Elements Of Statistical Learning Solutions

Solutions to select exercises

Krishan Bhalla

June 3, 2020

CONTENTS

1

INTRODUCTION

A selection of solutions to Elements of Statistical Learning by Hastie, Tibshirani, and Friedman. This will be continually updated as I work through the book.

I will not answer all exercises, but will endeavour to solve many.

This chapter largely exists to align latex chapter labels with those of the book.

2

CHAPTER 2 - OVERVIEW OF SUPERVISED LEARNING

2.1 Suppose each of K-classes has an associated target t_k which is a vector of all zeros except one in the kth position. Show that classifying to the largest element of $\hat{y}$ amounts to choosing the closest target $min_k \|t_k - \hat{y}\|$ if the elements of y sum to one.

Our norm here is the L^2 norm.

$$\begin{aligned} argmin_k \|t_k - \hat{y}\| &= argmin_k \|t_k - \hat{y}\|^2 \\ &= argmin_k \sum_i (\hat{y}_i - (t_k)_i)^2 \\ &= argmin_k \sum_i (\hat{y}_i - \delta_{i,k})^2 \\ &= argmin_k \left(1 - 2\hat{y}_k + \sum_i \hat{y}_i^2\right) \\ &= argmin_k (-2\hat{y}_k) \end{aligned}$$

Where the last line follows as it is the only term dependant on k. *argmin* is independent of scale, so the above states that the k corresponding to the minimum value of the norm is exactly the largest element of $\hat{y}$

2.2 Show how to compute the Bayes decision boundary for the example in Figure 2.5.

Setup: There are 10 means m_k from a bivariate guassian distribution $N((1,0)^T, I)$ and we label this class blue. We take 10 n_k from $N((0,1)^T, I)$ and label this class orange. For each class there were 100 observations, where each observation was generated by picking one of the m_k (n_k resp) each with equal probability, and taking a point randomly from a $N(m_k, I/5)$ distribution.

Let x be an observation. We equate posteriors for x being blue or yellow, and note that in our setup $\mathbb{P}(blue) = \mathbb{P}(orange)$ (priors are equal) to simplify to:

$$\sum_k exp(-5\|m_k - x\|^2) = \sum_k exp(-5\|n_k - x\|^2)$$

This defines a curve in the plane separating the two classes.

2.3 Derive equation 2.24.

Setup: Consider N data points uniformly distributed in a p-dimensional unit ball around the origin. Consider a nearest neighbour estimate at the origin. The median distance from the origin to the closest point is given by:

$$d(p_N) = \left(1 - \frac{1}{2}^{1/N}\right)^{1/p}$$

Solution: Let x be a point and let $y = \|x\|$. y has cdf equal to the ratio of a ball of radius y to a ball of radius 1, i.e. $F(y) = y^p$. The minimum over all x then has cdf $F_{ymin}(y) = 1 - (1 - F(y))^N$ (a general fact for order statistics). Thus $F_{ymin} = 1 - (1 - y^p)^N$.

The median distance for $ymin$ is when $F_{ymin}(y) = 1/2$. Solving for this yields the result.

2.4 Setup as in book. Projection in direction a.

Pick an orthonormal basis of $\mathbb{R}^p$ which includes the vector a, say $a_1, \ldots, a_p$ with $a_1 = a$. Then each $x_i = \sum_j X_{i,j} a_j$, and so $z_i = X_{i,1}$ where X is the matrix with rows x_i

The x_i have distribution $N(0, I_p)$, and under such a distribution each component of x_i has distribution $N(0, 1)$.

In particular this means that each $X_{i,j}$ has distribution $N(0, 1)$ and so the z_i do.

The squared distance from the origin is just z_i^2, with distribution χ_1^2, and this has mean 1.

2.6 Consider a regression problem with inputs x_i and outputs y_i, and a parameters model $f_\theta(x)$ to be fit by least squares. Show that if there are observations with tied or identical values of x then the fir can be obtained from a reduced weighted least squares problem.

