STA 325: Homework 3 (97 points + 6 bonus)

DUE: 11:59pm, Nov 4 (on Sakai)
COVERAGE: ISL Chapters 6.2, 7

1. **[28 points]** Let's dig deeper into the shrinkage behavior of ridge regression and Lasso. Consider the simple linear regression model:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad i = 1, \cdots, n,$$

where β_0 and β_1 are model parameters. For simplicity, suppose the predictor x is standardized such that $\sum_{i=1}^n (x_i - \bar{x})^2 = 1$.

(a) **[5 points]** Recall the residual-sum-of-squares (RSS) criterion:

$$\mathrm{RSS}(\beta_0, \beta_1) = \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2 .$$

Show that the least-squares-estimators (LSE) for (β_0, β_1), which minimize $\mathrm{RSS}(\beta_0, \beta_1)$, are given by:

$$\hat{\beta}_0^{\mathrm{LS}} = \bar{y} - \hat{\beta}_1^{\mathrm{LS}} \bar{x}, \quad \hat{\beta}_1^{\mathrm{LS}} = \sum_{i=1}^n x_i y_i - n\bar{x}\bar{y}.$$

Hint: Set the derivative of $\mathrm{RSS}(\beta_0, \beta_1)$ *with respect to* β_0 *to zero, then solve for* β_0*. Plug this expression for* β_0 *into the derivative of* $\mathrm{RSS}(\beta_0, \beta_1)$ *with respect to* β_1*, then set to zero and solve for* β_1*.*

(b) **[5 points]** Consider the ridge regression estimators $(\hat{\beta}_{0,\lambda}^{\mathrm{R}}, \hat{\beta}_{1,\lambda}^{\mathrm{R}})$, which minimize the following optimization problem:

$$\min_{\beta_0, \beta_1} \left\{ \mathrm{RSS}(\beta_0, \beta_1) + \lambda \beta_1^2 \right\}.$$

Show that:

$$\hat{\beta}_{0,\lambda}^{\mathrm{R}} = \bar{y} - \hat{\beta}_{1,\lambda}^{\mathrm{R}} \bar{x}, \quad \hat{\beta}_{1,\lambda}^{\mathrm{R}} = \frac{\hat{\beta}_1^{\mathrm{LS}}}{1 + \lambda}.$$

(c) **[5 points]** Suppose $\lambda = 1$. Plot the ridge regression estimator $\hat{\beta}_{1,\lambda}^{\mathrm{R}}$ (which is shrunk) as a function of the least-squares estimator $\hat{\beta}_1^{\mathrm{LS}}$, for $\hat{\beta}_1^{\mathrm{LS}} \geq 0$. Comment on the shrinkage behavior of ridge regression. Does this plot give any insight on its ability to select important variables?

(d) **[BONUS 3 points]** Consider the Lasso estimators $(\hat{\beta}_{0,\lambda}^{\mathrm{L}}, \hat{\beta}_{1,\lambda}^{\mathrm{L}})$, which minimize the following optimization problem:

$$\min_{\beta_0, \beta_1} \left\{ \mathrm{RSS}(\beta_0, \beta_1) + \lambda |\beta_1| \right\}.$$

Suppose $\hat{\beta}_1^{\mathrm{LS}} \geq 0$. Show that:

$$\hat{\beta}_{0,\lambda}^{\mathrm{L}} = \bar{y} - \hat{\beta}_{1,\lambda}^{\mathrm{L}} \bar{x}, \quad \hat{\beta}_{1,\lambda}^{\mathrm{L}} = (\hat{\beta}_1^{\mathrm{LS}} - \lambda/2)_+ := \max\{\hat{\beta}_1^{\mathrm{LS}} - \lambda/2, 0\}. \tag{1}$$

Hint: The key challenge here is that $|\beta_1|$ *is not differentiable, so we need to generalize the notion of a derivative a bit. One can show that the Lasso estimators*

$(\hat{\beta}^{\text{L}}_{0,\lambda}, \hat{\beta}^{\text{L}}_{1,\lambda})$ *solve the two equations:*

$$\begin{aligned} -2\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i) &= 0 \\ -2\sum_{i=1}^{n} x_i(y_i - \beta_0 - \beta_1 x_i) + \lambda\, \partial\beta_1 &\ni 0, \end{aligned} \tag{2}$$

where $\partial\beta_1$ *is the so-called subdifferential of* $|\beta_1|$*:*

$$\partial\beta_1 = \begin{cases} -1, & \beta_1 < 0, \\ [-1, +1], & \beta_1 = 0, \\ +1, & \beta_1 > 0. \end{cases}$$

