Data Mining and Statistic Learning 3rd Homework

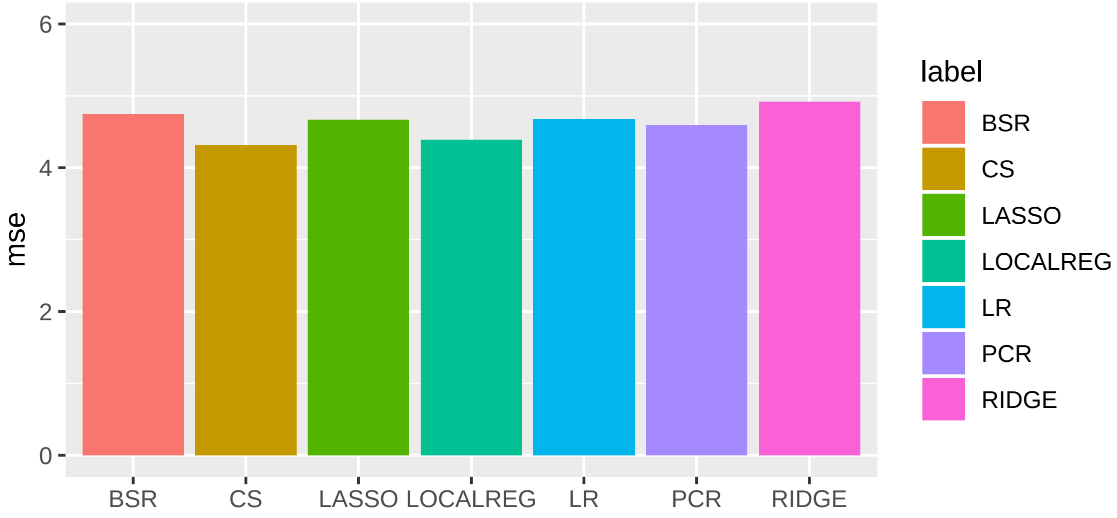
Daifeng Qi, 201657015

2019.6.11

```
# packages
library(dplyr)
library(leaps)
library (glmnet)
library(nnet)
library (pls)
library(splines)
library(ggplot2)
library(gam)
```

```
setwd('D://R/dmsl/dmsl_3')
dat <- read.csv('abalone.csv',stringsAsFactors = F)
# data preparation
dat <- as_tibble(dat)
train <- dat[1:2000,]
test_X <- dat[2001:4177,1:8]
test_y <- unlist(dat[2001:4177,9])
# LR
lr <- lm(Rings~.,data = train)
lr_pre <- predict(lr,newdata = test_X)
mse_lr <- sum((lr_pre-test_y)^2)/2177
# Best Subset regression
bsr <- regsubsets(Rings~., data=train ,nvmax =19)
bsrbic <- summary(bsr)$bic
bsr_coe <- coef(bsr,which(bsrbic == min(bsrbic)))
# Get best subset
bsr_mat <- cbind(rep(1,2177),test_X[['Sex']]=='I',test_X[['Diameter']],
                 test_X[['Height']],test_X[['Whole_weight']],test_X[['Shucked_weight']],
                 test_X[['Viscera_weight']])
bsr_pre <- bsr_mat %*% bsr_coe
mse_bsr <- sum((bsr_pre-test_y)^2)/2177
# LASSO
X <- as.matrix(train[,1:8])
y <- train[[9]]
X <- apply(cbind(class.ind(X[,1])[,1:2],X[,2:8]),2,as.numeric)
cv.out <- cv.glmnet(X, y, alpha =1)
bestlam <- cv.out$lambda.min
l1 <- glmnet(X,y, alpha =1, lambda = bestlam)
test_X_l <- as.matrix(test_X)
test_X_l <- apply(cbind(rep(1,2177),class.ind(test_X_l[,1])[,1:2],test_X_l[,2:8]),2,as.numeric)
l1_coe <- coef(l1)@x
l1_pre <- test_X_l %*% l1_coe
mse_l1 <- sum((l1_pre-test_y)^2)/2177
# Ridge
cv.out <- cv.glmnet(X, y, alpha =0)
bestlam <- cv.out$lambda.min
l2 <- glmnet(X,y, alpha =0, lambda = bestlam)
```

```
l2_coe <- coef(l2)@x
l2_pre <- test_X_l %*% l2_coe
mse_l2 <- sum((l2_pre-test_y)^2)/2177
# PCAR
pcr_model <- pcr(Rings~.,data=train)
pcr_pre <- predict(pcr_model,newdata = test_X)
mse_pcr <- vector('numeric',8)
for(i in 1:8){
  mse_pcr[i] <- sum((pcr_pre[1:2177,1,i]-test_y)^2)/2177
}
mse_pcr <- min(mse_pcr)
# Cubic Spline
# summary(dat)
cs <- gam(Rings~Sex + bs(Length ,knots =c(0.45 ,0.54 ,0.61))+
            bs(Diameter ,knots =c(0.34 ,0.42 ,0.48))+ # bs(Height ,knots =c(0.11 ,0.14 ,0.16))+
            bs(Whole_weight ,knots =c(0.43 ,0.8 ,1.13))+ bs(Shucked_weight ,knots =c(0.18 ,0.33 ,0.49))
            bs(Viscera_weight ,knots =c(0.09 ,0.17 ,0.24))+ bs(Shell_weight ,knots =c(0.12 ,0.23 ,0.32)
          data=train)
cs_pre <- predict(cs , newdata = test_X[,-4])
mse_cs <- sum((cs_pre-test_y)^2)/2177
# Local Regression
localreg <- gam(Rings~
                  Sex +
                  lo(Length ,span =.2) + lo(Diameter ,span =.2) + # lo(Height ,span =.2) +
                  lo(Whole_weight ,span =.2) + lo(Shucked_weight ,span =.2) +
                  lo(Viscera_weight ,span =.2) + lo(Shell_weight ,span =.2), data=train)
local_pre <- predict(localreg , newdata = test_X[,-4])
mse_local <- sum((local_pre-test_y)^2)/2177
# Visualization
label <- c('LR', 'BSR', 'LASSO', 'RIDGE', 'PCR', 'CS', 'LOCALREG')
mse <- c(mse_lr, mse_bsr, mse_l1, mse_l2, mse_pcr, mse_cs, mse_local)
ggplot() +geom_bar(mapping = aes(x = label, y = mse, fill = label), stat = 'identity')+
  xlab('') + ylim(c(0,6))
```


mse
BSR
CS
LASSO
LOCALREG
LR
PCR
RIDGE
label
BSR
CS
LASSO
LOCALREG
LR
PCR
RIDGE