The problem can of finding θ amounts to solving the following:

$$argmin_\theta \left(y - f_\theta(x)\right)^T \left(y - f_\theta(x)\right)$$

Denote by $z_1, \ldots, z_M$ the unique values of x in our training set, denote by n_j the number of occurrences of value z_j. Then let $t_j = \frac{1}{n_j} \sum_{i=1}^{n_j} y_i$ If we can get to the following, we're done (as it is in the form of weighted regression).

$$argmin_\theta \sum_j n_j \left(t_j - f_\theta(z_j)\right)^2$$

Expanding the initial expression we get (denoting by $y_{i,j}$ the ith value of y corresponding to input z_j):

$$\begin{aligned}
(y - f_\theta(x))^T (y - f_\theta(x)) &= \sum_i (y_i - f_\theta(x_i))^2 \\
&= \sum_i \left(y_i^2 + f_\theta(x_i)^2 - 2y_i f_\theta(x_i)\right) \\
&= \sum_i \left(y_i^2 + f_\theta(x_i)^2 - 2y_i f_\theta(x_i)\right) \\
&= \sum_{j=1}^{M} \sum_{i=1}^{n_j} y_{i,j}^2 - 2f_\theta(z_j)y_{i,j} + f_\theta(z_j)^2 \\
&= \sum_{j=1}^{M} \left(\sum_i y_{i,j}^2\right) - 2n_j f_\theta(z_j)t_j + n_j f_\theta(z_j)^2 \\
&= \sum_{j=1}^{M} n_j(t_j - f_\theta(z_j))^2 - \sum_{j=1}^{M} n_j t_j^2 + \sum_{j=1}^{M} \left(\sum_i y_{i,j}^2\right)
\end{aligned}$$

This last trick of adding 0 leaves us with an expression where only the first sum is dependant on θ Thus when taking $argmin_\theta$ we can ignore the last two terms.

This leaves us with the required equivalence.:

$$argmin_\theta \, (y - f_\theta(x))^T (y - f_\theta(x)) = argmin_\theta \sum_j n_j \left(t_j - f_\theta(z_j)\right)^2$$

3

CHAPTER 3 - LINEAR METHODS FOR REGRESSION

3.1 Show that the F-statistic for dropping a single coefficient of a model is equivalent to the square of the corresponding z score

Let X be our data, and let $v_{j,j}$ be the jth diagonal element of $V = (X^T X)^{-1}$. $z_j = \hat{\beta}_j / \hat{\sigma}\sqrt{v_{j,j}}$ is the z score.

The F statistic is

$$F = \frac{(RSS_0 - RSS_1)/(p1 - p0)}{RSS_1/(N - p_1 - 1)}$$

Where the regression models are have $p_1 + 1$ and $p_0 + 1$ degrees of freedom respectively. We also know that $\hat{\sigma}^2$ is equivalent to the denominator. In the case of dropping a single variable, this simplifies to:

$$F = \frac{RSS_0 - RSS_1}{\hat{\sigma}^2}$$

Where $\hat{\sigma}$ is derived from the bigger model.

Thus our question can be simplifed to showing that:

$$RSS_0 - RSS_1 = \hat{\beta}_n^2 / v_{j,j}$$

We know $\hat{\beta}$ $N(\beta, \sigma^2 V)$ and so $\hat{\beta}_j$ $N(\beta_j, \sigma^2 v_{j,j})$.

Under the null-hypothesis $\beta_j = 0$ and so: $\hat{\beta}_j = \sigma\sqrt{v_{j,j}} Z$ where $Z \sim N(0,1)$ Thus $\hat{\beta}_j^2 = \sigma^2 v_{j,j} Z^2 = \sigma^2 v_{j,j} Q$ where $Q \sim \chi_1^2$

Similarly RSS_0, RSS_1 have distribution $\sigma^2 \chi^2_{N_i}$ where N_i is the number of degrees of freedom. Thus $RSS_0 - RSS_1 \sim \sigma^2 \chi_1^2$

Hence $\hat{\beta}_j / v_{j,j}$ and $RSS_0 - RSS_1$ have the same distribution. They further test the same hypothesis and thus must be identical.

