HW 7

Ian Douglas

4/7/2019

library(glmnet)

## Loading required package: Matrix

## Loading required package: foreach

## Loaded glmnet 2.0-16

library(ISLR)  
attach(College)

# Part A

set.seed(25)  
#randomize the test and train sets  
i <- sample(1:nrow(College), size=.5\*nrow(College))  
train <- College[i,]  
test <- College[-i,]

# Part B

#OLS models with numeric predictors only  
trainmod1 <- lm(Apps ~ .-Private, data = train)  
#Compute prediction error on test set  
MSPE.test <- mean(((predict(trainmod1, newdata=test) - test$Apps))^2)  
MSPE.test

## [1] 1425733

# Part C

#Scale predictors before running ridge regression  
X <- apply(College[3:ncol(College)], 2,   
 function(x) x/mean((x-mean(x))^2))  
Y <- as.matrix(College[,2])  
colnames(X) <- names(College)[3:ncol(College)]  
colnames(Y) <- names(College)[2]  
X.train.scaled <- X[i,]  
X.test.scaled <- X[-i,]  
#run the Ridge reression (alpha = 0), and use the cv.glmnet  
#function to simultaneously select the optimal lambda parameter  
ridge1 <- cv.glmnet(x = X.train.scaled, y = Y[i,],  
 family = "gaussian", alpha = 0)  
#Compute MSPE on test set  
MSPE.test.ridge <- mean(  
 ((predict(ridge1, newx = X.test.scaled, s = "lambda.min")  
 - Y[-i,]))^2  
 )  
MSPE.test.ridge

## [1] 2185850

# Part D

#repeat part C with the lasso  
lasso1 <- cv.glmnet(x = X.train.scaled, y = Y[i,],  
 family = "gaussian", alpha = 1)  
#Compute MSPE on test set  
MSPE.test.lasso <- mean(  
 ((predict(lasso1, newx = X[-i,], s = "lambda.min")  
 - Y[-i,]))^2  
 )  
MSPE.test.lasso

## [1] 1472055

# Part E

#Compute R^2 for each  
OLS <- summary(trainmod1)$r.squared  
ridge <- ridge1$glmnet.fit$dev.ratio[which(ridge1$glmnet.fit$lambda == ridge1$lambda.min)]  
lasso <- lasso1$glmnet.fit$dev.ratio[which(lasso1$glmnet.fit$lambda == lasso1$lambda.min)]  
R2 <- cbind(OLS,ridge,lasso)  
colnames(R2) <- c("OLS","ridge","lasso")  
R2

## OLS ridge lasso  
## [1,] 0.9234454 0.9102917 0.9223007

# Part F

The for the Lasso regression was comparable to that of the OLS model, and both were closer to 1 than that of the ridge model. This pattern was also observed for the mean squared prediction error across the three models, suggesting that Lasso regression is the best approach when taking into account both fit on the training set in addition to the prediction error on the test dataset.