DS 301: Homework 7
Due: March 30, 2022 on Canvas by 11:59 pm (CT)

Instructions: Homework is to be submitted on Canvas by the deadline stated above. Please clearly print your name and student ID number on your HW.

Show your work (including calculations) to receive full credit. Please work hard to make your submission as readable as you possibly can - **this means no raw `R` code or raw `R` output** (unless it is asked for specifically or needed for clarity).

Code should be submitted with your homework as a separate file (for example, an R file, text file or .Rmd are all acceptable). You should mark sections of the code that correspond to different homework problems using comments (e.g. `##### Problem 1 #####`).

Problem 1: Concept Review

a. For the lasso regression model, what is the bias of $\hat{\beta}_{\text{lasso}}$ when $\lambda = 0$?

b. For the lasso regression model, what is the variance of $\hat{\beta}_{\text{lasso}}$ when $\lambda = \infty$?

c. For ridge regression, as we increases λ from 0, draw a single plot (by hand is fine) showing how the following quantities will behave: (1) training MSE, (2) test MSE, (3) variance, (4) (squared) bias, and (5) the irreducible error.

d. For the lasso regression model, we can estimate the regression coefficients in a linear regression model by minimizing

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

This is **exactly equivalent** to the formulation presented in class, except now instead of a penalty λ we have a constraint controlled by s. As we increase s from 0, draw a single plot showing how the following quantities will behave: (1) training MSE, (2) test MSE, (3) variance, (4) (squared) bias, and (5) the irreducible error.

Problem 2: Simulation Studies

a. Use the `rnorm()` function to generate a predictor X of length $n = 100$, as well as error vector ϵ of length $n = 100$. Assume that ϵ has variance 1.

b. 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 β_3 are constants of your choice.

c. Implement best subset selection in order to choose the best model containing the predictors $X, X^2, \ldots, X^{10}$. List all the models subset selection searches through of size 9. How many models are there? Calculate RSS/AIC/BIC for each model. Use this as an example to justify that when the models are of the same size, AIC/BIC will lead you to pick the same optimal model. Note: you will need to use the `data.frame()` function to create a single data set containing both X and Y.

d. Consider all models $M_1, \ldots, M_{10}$. What is the best model obtained according to BIC and adjusted R^2? Report the coefficients of the best model obtained.

e. Repeat model selection using forward selection and also using backward selection. Report the best models obtains according to BIC and adjusted R^2 for both approaches. How does your answer compared to the results in part (d)?

f. Now fit a lasso model to the simulated data, again using $X, X^2, \ldots, X^{10}$ as predictors. Use 10-fold cross-validation to select the optimal value of λ. Present a plot of the cross-validation error as a function of λ. Report the resulting coefficient estimates and discuss the results obtained.

g. 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 (again using predictors $X, X^2, \ldots, X^{10}$). Discuss the results obtained.

Problem 3: Regularized Regression Models

For this problem, we will continue with the `Hitters` example from lecture. Our aim is to predict the salary of baseball players based on their career statistics.

a. We will start with a little data cleaning. We'll also split the data into a training and test set. So that we all get the same results, please use the following code:

```
library(ISLR2)
Hitters = na.omit(Hitters)
n = nrow(Hitters) #there are 263 observations
x = model.matrix(Salary ~.,data=Hitters)[,-1]  #19 predictors
Y = Hitters$Salary
set.seed(1)
train = sample(1:nrow(x), nrow(x)/2)
test=(-train)
Y.test = Y[test]
```

b. Replicate the example we had in class and fit a ridge regression model on a grid of values (on the training set only). What are the estimates of our regression coefficients when $\lambda = 0.013$? Report them here.

c. What are the estimates of our regression coefficients when $\lambda = 10^{10}$? Report them here. How do these coefficients compare to the ones obtained in (b)?

d. Use your results in part (b) and (c) to explain what happens to the regression coefficients as λ increases.

e. Report the l_2 norm (defined as $\sum_{j=1}^{p} \beta_j^2$) when $\lambda = 0.013$. Do you expect this to be larger or smaller then the l_2 norm with $\lambda = 10^{10}$? Explain.

f. Use 10-fold CV to obtain the optimal λ. Present a plot of the cross-validation error as a function of λ. Report that value here and call it $\lambda_{\text{min}}^{\text{ridge}}$.

g. Naturally, if we had taken a different training/test set or a different set of folds to carry out cross-validation, our optimal λ and therefore test error would change. An alternative is to select λ using the *one-standard error rule*. The idea is, instead of picking the λ that produces the smallest CV error, we pick the model whose CV error is within one standard error of the lowest point on the curve you produced in part (f). The intention is to produce a more **parimonious** and **robust** model. The `glmnet` function does all of this hard work for you and we can extract the λ based on this rule using the following code: `cv.out$lambda.1se` (assuming your `cv.glmnet` object is named `cv.out`). Report your that λ here and call it $\lambda_{\text{1se}}^{\text{ridge}}$.

h. Fit a lasso regression model. Replicate the example we had in class to obtain the the optimal λ using 10-fold CV. Report that value here and call it $\lambda_{\text{min}}^{\text{lasso}}$. Also report the optimal λ using the smallest standard error rule and called it $\lambda_{\text{1se}}^{\text{lasso}}$.

i. You now have 4 values for the tuning parameter:

$$\lambda_{\text{min}}^{\text{ridge}}, \lambda_{\text{1se}}^{\text{ridge}}, \lambda_{\text{min}}^{\text{lasso}}, \lambda_{\text{1se}}^{\text{lasso}}.$$

Now evaluate the ridge regression models on your test set using $\lambda = \lambda_{\text{min}}^{\text{ridge}}$ and $\lambda = \lambda_{\text{1se}}^{\text{ridge}}$. Evaluate the lasso models on your test set using $\lambda_{\text{min}}^{\text{lasso}}$ and $\lambda_{\text{1se}}^{\text{lasso}}$. Compare the obtained test errors and report them here. Which model performs the best in terms of prediction? Do you have any intuition as to why?

j. Report the coefficient estimates coming from ridge using $\lambda_{\text{min}}^{\text{ridge}}$ and $\lambda_{\text{1se}}^{\text{ridge}}$ and likewise for the lasso models. How do the ridge regression estimates compare to those from the lasso? How do the coefficient estimates from using λ_{min} compare to those from the one-standard error rule?

k. If you were to make a recommendation to an upcoming baseball player who wants to make it big in the major leagues, what handful of features would you tell this player to focus on?

Problem 4: Comparing predictive models

For this problem, we will use the `College` data set in the `ISLR2` R package. Our aim is to predict the number of applications (`Apps`) received using the other variables in the dataset.

a. Split the data set into a training and a test set.

b. Fit a least squares linear model (using all predictors) on the training set. Carry out model selection on the training set only and report the test MSE obtained. You can choose between subset/forward/ or backward selection.

c. Fit a ridge regression model (using all predictors) on the training set. The function `glmnet`, by default, internally scales the predictor variables so that they will have standard deviation 1. Explain why this scaling is necessary when implementing regularized models.

d. Find an optimal λ for the ridge regression model on the training set by using 10-fold cross-validation. Report the optimal λ here.

e. Using that optimal λ, evaluate your trained ridge regression model on the test set. Report the test MSE obtained. Is there an improvement over the model from part (b)?

f. Fit a lasso regression model on the training set. Find the optimal lambda using 10-fold cross-validation. Report the optimal λ, the number of non-zero coefficient estimates, and the test MSE obtained.

g. From the 3 models, report your final model. Comment on your results. Is there much difference among the test errors resulting from these 3 approaches?