3.3 a. Prove the Gauss-Markov theorem: the least squares estimate of a parameter $a^T\beta$ has a variance no bigger than that of any other linear unbiased estimate of $a^T\beta$.
b. Secondly, show that if $\hat{V}$ is the variance-covariance matrix of the least squares estimate of β and $\tilde{V}$ is the variance covariance matrix of any other linear unbiased estimate, then $\hat{V} \leq \tilde{V}$, where $B \leq A$ if $A - B$ is positive semidefinite.

First note that part b implies part a. If β has dimension 1, then V is just the variance of beta, and $\leq$ is equivalent to the normal $\leq$ operator. Taking the inner product with a is just a linear operation. We thus only need to show b.

Suppose $\hat{\beta}$ is the OLS estimate of β and that $\tilde{\beta}$ is another linear unbiased estimate. The the variance-covariance matrices be $\hat{V}$ and $\tilde{V}$ resp. $\hat{\beta} = \left(X^TX\right)^{-1}X^Ty$ so $\hat{\beta} = Cy$ say, and write $\tilde{\beta} = (C+D)y$ for some non-zero mxn matrix D.

$$\begin{aligned}
\mathbb{E}\left(\tilde{\beta}\right) &= \mathbb{E}\left((C+D)\left(X\beta+\epsilon\right)\right) \\
&= \mathbb{E}\left(\left(X^TX\right)^{-1}X^T + D\right)\left(X\beta+\epsilon\right) \\
&= \mathbb{E}\left((\beta + DX\beta)\right) + \zeta\mathbb{E}(\epsilon) \\
&= \mathbb{E}\left((\beta + DX\beta)\right) \\
&= \beta + DX\beta \\
&= \beta
\end{aligned}$$

Where the last lines follow as $\tilde{\beta}$ is unbiased. In particular $DX\beta = 0$ and so $DX = 0$ as beta is an unobserved parameter to be estimated.

$$\begin{aligned}
\tilde{V} &= Var\left(\tilde{\beta}\right) \\
&= Var\left((C+D)\,y\right) \\
&= (C+D)\,(C+D)^T\,Var\,(y) \\
&= \sigma^2\,(C+D)\,(C+D)^T \\
&= \sigma^2\left(CC^T + CD^T + DC^T + DD^T\right) \\
&= \sigma^2\left(CC^T + CD^T + DC^T + DD^T\right) \\
&= \hat{V} + \sigma^2\left(+CD^T + DC^T + DD^T\right) \\
&= \hat{V} + \sigma^2\left(\left(X^TX\right)^{-1}X^TD^T + DX\left(X^TX\right)^{-1} + D^TD\right) \\
&= \hat{V} + \sigma^2DD^T
\end{aligned}$$

Where we know that DD^T is positive semi-definite[1] and so are done.

1 $v^TDD^Tv = \left\|D^Tv\right\| >= 0\ \forall v$

3.4 Show how the vector of least squares coefficients can be obtained from a single pass of the Gram–Schmidt procedure (Algorithm 3.1). Represent your solution in terms of the QR decomposition of X.

Let $X = QR$ be the QR decomposition of X. This can be attained via Gram-Schmidt. We assume that R has no zeros on the diagonal (i.e. the variables are linearly independent). Then

$$\begin{aligned}\hat{\beta} &= (X^T X)^{-1} X^T y \\ &= (R^T R)^{-1} R^T Q^T y \\ &= R^{-1} Q^T y\end{aligned}$$

We can invert R via backpropogation, and $Q^T y$ is easily calculable.

