_2$.

Solution. To find $(\hat{\beta}_1, \hat{\beta}_2)$ solutions of (7), we derivate J_R with reapect to β_1, β_2 and impose them to be zero when evaluated in $(\hat{\beta}_1, \hat{\beta}_2)$:

$$\frac{\partial J_R}{\partial \beta_1} = 4x_{11}^2(\beta_1 + \beta_2) - 4y_1 x_{11} + 2\lambda\beta_1 = 0 \tag{4}$$

$$\frac{\partial J_R}{\partial \beta_2} = 4x_{11}^2(\beta_1 + \beta_2) - 4y_1 x_{11} + 2\lambda\beta_2 = 0 \tag{5}$$

Now if we consider (4) - (5) = 0 we obtain that $2\lambda(\hat{\beta}_1 - \hat{\beta}_2) = 0$, and the only possiblity that holds for any λ is the condition

$$\hat{\beta}_1 = \hat{\beta}_2 \tag{6}$$

◀

(c) Write out the lasso optimization problem in this setting.

Solution. The lasso regression problem aims to minimize

$$J_L(\beta_0, ..., \beta_p) = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \tag{7}$$

given a train data set.

Equivalently, [1] this problem is equivalent to solve equation (1). In this setting it is to minimize

$$2\left(y_1 - x_{11}(\beta_1 + \beta_2)\right)^2 \text{ subject to } |\beta_1| + |\beta_2| \leq s. \tag{8}$$

which corresponds to equation (8) for some $\lambda = \lambda(s)$. ◀

(d) Argue that in this setting, the lasso coefficients $\hat{\beta}_1$ and $\hat{\beta}_2$ are not unique; in other words, there are many possible solutions to the optimization problem in (c). Describe these solutions.

Solution. The minimum value of the cost function $2\left(y_1 - x_{11}(\beta_1 + \beta_2)\right)^2$ is clearly taking when its value is zero, i.e., for (β_1, β_2) such that

$$\beta_1 + \beta_2 = \frac{y_1}{x_{11}} := c$$

In figure 2 we have represented the cost function to have an intuitive idea of its shape. We can see that the minima is taken along a straight line, as we have already said.

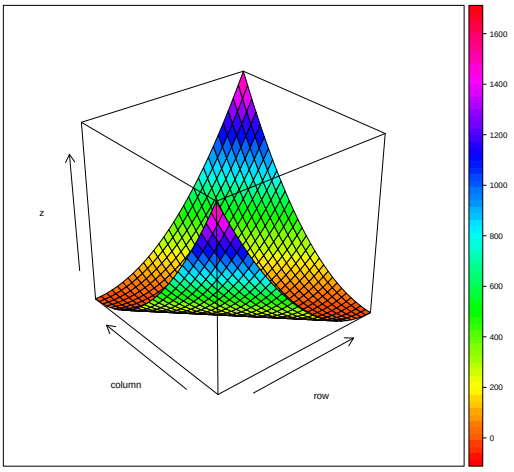

Figure 2: Graphical representation of the cost function in (8).

However, we want to find the optimal solution under the restriction $|\beta_1|+|\beta_2| \leq s$. In order to do so, fixed the value of s (or equivalently the value of λ), we distinguish two cases:

(a) If $|c| > s$, then due to the shape of the cost function (see Figure 2) the minimum is taken in the boundary of the rhomboidal region defined by the restriction. We distinguish two sub-cases depending on the sign of c:

(a.1) If $c > 0$ then the multiple possible solutions are those such that $\beta_1+\beta_2 = s$ and $0 \leq \beta_1, \beta_2 \leq s$.

(a.2) If $c < 0$ then the multiple possible solutions are those such that $\beta_1 + \beta_2 = -s$ and $-s \leq \beta_1, \beta_2 \leq s$.

(b) If $|c| \leq s$ then the minima is not always taken at the boundary (for $|c| < s$ it is inside the region defined by the restriction). The possible solutions are the pairs (β_1, β_2) such that $\beta_1 + \beta_2 = c$ and $|\beta_1| + |\beta_2| \leq s$.

In particular, $\hat{\beta}_1, \hat{\beta}_2$ are not unique.

◀

References

[1] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2014. Chapters 5 and 6, *An Introduction to Statistical Learning: With Applications in R*. Springer Publishing Company, Incorporated.

[2] Lecture notes on ridge regression, Wessel N. van Wieringen. Version 0.10, January 01, 2018. https://arxiv.org/pdf/1509.09169.pdf