DS 303 Homework 7
Due: Oct. 16, 2023 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` output or code** (unless it is asked for specifically or needed for clarity).

Code should be submitted with your homework as a separate file (for example, a .R file, text file, Word 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) Explain in plain language (using limited statistics terminology) why lasso can set some of the regression coefficients to be 0 exactly, while ridge regression cannot. You may include a figure if that is helpful.

(b) Suppose we estimate the regression coefficients in a linear regression model by minimizing

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}\beta_j^2$$

for a particular value of λ. For parts (a) through (e), indicate which of i. through v. is correct. Justify your answer.

a. As we increase λ from 0, the training MSE will:

i. Increase initially, and then eventually start decreasing in an inverted U shape.

ii. Decrease initially, and then eventually start increasing in a U shape.

iii. Steadily increase.

iv. Steadily decrease.

v. Remain constant.

b. Repeat (a) for test MSE.

c. Repeat (a) for variance.

d. Repeat (a) for (squared) bias.

e. Repeat (a) for irreducible error.

Problem 2: 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) Fit a ridge regression model. Replicate the example we had in class to obtain the the optimal λ that minimizes the 10-fold CV. Present a plot of the cross-validation error as a function of λ. Report that value here and call it $\lambda^{\text{ridge}}_{\text{min}}$.

(c) 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 (b). The intention is to produce a more **parimonious** 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 $\lambda^{\text{ridge}}_{\text{1se}}$.

(d) Fit a lasso regression model. Replicate the example we had in class to obtain the the optimal λ that minimizes the 10-fold CV. Present a plot of the cross-validation error as a function of λ. Report that value here and call it $\lambda^{\text{lasso}}_{\text{min}}$.

(e) For lasso, report the optimal λ using the smallest standard error rule and called it $\lambda^{\text{lasso}}_{\text{1se}}$.

(f) Evaluate the ridge regression models on your test set using $\lambda = \lambda^{\text{ridge}}_{\text{min}}$ and $\lambda = \lambda^{\text{ridge}}_{\text{1se}}$. Evaluate the lasso models on your test set using $\lambda^{\text{lasso}}_{\text{min}}$ and $\lambda^{\text{lasso}}_{\text{1se}}$. Report the test MSEs from these 4 models.

(g) Report the coefficient estimates from ridge using $\lambda^{\text{ridge}}_{\text{min}}$ and $\lambda^{\text{ridge}}_{\text{1se}}$ 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? Discuss what you observe.

(h) Train and implement elastic net. Report the optimal values for α and λ that produce the smallest 10-fold cross-validation error. Call these values α^{enet} and λ^{enet}.

(i) Evaluate the elastic net on your test set using the optimal values for α^{enet} and λ^{enet}.

(j) Which model performs the best in terms of prediction? Explain any intuition as to why.

(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 3: Bootstrap

We will work with the `Boston` housing data set from the `ISLR2` library. The first part of this problem is similar to our in-class activity on Bootstrap.

(a) Based on this data set, provide an estimate for the population mean of `medv`. Call this estimate $\hat{\mu}$.

(b) Provide an estimate of the standard error of $\hat{\mu}$ using an analytical formula. Interpret this result.

(c) Now the estimate the standard error $\hat{\mu}$ using the bootstrap. How does this compare to your answer from (b)?

(d) Using bootstrap, provide a 95% confidence interval for the mean of `medv`. Compare it to results using analytical formulas.

(e) Based on this data set, provide an estimate $\hat{\mu}_{\text{med}}$ for the median value of `medv`.

(f) We would like to estimate the standard error of $\hat{\mu}_{\text{med}}$. Since there is no simple formula for computing the standard error of the median, use bootstrap. Comment on your findings.

(g) Based on this data set, provide an estimate $\hat{\mu}_{0.1}$, the 10th percentile of `medv`.

(h) Use bootstrap to estimate the standard error of $\hat{\mu}_{0.1}$. Comment on your findings.

Problem 4: Properties of Bootstrap

(a) What is the probability that the first bootstrap observation is the jth observation from the original sample? Justify your answer.

(b) What is the probability that the first bootstrap observation is *not* the jth observation from the original sample? Justify your answer.

(c) What is the probability that the jth observation from the original sample is *not* in the bootstrap sample?

(d) When $n = 5$, what is the probability that the jth observation is in the bootstrap sample?

(e) When $n = 100$, what is the probability that the jth observation is in the bootstrap sample?

(f) When $n = 10,000$, what is the probability that the jth observation is in the bootstrap sample?

(g) Create a plot (in R) that displays, for each integer value of n from 1 to 100,000, the probability that the jth observation is in the bootstrap sample. Comment on what you observe.

(h) Let's investigate this numerically. What is the probability that the jth observation is in a bootstrap sample of size $n = 100$? Suppose $j = 5$. Repeatedly create bootstrap samples, and each time we record whether or not the fifth observation is contained in the bootstrap sample. The following code may help get you started:

```
results <- rep(NA, 10000)
for(i in 1:10000){
      results[i] <- sum(sample(1:100, rep=TRUE) == 5) > 0
}
```

Comment on your findings.

End of assignment