3.5 Show that the ridge regression problem (using α to denote the constant intercept vector)

$$\hat{\beta} = argmin_{\beta,\alpha} (y - \alpha - X\beta)^T (y - \alpha - X\beta) + \lambda \|\beta\|^2$$

is equivalent to the problem:

$$\hat{\beta^c} = argmin_{\beta^c,\alpha^c} \left(y - \alpha^c - \tilde{X}\beta^c\right)^T \left(y - \alpha^c - \tilde{X}\beta^c\right) + \lambda \|\beta^c\|^2$$

Where $\tilde{X} = X - \bar{X}$, and bar represents the the $Nxp)$ matrix where each value in the jth column is the mean x_j. Give the correspondence between β^c and the original β. Do the same for the lasso.[2]

This problem is easier with summation

$$\begin{aligned}&(y - \alpha - X\beta)^T (y - \alpha - X\beta) + \lambda \|\beta\|^2 \\ &= (y - \alpha - \bar{X}\beta - (X - \bar{X})\beta)^T (y - \alpha - \bar{X}\beta - (X - \bar{X})\beta) + \lambda \|\beta\|^2 \\ &= (y - (\alpha + \bar{X}\beta) - (X - \bar{X})\beta)^T (y - (\alpha + \bar{X}\beta) - (X - \bar{X})\beta) + \lambda \|\beta\|^2 \\ &= \left(y - \alpha^c - \tilde{X}\beta^c\right)^T \left(y - \alpha^c - \tilde{X}\beta^c\right) + \lambda \|\beta^c\|^2\end{aligned}$$

Where $\alpha_i^c = \alpha_i + \sum_{j=1}^p \bar{x}_j \beta_j$ in all coordinates, and $\beta^c = \beta$. This is an expression of our desired form. This problem is equivalent to demeaning the data and adjusting the intercept. One can do exactly the same for the lasso.

2 This is all a tad odd as regression with intercept is desired, but we only penalise the non-intercept terms.

3.6 Show that the ridge regression estimate is the mean (and mode) of the posterior distribution, under a Gaussian prior β $N(0, \tau I)$ and Gaussian sampling model y $N(X\beta, \sigma^2 I)$. Find the relationship between the ridge parameter λ and the variances τ and σ^2

This question really states that the pdf of the posterior is proportional to the pdfs of y given β and β. Hence:

$$
\begin{aligned}
f(\beta|y) &\propto f(\beta)f(y|\beta) && (1)\\
log(f(\beta|y)) &= C + log(f(\beta)) + log(f(y|\beta)) && (2)\\
&= C + -\frac{1}{2\tau}\beta^T\beta * log\left(\frac{1}{\sqrt{2*\pi*\tau}}\right) && (3)\\
&- \frac{1}{2\sigma^2}\left(y - X\beta\right)^T\left(y - X\beta\right) * log\left(\frac{1}{\sqrt{2*\pi*\sigma^2}}\right) && (4)\\
&= C + -\frac{1}{2\tau}\beta^T\beta + -\frac{1}{2\sigma^2}\left(y - X\beta\right)^T\left(y - X\beta\right) && (5)
\end{aligned}
$$

Where we absorb terms into the constant as needed. We have recovered that:

$$
\begin{aligned}
f(\beta|y) &= C_1 e^{-\frac{1}{2\sigma^2}\left(\frac{\sigma^2}{\tau}\beta^T\beta + (y-X\beta)^T(y-X\beta)\right)} && (6)\\
&= C_1 e^{-\frac{1}{2\sigma^2}\left(\beta^T\left(X^TX + \frac{\sigma^2}{\tau}I\right)\beta + y^Ty - y^TX\beta - \beta^TX^Ty\right)} && (7)\\
&= C_1 e^{-\frac{1}{2\sigma^2}\left(\beta^T\Sigma\beta + y^Ty - y^TX\beta - \beta^TX^Ty\right)} && (8)\\
&= C_1 e^{-\frac{1}{2\sigma^2}\left(\left(\Sigma\beta - X^Ty\right)^T\Sigma^{-1}\left(\Sigma\beta - X^Ty\right) + y^Ty - y^TX^TXy\right)} && (9)\\
&= C_2 e^{-\frac{1}{2\sigma^2}\left(\Sigma\beta - X^Ty\right)^T\Sigma^{-1}\left(\Sigma\beta - X^Ty\right)} && (10)\\
& && (11)
\end{aligned}
$$

