# Homework 2

E6690: Statistical Learning for Bio & Info Systems

**P1.** 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) (2pt) Write out the ridge regression optimization problem in this setting.

(b) (2pt) Argue that in this setting, the ridge coefficient estimates satisfy $\hat{\beta}_1 = \hat{\beta}_2$.

(c) (2pt) Write out the lasso optimization problem in this setting.

(d) (4pt) 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.

**P2.** Consider a simple regression with $n = p$, and $\boldsymbol{X}$ a diagonal matrix with 1's on the diagonal and 0's in all off-diagonal elements. To simplify the problem further, assume that we are performing regression without an intercept. In this case, the usual least squares problem simplifies to finding $\beta_1, \ldots, \beta_p$ that minimize

$$\sum_{j=1}^{p} (y_j - \beta_j)^2.$$

The least squares solution is given by $\hat{\beta}_j = y_j$.
And in this setting, ridge regression amounts to finding $\beta_1, \ldots, \beta_p$ such that

$$\sum_{j=1}^{p} (y_j - \beta_j)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \tag{1}$$

is minimized, and the lasso amounts to finding the coefficients such that

$$\sum_{j=1}^{p} (y_j - \beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \tag{2}$$

is minimized. One can show that in this setting, the ridge regression estimates take the form

$$\hat{\beta}_j^R = y_j/(1 + \lambda), \tag{3}$$

and the lasso estimates take the form

$$\hat{\beta}_j^L = \begin{cases} y_j - \lambda/2 & \text{if } y_j > \lambda/2; \\ y_j + \lambda/2 & \text{if } y_j < -\lambda/2; \\ 0 & \text{if } |y_j| \leq \lambda/2. \end{cases} \tag{4}$$

(a) (5pt) Consider $(1)$ with $p = 1$. For some choice of $y_1$ and $\lambda > 0$, plot $(1)$ as a function of $\beta_1$. Your plot should confirm that $(1)$ is solved by $(3)$.

(b) (5pt) Consider $(2)$ with $p = 1$. For some choice of $y_1$ and $\lambda > 0$, plot $(2)$ as a function of $\beta_1$. Your plot should confirm that $(2)$ is solved by $(4)$.

**P3.** In this problem, we will derive the Bayesian connection to the lasso and ridge regression.

(a) (2pt) Suppose that $y_i = \beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j + \epsilon_i$ where $\epsilon_1, \ldots, \epsilon_n$ are independent and identically distributed from a $\mathcal{N}(0, \sigma^2)$ distribution. Write out the likelihood for the data.

(b) (3pt) Assume the following prior for $\boldsymbol{\beta} : \beta_1, \ldots, \beta_p$ are independent and identically distributed according to a double-exponential distribution with mean $0$ and common scale parameter $b$; i.e, $p(\boldsymbol{\beta}) = \frac{1}{2b}\exp(-|\boldsymbol{\beta}|/b)$ and $|\boldsymbol{\beta}| = \sum_{j=1}^{p} |\beta_j|$. Write out the posterior for $\boldsymbol{\beta}$ in this setting.

(c) (5pt) Argue that the lasso estimate is the mode for $\beta$ under this posterior distribution.

(d) (5pt) Now assume that the following prior for $\beta : \beta_1, \ldots, \beta_p$ are independent and identically distributed according to a normal distribution with mean $0$ and variance $c$. write out the posterior for $\beta$ in this setting.

(e) (5pt) Argue that the ridge regression estimate is both the mode and the mean for $\boldsymbol{\beta}$ under this posterior distribution.

**P4.** In this exercise, we will generate simulated data, and will then use this data to perform best subset selection.

(a) (2pt) Use the `rnorm()` function to generate a predictor $X$ of length $n = 100$, as well as a noise vector $\epsilon$ of length $n = 100$.

(b) (3pt) Generate a response vector $Y$ of length $n = 100$ according to the model

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \epsilon,$$

where $\beta_0$, $\beta_1$, $\beta_2$, and $\beta_3$ are constants of your choice.

(c) (5pt) Use the `regsubsets()` function to perform best subset selection in order to choose the best model containing the predictors $X, X^2, \ldots, X^{10}$. What is the best model obtained according to $C_p$, BIC, and adjusted $R^2$? Show some plots to provide evidence for your answer, and report the coefficients of the best model obtained. Note you will need to use the `data.frame()` function to create a single data set containing both $X$ and $Y$.

(d) (5pt) Repeat (c), using forward stepwise selection and also using backwards stepwise selection. How does your answer compare to the result in (c)?

(e) (5pt) Now fit a lasso model to the simulated data, again using $X, X^2, \ldots, X^{10}$ as predictors. Use cross-validation to select the optimal value of $\lambda$. Create plots of the cross-validation error as a function of $\lambda$. Report the resulting coefficient estimates, and discuss the results obtained.

(f) (5pt) Now generate a response vector $Y$ according to the model

$$Y = \beta_0 + \beta_7 X^7 + \epsilon,$$

and perform best subset selection and the lasso. Discuss the results obtained.

**P5.** In this exercise, we will predict the number of applications received using the other variables in the `college` data set.

(a) (2pt) Split the data set into a training set and a test set.

(b) (3pt) Fit a linear model using least squares on the training set, and report the test error obtained.

(c) (5pt) Fit a ridge regression model on the training set, with $\lambda$ chosen by cross-validation. Report the test error obtained.

(d) (5pt) Fit a lasso model on the training set, with $\lambda$ chosen by cross-validation. Report the test error obtained, along with the number of non-zero coefficient estimates.

**P6.** We have seen that as the number of features used in a model increases, the training error will necessarily decrease, but the test error may not. We will now explore this in a simulated data set.

(a) (2pt) Generate a data set with $p = 20$ features, $n = 1000$ observations, and an associated quantitative response vector generated according to the model

$$Y = X\beta + \epsilon,$$

where $\beta$ has some elements that are exactly equal to zero.

(b) (2pt) Split your data set into a training set containing $100$ observations and a test set containing $900$ observations.

(c) (2pt) Perform best subset selection on the training set, and plot the training set MSE associated with the best model of each size.

(d) (2pt) Plot the test set MSE associated with the best model of each size.

(e) (2pt) For which model size does the test set MSE take on its minimum value? Comment on your results. If it takes on its minimum value for a model containing only an intercept or a model containing all of the features, then play around with the way that you are generating the data in (a) until you come up with a scenario in which the test set MSE is minimized for an intermediate model size.

(f) (2pt) How does the model at which the test set MSE is minimized compare to the true model used to generate the data? Comment on the coefficient values.

(g) (3pt) Create a plot displaying $\sqrt{\sum_{j=1}^{p}(\beta_j - \hat{\beta}_j^r)^2}$ for a range of values of $r$, where $\hat{\beta}_j^r$ is the $j$th coefficient estimate for the best model containing $r$ coefficients. Comment on what you observe. How does this compare to the test MSE plot from (d)?