Summer Poissonnier
April 3, 2021
CAP 4773
Assignment 4

Homework 4: Due Thursday, April 15 at 11:59pm

There are four parts to this homework assignment, each with multiple questions. Please insert answers under corresponding questions, then save the document as a pdf and upload it to CANVAS. *Providing your R code is not required, but it may be helpful when assigning partial credit.*

We will again be using the **`College`** dataset in the **`ISLR`** package.

1) First, we will use best subset, forward, and backward selection to perform feature selection and construct models for predicting a college's out-of-state tuition from subsets of variables in **`College`**.

   a. (5 points) Using adjusted R^2 to estimate test MSE, how many features are in the best models for best subset, forward, and backward selection?

      Best subset selection: **15**

      Forward selection: **15**

      Backward selection: **15**

   b. (5 points) Using Mallows' C_p to estimate test MSE, how many features are in the best models for best subset, forward, and backward selection?

      Best subset selection: **14**

      Forward selection: **14**

      Backward selection: **14**

   c. (5 points) Using BIC to estimate test MSE, how many features are in the best models for best subset, forward, and backward selection?

      Best subset selection: **10**

      Forward selection: **13**

      Backward selection: **10**

d. (5 points) Based on your knowledge of these feature selection algorithms, which do you believe is best to use in this scenario? Briefly explain your reasoning.

Based on what I know, I would say the best algorithm is the best subset algorithm. I believe this because the best subset algorithm compares all 2^p models.

e. (5 points) Examine the output of the feature selection algorithm from 1d. Which feature is present in all models? Briefly describe what this feature represents.

The feature that is present in all models is: PrivateYes. This feature represents if the college is private or public.

f. (5 points) Based on your findings from 1a – c and your answer to 1d, provide **three** estimates for the number of features in the overall "best" model. Briefly explain your reasoning.

The estimates for the overall best model are: 15, 14, and 10. This is because the best model is best subset and the estimates for the features in each model are as stated above.

2) Next, we will use ridge regression to perform regularization and construct a model for predicting a college's out-of-state tuition from the other variables in `College`. ***Before performing this analysis, set the seed to "1", and then randomly split the data into training and test datasets.***

a. (5 points) What value of λ yields the smallest $\mathrm{CV}_{(10)}$ for ridge regression? ***Remember to set the seed to "1" again before performing this analysis.***

The value of lambda that yields the smallest CV(10) for ridge regression is: 665.1587

b. (5 points) Compute the test MSE for ridge regression with the value of λ from 2a.

The MSE for ridge regression is: 4020611.

c. (5 points) Use your answer from 2b to estimate the average difference ***in dollars*** between observed and predicted out-of-state tuition.

The average difference in dollars between observed and predicted out-of-state tuition is: $4020611.

d. (10 points) Are any of the features removed from the model? If so, which ones? If not, why do you think all features are included in the model?

None of the features are removed from the model. I think all of the features are included in the model because none of them are 0.

3) Third, we will use lasso to perform regularization and construct a model for predicting a college's out-of-state tuition from the other variables in `College`. ***Note that we will be using the same training and test data as for question 2 above.***

a. (5 points) What value of λ yields the smallest $\mathrm{CV}_{(10)}$ for lasso? ***Remember to set the seed to "1" again before performing this analysis.***

The value of lambda that yields the smallest CV(10) for lasso is: 83.93351.

b. (5 points) Compute the test MSE for lasso with the value of λ from 3a.

The MSE for lasso is: 4073819.

c. (5 points) Use your answer from 3b to estimate the average difference ***in dollars*** between observed and predicted out-of-state tuition.

The average difference in dollars between observed and predicted out-of-state tuition is: $4073819.

d. (10 points) Are any of the features removed from the model? If so, which ones? If not, why do you think all features are included in the model?

The features that are removed from the model are: Apps, Enroll, F.undergrad, and Books.

4) Last, we will compare our findings from parts 1-3 above.

a. (5 points) Which of the four statistics used in this assignment (three in part 1, and one in parts 2 and 3) likely provides the best estimate of test MSE? Briefly justify your reasoning.

Best subset/feature selection provides the best estimate of test MSE because a model is fit to different combinations of features and the one with the lowest test error is selected.

b. (5 points) Which of the two regularization techniques has better performance on the test data in this example? Briefly justify your answer based on findings from your analysis.

Lasso has a better performance on the test data in this example. This model has a smaller lambda value.

c. (5 points) Propose an explanation for why you believe that the regularization technique from 4b performed better on the test data in this example. *Hint: Consider your answers to 2d and 3d.*

This is because lasso in the College dataset tested all features and then got rid of the features that were not necessary to the model.

d. (5 points) What is the advantage of feature selection and regularization over the traditional linear regression that we performed in HW 3? *Hint: Think about question 1e in HW 3.*

The advantage of feature selection and regularization over the traditional linear regression is that these algorithms only include features that are linearly related to the test data.

Code:

```
install.packages("leaps")
library(leaps)

#Question 1 a-c
bestsub <- regsubsets(Outstate ~. , data = College, nvmax = 17)
forward <- regsubsets(Outstate ~. , data = College,
                      nvmax=17, method = "forward")
backward <- regsubsets(Outstate ~. , data = College,
                       nvmax=17, method = "backward")

bestsub.sum <- summary(bestsub)
forward.sum <- summary(forward)
backward.sum <- summary (backward)

names(bestsub.sum)

c(which.max(bestsub.sum$adjr2),
  which.max(forward.sum$adjr2), which.max(backward.sum$adjr2))

c(which.min(bestsub.sum$cp),
  which.min(forward.sum$cp), which.min(backward.sum$cp))

c(which.min(bestsub.sum$bic),
  which.min(forward.sum$bic), which.min(backward.sum$bic))

#Question 1d
coef(bestsub, 15)
coef(bestsub,14)
coef(bestsub,10)

#Question 2
set.seed(1)
train <- sample(c(TRUE, FALSE),nrow(College), rep=TRUE)
test <- (!train)
install.packages("glmnet")
library(glmnet)
features <- model.matrix(Outstate~. , data =College)[,-1]
set.seed(1)
cv.out <- cv.glmnet(features[train,],
                    College$Outstate[train], alpha=0)

#Question 2a
set.seed(1)
bestlam <- cv.out$lambda.min
bestlam
```

```
#Question 2b
ridge <- glmnet(x = features[train,],y = College$Outstate[train],
            alpha = 0, lambda = bestlam)
ridge.pred <- predict(ridge, s = bestlam, newx=features[test,])
mean((ridge.pred - College$Outstate[test])^2)
coef(ridge)

#Question 3a
set.seed(1)
cv.out <- cv.glmnet(features[train,],
                College$Outstate[train], alpha=1)
bestlam <- cv.out$lambda.min
bestlam

#Question 3b
lasso <- glmnet(x = features[train,],y = College$Outstate[train],
            alpha = 1, lambda = bestlam)
lasso.pred <- predict(lasso,s=bestlam,newx=features[test,])
mean((lasso.pred - College$Outstate[test])^2)
coef(lasso)
```