Where $\Sigma = \left(X^TX + \frac{\sigma^2}{\tau}I\right)$ is a $(p+1)*(p+1)$ matrix, and for (11) we absorbed into the constant any terms not dependant on β

Thus the posterior has a multivariate gaussian pdf (up to scaling), so the mean and mode of the distribution are identical, and can be found by maximising (5) over β, which due to the minus sign is sufficient.

This amounts to

$$argmin_\beta\left(\frac{1}{\tau}\beta^T\beta + \frac{1}{\sigma^2}\left(y - X\beta\right)^T\left(y - X\beta\right)\right)$$

And hence is equivalent to solving:

$$argmin_\beta\left(\frac{\sigma^2}{\tau}\beta^T\beta + \left(y - X\beta\right)^T\left(y - X\beta\right)\right)$$

Hence the ridge regression parameter is σ^2/τ

3.9 Forward stepwise regression: Suppose we have the QR decomposition for the N x q matrix X_1 in a multiple regression problem with response y, and suppose we have an additional $p - q$ predictors in the matrix X_2. Denote the residual by r. Describe an efficient procedure for establishing which additional variable will reduce the residual sum of squares the most.

Intuition - pick the column $\hat{v}$ such that v has the least angle with r. i.e.

$$\hat{v} = argmax_{v \in \{columns X_2\}} \frac{|r^T v|}{\|v\|}$$

With this in mind, let $u_j = x_j - \frac{|r^T x_j|}{\|x_j\|}$ where the x_j are the columns of X_2, and let $v_j = \frac{u_j}{\|u_j\|}$

$$\begin{aligned}
RSS &= r^T r \\
r &= y - \hat{y} \\
&= y - R^{-1} Q^T y \\
let\ r_j &= r - \left(r^T u_j\right) u_j \\
and\ RSS_j &= RSS - 2\left(r^T u_j\right)^2 + \left(r^T u_j\right)^2 \\
then\ RSS_j &= RSS - \left(r^T u_j\right)^2
\end{aligned}$$

Where RSS_j is the residual sum of squares for our new model. This verifies our intuition, RSS_j is minimised when we pick the column of X_2 with least angle to r. I assume the efficiency in the question comes from the ease of inverting R compared to $X^T X$ (backpropogation will do), and that we can extend our QR decomposition to include the new variable easily via Gram-Schmidt. In particular most of what is needed for Gram-Schmidt is already computed when taking inner product with the residual.

3.10 Backward stepwise regression. Suppose we have the multiple regression fit of y on X along with the standard errors and Z-scores. We wish to establish which variable, when dropped, will increase the RSS the least. How would you do this?

From question 3.1 we know that the F-statistic for dropping a single coefficient of a model is equivalent to the square of the corresponding Z score. Further, we know the F statistic in the case of dropping 1 variable is:

$$\frac{(RSS_0 - RSS_1)}{RSS_1/(N - p1 - 1)}$$

where N, p_1 and RSS_1 are constant. In particular, the change is RSS is proportional to the F-score with a constant that does not depend on choice of variable to be dropped, and so the change in RSS is proportional to the square of the z-score in the same manner. Thus the difference will be smallest (smallest increase in RSS) if the variable has minimal z-score in our model.

3.12 Show the ridge regression estimates can be obtained by OLS regression on an augmented data set. Add p rows to the centered matrix X, $\sqrt{\lambda}I_p$, and augment y with p zeros.

