Homework_1_mp3653

Matthew Perrotta

March 1, 2019

Load Packages

```
library(tidyverse)
library(ISLR)
library(glmnet)
library(caret)
library(corrplot)
library(plotmo)
library(boot)
library(pls)
```

Load Data

```
test = read.csv('./data/solubility_test.csv')
train = read.csv('./data/solubility_train.csv')
```

```
# Validation Control
ctrl1 <- trainControl(method = "repeatedcv", number = 10, repeats = 5)

# Train Predictor Matrix
trX = model.matrix(Solubility~., train)[, -1]
# Train Response
trY = train$Solubility

# Test Predictor Matrix
teX = model.matrix(Solubility~., test)[, -1]
# Test Response
teY = test$Solubility
```

Q1 Linear Model

```
set.seed(1)
lm.fit <- train(trX, trY,
                method = "lm",
                trControl = ctrl1)

lm.pred <- predict(lm.fit$finalModel, newdata = data.frame(teX))

mean((lm.pred - teY)^2)
```

MSE = 0.5558898

Q2 Ridge Regression Model

```r
set.seed(1)
ridge.fit = train(trX, trY,
              method = "glmnet",
              tuneGrid = expand.grid(alpha = 0,
                                     lambda = exp(seq(-1, 10, length = 100))),
              trControl = ctrl1)

best.lambda.ridge <- ridge.fit$bestTune$lambda

ridge.pred = predict(ridge.fit$finalModel, s = best.lambda.ridge, newx = teX)

mean((ridge.pred - teY)^2)
```

MSE = 0.545737

Q3 Lasso Regression Model

```r
set.seed(1)
lasso.fit = train(trX, trY,
              method = "glmnet",
              tuneGrid = expand.grid(alpha = 1,
                                     lambda = exp(seq(-10, 10, length = 200))),
              trControl = ctrl1)

best.lambda.lasso = lasso.fit$bestTune$lambda

lasso.pred = predict(lasso.fit$finalModel, s = best.lambda.lasso, newx = teX)

mean((lasso.pred - teY)^2)
```

MSE = 0.4987333

number of non-zero coefficients

```r
lasso.coef = predict(lasso.fit$finalModel, s = best.lambda.lasso, type = 'coefficients')
length(lasso.coef[lasso.coef != 0])
```

```
## <sparse>[ <logic> ] : .M.sub.i.logical() maybe inefficient
```

The number of nonzero coefficients is 144

Q4 PCR Model

```r
set.seed(1)
pcr.fit <- train(trX, trY,
                 method = "pcr",
                 tuneLength = 228,
                 trControl = ctrl1,
                 scale = T)

pcr.pred <- predict(pcr.fit$finalModel, newdata = teX,
                       ncomp = pcr.fit$bestTune$ncomp)

pcr.fit$bestTune$ncomp

mean((pcr.pred - teY)^2)
```

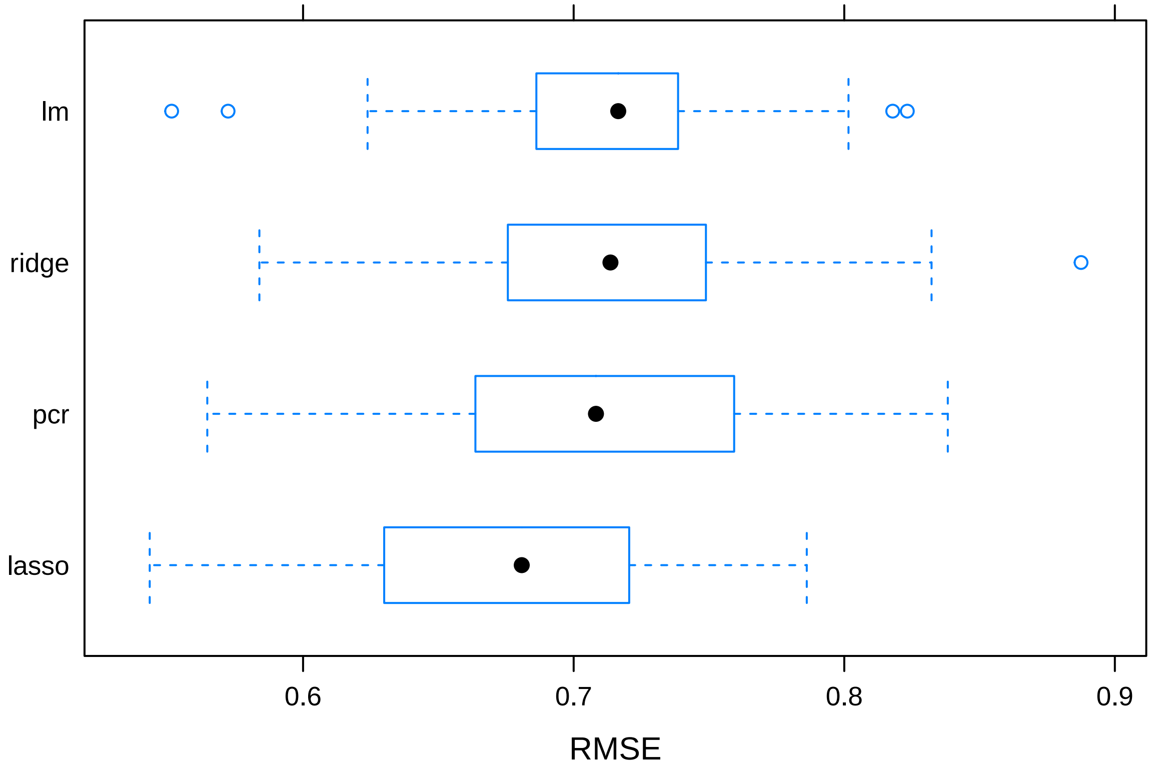
MSE = 0.5490447

value of M = 158

Q5 Discussion

```r
resamp <- resamples(list(lasso = lasso.fit,
                         ridge = ridge.fit,
                         pcr = pcr.fit,
                         lm = lm.fit))

bwplot(resamp, metric = "RMSE")
```



The RMSE value is smallest when using lasso regression compared to the other 3 models, with ridge, pcr and lm having increasing RMSE values respectively. Ridge regression assumes that all predictors are necessary, while lasso assumes that some coefficients are equal to zero. Lasso, with the smallest RMSE value, demonstrates that some of the cofficients for the predictors are truly zero.