STATG006: Solutions to Exercise Sheet #6

The exercises in this sheet focus on model selection and sparse regression. As before, we provide solutions, sometimes detailed, sometimes a sketch that should point you to the complete solution. Skeetches should not be taken at face value as the level of detail required for an exam answer.

1. (a) Best subset, by definition, does an exaustive search on the space of models with k predictors. Therefore, by construction it will be the best.

   (b) This is not possible to say without actually testing them. It may be fair to say that there is no a priori reason to think it would be the one given by best subset search. This is the discrete counterpart to the problem we saw in Exercise Sheet #4, where a cubic polynomial was not guaranteed to get best test error than a fully linear model, even if the truth was a cubic polynomial.

   (c.i) True, as by construction this is a greedy search algorithm that builds solutions in stage $k + 1$ by modifying the solution found in stage k.

   (c.ii) True, by a similar reasoning.

   (c.iii) False, this would be equivalent to saying that both methods present the same solution, which is not true.

   (c.iv) False, by a similar reasoning.

   (c.v) False. The search space of subset selection is complex, without any special structure that would have this property. If this was true, best subset would be identical to backward seletion (why?).

2. (a) The lasso is more flexible in the sense that least-squares is the special case where the regularizer is given a weight of zero. The regularizer increases bias (since it moves the estimator way from the minimiser of RSS) in an attempt to reduce variance (shrinking coefficients towards zero). Hence, the correct alternative is (i).

   (b) The reasoning is exactly the same as in (a).

   (c) This depends on our assumptions. If we assume it is possible for the truth to be non-linear (not an unreasonable assumption!), then there will be less bias. It is less obvious whether variance will be reduced, depending of the sample size and dimensionality of the problem.

3. (a) It will steadily decrease (until it plateaus), as this is equivalent to expanding the feasible region of a constrained least-squares problem – and we cannot have a worse solution in a constrained optimisation method by expanding the feasible region. Hence, answer (iv).

(b) The answer is (ii). I hope by now this is obvious!

(c) The answer is (iii). This is closely related to the reason why test set error increases: a bias-variance trade-off that starts to go bad as variance dominates bias.

(d) The answer is (iv). This is closely related to the reason why training set error decreases: the bias-variance trade-off does not apply to the training error.

(e) The answer is (v). Irreducible error is a population concept (see Chapter 2 of ISLR), and as such it is not affected by how we estimate our model.

4. This question is entirely analogous to the previous question, the only difference being that the optimisation problem is cast in a Lagrange multiplier formulation instead of an explicit constrained optimisation problem. The fact that this is not l_1 regularisation, but something else, does not affect the answers.

5. Here I am going to use my notation, where $x_1^{(1)}$ and $x_2^{(1)}$ mean "measurements of covariate X_1 at data points 1 and 2", in contrast with the book's notation x_{11}, x_{12}.

   (a) Using the assumptions, $(y^{(1)} - \beta_0 - \beta_1 x_1^{(1)} - \beta_2 x_2^{(1)})^2 = (-y^{(2)} - \beta_0 + \beta_1 x_1^{(2)} + \beta_2 x_2^{(2)})^2 = (y^{(2)} + \beta_0 - \beta_1 x_1^{(2)} - \beta_2 x_1^{(2)})^2$. For simplicity, let's drop the superscript and define $x \equiv x_1 = x_2$. We can write the optimisation problem as

   $$(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2) = \text{argmin}_{\beta_0,\beta_1,\beta_2}(y-\beta_0-\beta_1 x-\beta_2 x)^2+(y+\beta_0-\beta_1 x-\beta_2 x)^2+\lambda(\beta_1^2+\beta_2^2),$$

   which simplifies to

   $$(\hat{\beta}_1, \hat{\beta}_2) = \text{argmin}_{\beta_1,\beta_2} 2(y - \beta_1 x - \beta_2 x)^2 + \lambda(\beta_1^2 + \beta_2^2),$$

   since we know that $\hat{\beta}_0 = 0$ (which you can verify by differentiation).

   (b) Just differentiate the objective function with respect to β_1 and β_2, setting the expressions to zero:

   $$\begin{aligned} -4(y - \beta_1 x - \beta_2 x)x + 2\lambda\beta_1 &= 0 \\ -4(y - \beta_1 x - \beta_2 x)x + 2\lambda\beta_2 &= 0 \end{aligned}$$

   so that

   $$\begin{aligned} 2\lambda\beta_1 &= 4(y - \beta_1 x - \beta_2 x)x \\ 2\lambda\beta_2 &= 4(y - \beta_1 x - \beta_2 x)x \end{aligned}$$

   and the result follows from taking the ratios.