Let $X_2 = [X^T, \sqrt{\lambda}I_p]^T$ be our augmented matrix, and let $y_2 = [y^T, 0]^T$ be the augmented response. Under OLS, we have $\beta_2 = (X_2^T X_2)^{-1} X_2^T Y$.

$$\begin{aligned} X_2^T X_2 &= [X^T, \sqrt{\lambda}I_p][X^T, \sqrt{\lambda}I_p]^T \\ &= X^T X + \lambda I_p \\ X_2^T y_2 &= [X^T, \sqrt{\lambda}I_p][y^T, 0]^T \\ &= X^T y \end{aligned}$$

Thus $\beta = \left(X^T X + \lambda I_p\right) X^T y$ which is exactly the ridge regression beta for our non-augmented dataset.

3.13 Derive expression 3.62 and show that $\hat{\beta}^{pcr}(p) = \hat{\beta}^{ls}$

Given an SVD of X, $X = UDV^T$ say, with $V = [v_1, \ldots, v_p]$, the principal components z_m in equation 3.61 are defined as Xv_m for all m. The principal component regression is, for any $M <= p$:

$$\begin{aligned} \hat{y}^{pcr} &= \bar{y} + \sum_{m=1}^{M} \hat{\theta_m} z_m \\ &= \bar{y} + X \sum_{m=1}^{M} \hat{\theta_m} v_m \end{aligned}$$

So we can just set $hat\beta^{pcr}(M) = \sum_{m=1}^{M} \hat{\theta_m} v_m$ and we're done. Now it remains to show that $\hat{\beta}^{pcr}(p) = \hat{\beta}^{ls}$.

$$\begin{aligned}
\hat{\beta}^{pcr}(p) &= \sum_{m=1}^{p} \hat{\theta_m} v_m \\
&= V[\frac{z_1^T y}{z_1^T z_1}, \ldots, \frac{z_p^T y}{z_p^T z_p}]^T \\
&= V[\frac{z_1^T y}{d_1^2}, \ldots, \frac{z_p^T y}{d_p^2}^T \\
&= VD^{-2}[z_1^T y, \ldots, z_p^T y]^T \\
&= VD^{-2}[u_1^T d_1^T y, \ldots, u_p^T d_p^T y]^T \\
&= VD^{-2}D[u_1^T y, \ldots, u_p^T y]^T \\
&= VD^{-1}U^T y
\end{aligned}$$

Where we made use of the SVD. Now:

$$\begin{aligned}
\hat{\beta}^{ls} &= \left(X^T X\right)^{-1} X^T y \\
&= \left(VDU^T UDV^T\right)^{-1} UDV^T y \\
&= \left(VDDV^T\right)^{-1} VDU^T y \\
&= D^{-2}\left(VV^T\right)^{-1} VDU^T y \\
&= D^{-2} VDU^T y \\
&= VD^{-1}U^T y
\end{aligned}$$

Using the orthonomality of U and V.

3.14 Show that in the orthogonal case, partial least squares stops after m = 1 steps.

Assume X is such that each column has mean zero, unit variance, and the colums are orthogonal. let $z = \sum_i \left(x_i^T y\right) x_i$ and $\hat{\theta} = \frac{z^T y}{z^T z}$.

In this case

$$\begin{aligned}
z^T z &= \sum_i \sum_j \left(x_i^T y\right)\left(x_j^T y\right) x_i^T x_j \\
&= \sum_i \left(x_i^T y\right)^2 \left(x_i^T x_i\right) \\
&= \sum_i \left(x_i^T y\right)^2
\end{aligned}$$

Similarly $z^T y = \sum_i \left(x_i^T y\right)^2$. Now let $x_j^{(1)} = x_j - \frac{z^T x_j}{z^T z}$ Then let:

$$x_j^{(1)} = x_j - \frac{z^T x_j}{z^T z} z$$

In the next iteration, we have:

$$\begin{aligned}
\langle x_j^{(1)}, y\rangle &= x_j^T y - \frac{\left(x_j^T y\right)\left(x_j^T x_j\right)}{\sum_i \left(x_i^T y\right)^2} z^T y \\
&= x_j^T y - \frac{x_j^T y}{\sum_i \left(x_i^T y\right)^2} \sum_i \left(x_i^T y\right)^2 \\
&= x_j^T y - x_j^T y \\
&= 0
\end{aligned}$$

So algorithm 3.3 (page 81 in my edition) terminates after 1 step.