From (2)*, the Lasso estimator can be derived using the following two steps:*

- *Suppose the least-squares estimate* $\hat{\beta}^{\text{LS}}_1 > \lambda/2$*. What do the estimators in* (1) *simplify to? Do the simplified estimators solve* (2)*?*
- *Suppose the least-squares estimate* $\hat{\beta}^{\text{LS}}_1 \leq \lambda/2$*. What do the estimators in* (1) *simplify to? Do the simplified estimators solve* (2)*?*

(e) [**5 points**] Suppose $\lambda = 1$. Plot the lasso estimator $\hat{\beta}^{\text{L}}_{1,\lambda}$ (which is shrunk) as a function of the least-squares estimator $\hat{\beta}^{\text{LS}}_1$, for $\hat{\beta}^{\text{LS}}_1 \geq 0$. Comment on the shrinkage behavior of Lasso. Does this plot give any insight on its ability to select important variables?

(f) [**3 points**] Having used the squared-l_2 norm (part (b)) and the l_1-norm (part (d)), let's now try the so-called l_0-norm on β_1: $I(\beta_1 \neq 0)$. This new "norm" gives a value of 1 whenever β_1 is non-zero (i.e., the variable is active), and a value of 0 whenever β_1 equals zero (i.e., the variable is inert). Using this, the penalized regression problem becomes:

$$\min_{\beta_0, \beta_1} \left\{ \text{RSS}(\beta_0, \beta_1) + \lambda I(\beta_1 \neq 0) \right\}.$$

Reformulate this penalized problem into its constrained form with radius s (see Equations (6.8) or (6.9) in ISL). We've seen this constrained problem before for variable selection. What is it? Explain.

(g) [**BONUS 3 points**] Suppose $\hat{\beta}^{\text{LS}}_1 \geq 0$. Show that the estimators $(\hat{\beta}^{\text{S}}_{1,\lambda}, \hat{\beta}^{\text{S}}_{0,\lambda})$ which minimize the constrained problem in part (f) are given by:

$$\hat{\beta}^{\text{S}}_{0,\lambda} = \bar{y} - \hat{\beta}^{\text{S}}_{1,\lambda}\bar{x}, \quad \hat{\beta}^{\text{S}}_{1,\lambda} = \hat{\beta}^{\text{LS}}_1 \cdot I(\hat{\beta}^{\text{LS}}_1 \geq \sqrt{\lambda}).$$

(h) [**5 points**] Suppose $\lambda = 1$. Plot the estimator $\hat{\beta}^{\text{S}}_{1,\lambda}$ as a function of the least-squares estimator $\hat{\beta}^{\text{LS}}_1$, for $\hat{\beta}^{\text{LS}}_1 \geq 0$. Comment on the shrinkage behavior of this method. Does this plot give any insight on its ability to select important variables?

2. **[21 points]** State whether each of the following statements are TRUE or FALSE. Briefly justify why in a couple of sentences.

   (a) Least-squares estimation should be used over ridge regression when there is high multi-collinearity in the data.

   (b) Lasso should be used over ridge regression when we know a priori that only a small handful of predictors are active.

   (c) Piecewise polynomial models can be discontinuous without constraints.

   (d) For cubic splines, the variance of the fitted model decreases as more knots are added.

   (e) Splines provide greater model flexibility in regions with many knots.

   (f) A model with high degrees-of-freedom implies a greater bias in its fit.

3. **[21 points]** Consider a *quartic spline* model with distinct knots ξ_k, $k = 1, \cdots, K$. A quartic spline satisfies two properties: (i) it is a quartic (i.e., degree-4) polynomial between any two neighboring knots, and (ii) it has continuous derivatives of up to order 3 at each knot. Note that property (ii) includes derivatives of order 0, meaning the quartic spline should be continuous at knots.

   (a) **[5 points]** Write out the full model specification for the quartic spline, including model parameters and basis functions (see Equation (7.9) in ISL). How many degrees-of-freedom (d.f.s) are in your model?

   (b) **[6 points]** Prove that properties (i) and (ii) hold for your model in (a).

   (c) **[5 points]** Suppose you present this model to your boss. Her initial reaction was that, while she likes the flexibility of your model, she is afraid that this comes at a huge computational cost. She is worried that model fitting (e.g., estimation, prediction, computing confidence intervals) will be too time-consuming for large datasets. Because of this, she suggests you try a simpler linear model instead, which can be fit efficiently. Should you agree with her? Explain why or why not.

   (d) **[5 points]** Suppose, after some discussion, she begrudgingly adopts your quartic spline model. After seeing your R output, however, she complains that your fit requires 16 d.f.s, which she believes to be too many. She claims that, with that many d.f.s, a degree-15 polynomial model can be fit, which can capture higher order effects than your quartic model. Should you agree with her? Explain why or why not.

4. [**30 points**] ISL Chapter 7, Exercise 9.