   (c)

   $$(\hat{\beta}_1, \hat{\beta}_2) = \text{argmin}_{\beta_1,\beta_2} 2(y - \beta_1 x - \beta_2 x)^2 + \lambda(|\beta_1| + |\beta_2|).$$

(d) First show that in the unconstrained RSS problem, we get as a set of solutions $\beta_1 + \beta_2 = y/x$. This would correspond to a line passing through the points $(0, y/x)$ and $(y/x, 0)$ in a $\beta_1 \times \beta_2$ coordinate. Now think of the problem finding which would be the smallest feasible region that would contain at least one of the points in this line? We know that the feasible region is shaped like a diamond, where the side on the upper right quadrant is also given by a line of slope -1, just like the line $\beta_2 = -\beta_1 + y/x$. So for every feasible region of this size or smaller (there is no reason for making it larger, as this will not decrease the RSS but will increase the penalty), all points lying on the upper right boundary of the diamond will be optimal solutions as these will be the points equally close to the line $\beta_2 = -\beta_1 + y/x$ (contrast this to the ridge regression feasible region). See Figure 6.7 or ISLR, where the RSS minimiser is now a whole line instead of a single point, and the ellipsoid is a degenerate set of parallel lines $\beta_1 + \beta_2 = constant$.

6. I will not draw the graphs here, but just hint how it could be solved analytically. I hope you find (a) obvious: it is just a matter of taking the derivatives and setting it to 0. For (b), it will pay off to split the possible values of β_1 into two: for $\beta_1 > 0$, verify by differentiation of $(y_1 - \beta_1)^2$ that $-2y_1 + 2\beta - \lambda = 0$. If the solution of that is positive, then this will be the optimal value of β_1. If not, the closest point to it in the feasible region $[0, \infty)$ will be zero itself. Put these ideas together to get (6.15).

7. For that, first define x_1 and x_2 as the binary indicators of the factor level, so that $x_1^{(i)} + x_2^{(i)} = 0$ for any data point i. Realize that $(x_1^{(i)})^2 = x_1^{(i)}$ and that $x_1^{(i)} x_2^{(i)} = 0$. Denoting $\bar{z}$ as the sample average of any sample $z^{(1)}, \ldots, z^{(n)}$, first optimise the objective function of group lasso,

$$\sum_{i=1}^{n} (y^{(i)} - \beta_0 - \beta_1 x_1^{(i)} - \beta_2 x_2^{(i)})^2 + \lambda ||\beta||_2,$$

with respect to β_0, to get that

$$\hat{\beta}_0 = \bar{y} - \beta_1 \bar{x}_1 - \beta_2 \bar{x}_2.$$

Now, let us redefine the objective function by dividing its first term by n – this does not change the optimal solution (λ is rescaled implicitly). Take the derivative of this rescaled objective function with respect to β_1 to get

$$\frac{1}{n} \sum_{i=1}^{n} (y^{(i)} - \beta_0 - \beta_1 x_1^{(i)} - \beta_2 x_2^{(i)})^2 (-x_1^{(i)}) + 2\lambda \beta_1 / ||\beta||_2,$$

where we should recall that $||\beta||_2 \equiv \sqrt{\beta_1^2 + \beta_2^2}$. Using the facts mentioned at the beginning, we get

$$-\overline{yx_1} - \beta_0 \bar{x}_1 - \beta_1 \bar{x}_1) + 2\lambda \beta_1 / ||\beta||_2 = 0.$$

where $\overline{yx_1} = \sum_i y^{(i)} x_1^{(i)}/n$. Analogously, deriving with respect to β_2 gives

$$-(\overline{yx_2} - \beta_0 \bar{x}_2 - \beta_2 \bar{x}_2) + 2\lambda\beta_2/||\beta||_2 = 0.$$

Now, add these two equations, and by using the following facts (verify them yourself),

$$\begin{aligned} \overline{yx_1} + \overline{yx_2} &= \bar{y} \\ \bar{x}_1 + \bar{x}_2 &= 0 \end{aligned}$$

along with $\hat{\beta}_0 = \bar{y} - \beta_1 \bar{x}_1 - \beta_2 \bar{x}_2$, get to the conclusion that $\hat{\beta}_1 + \hat{\beta}_2 = 0$.

8. See EX6.R.
9. See EX6.R.
10. See EX6.R.
11. See EX6.R.