3.19 Show that $\|\hat{\beta}^{ridge}\|$ *increasesasthetuningparameter*$\lambda \to 0$. *Does the same property hold for the Lasso and PLS?*

Throughout this question I use regression without intercept as the intercept is not included in the authors formulation of the penalty in the Lagrangian form.

a) Ridge

For data X with mean 0 and unit variance, we have

$$\hat{\beta}^{ridge} = (X^T X + \lambda I)^{-1} X^T y$$

Then:

$$\begin{aligned}
\left\|\hat{\beta}^{ridge}\right\|^2 &= \left((X^T X + \lambda I)^{-1} X^T y\right)^T (X^T X + \lambda I)^{-1} X^T y \\
&= \left(V(D^T D + \lambda I)^{-1} D U^T y\right)^T V(D^T D + \lambda I)^{-1} D U^T y \\
&= y^T U D^2 (D^T D + \lambda I)^{-2} U^T y
\end{aligned}$$

Using the SVD, the commutativity of D and V, and the orthonormality of V. $D^T D + \lambda I$ is diagonal, and so $D^2(D^T D + \lambda I)^{-2}$ is too with entries $\frac{d_j^2}{d_j^2+\lambda}$ on the diagonal. Let $z_j = (U^T y)_j$ Then we see:

$$\begin{aligned}
\left\|\hat{\beta}^{ridge}\right\|^2 &= \sum_i \sum_j \frac{d_j^2 z_i^T z_j}{d_j^2 + \lambda} \\
&= \sum_j \frac{d_j^2 z_j^2}{d_j^2 + \lambda}
\end{aligned}$$

All terms are non-negative, and increase with decreasing λ so $\hat{\beta}^{ridge}$ must too. Recall that we can view the ridge regression estimate as the solution to

$$\hat{\beta}^{ridge} =_\beta (y - X\beta)^T (y - X\beta) + \lambda \|\beta\|_2^2$$

b) Lasso

Under a similar formulation, we have

$$\hat{\beta}^{lasso} = argmin_{\beta}(y - X\beta)^T(y - X\beta) + \lambda \|\beta\|_1$$

In a similar way one can see that if λ is sufficiently large, it dominates this expression (for fixed X and y), and so $\hat{\beta}^{lasso}$ will decrease in norm with increasing λ.

c) PLS

?

3.23. *Please refer to the book. LAR*

a)Correlation with residuals remains constant in absolute value.

$$\begin{aligned}
\frac{1}{N}X^T\left(y - u(\alpha)\right) &= \frac{1}{N}\left(X^Ty - \alpha X^TX\hat{\beta}\right) \\
&= \frac{1}{N}\left(X^Ty - \alpha X^TX\left(X^TX\right)^{-1}X^Ty)\right) \\
&= \frac{1}{N}\left(X^Ty - \alpha X^Ty)\right) \\
&= \frac{1-\alpha}{N}X^Ty
\end{aligned}$$

Thus $|\frac{1}{N}X^T\left(y - u(\alpha)\right)| = (1-\alpha)[\lambda, \ldots, \lambda]^T$

Which is exactly the required result in vector notation.

b)Explicit form of correlation

The question as stated ignores the need for an absolute value, so we shall assume that $\langle x_j, y - u(\alpha)\rangle >= 0$ for every j, else replace x_j by $-x_j$ in our data.

Correlations are given by covariance divided by the product of the standard deviations. as everything (data and response) is assumed to be standardised, we have:

$$\begin{aligned}
\left(y - u(\alpha))^T(y - u(\alpha)\right) &= y^Ty - 2y^Tu(\alpha) + u(\alpha)^Tu(\alpha) \\
&= y^Ty - 2\alpha y^TX\hat{\beta} + \alpha^2\hat{\beta}^TX^TX\hat{\beta} \\
&= y^Ty - 2\alpha y^TX\hat{\beta} + \alpha^2\hat{\beta}^TX^TX\left(X^TX\right)^{-1}X^Ty \\
&= y^Ty - 2\alpha y^TX\hat{\beta} + \alpha^2\hat{\beta}^TX^Ty \\
&= y^Ty + \alpha(\alpha - 2)y^TX\hat{\beta}
\end{aligned}$$

Setting $\alpha = 1$ we get $RSS = y^T y - y^T X\hat{\beta}$. Thus:

$$\left(y - u(\alpha)\right)^T (y - u(\alpha)\Big) = y^T y + \alpha(\alpha - 2)(y^T y - RSS)$$

And so:

$$\begin{aligned}\Big(y - u(\alpha)\big)^T (y - u(\alpha)\Big) &= (1-\alpha)^2 y^T y + \alpha(2-\alpha) RSS \\ &= N(1-\alpha)^2 + \alpha(2-\alpha)RSS\end{aligned}$$

$$\begin{aligned}Corr\left(x_j, y - u(\alpha)\right) &= \frac{\langle x_j, y - u(\alpha)\rangle / N}{\sqrt{\langle x_j, x_j\rangle / N}\sqrt{\langle y - u(\alpha), y - u(\alpha)\rangle / N}} \\ &= \frac{\lambda(1-\alpha)}{1 \cdot \sqrt{(1-\alpha)^2 + \alpha(2-\alpha)RSS/N}} \\ &= \frac{(1-\alpha)}{\sqrt{(1-\alpha)^2 + \alpha(2-\alpha)RSS/N}}\lambda\end{aligned}$$

As required.

c) Show the LAR algorithm keeps correlations tied and monotonically decreasing.

Part a) showed that the LAR algorithm keeps correlations tied ad this is exactly the step taken for an active set of variables in the *kth* step. Part b) shows that correlations are monotonically decreasing with α as numerator falls faster than the numerator for α in $[0, 1]$[3]. In particular $\lambda(0) = \lambda$ and $\lambda(1) = 0$.

3.29 Please see the book. Ridge regression with duplicate variables

Setup: The data X has identical columns x.

Let $X = [x, \ldots, x]$ say, for x some column vector, where X has dimension nxm. The ridge regression beta is given by

$$\left(X^T X + \lambda I\right)^{-1} X^T y$$

Note $X^T y = [x^T y, \ldots, x^T y]^T$ $X^T X$ is a constant matrix with every value $x^T x$.

Let $M = X^T X / (x^T x)$ be the matrix of ones.

Looking for an inverse of the form $sM + \lambda^{-1} I$ yields

$$s = \frac{-x^T x}{\lambda(\lambda + m x^T x)}$$

3 One can and should compute the gradient and check that it is negative on paper.

Where one uses that $M^T M = mM$. We have

$$(sM + \lambda^{-1} I)[x^T y, \ldots, x^T y]^T = [c, \ldots, c]^T$$

Where

$$\begin{aligned} c &= \frac{-m \cdot x^T x \cdot x^T y}{\lambda(\lambda + m \cdot x^T x)} + \frac{x^T y}{\lambda} \\ &= \frac{(\lambda + m \cdot x^T x) x^T y - m \cdot x^T x \cdot x^T y}{\lambda(\lambda + m \cdot x^T x)} \\ &= \frac{\lambda x^T y}{\lambda(\lambda + m \cdot x^T x)} \\ &= \frac{x^T y}{\lambda + m \cdot x^T x} \end{aligned}$$