FTAP HW 7

Zachary Fogelson

July 21, 2015

Lasso Homework

Problem 1

1.

```
train <- read.csv("train.csv")
trainMeaningFull <-subset(train, select=-c(id, member_id))
trainMeaningFull <- trainMeaningFull[complete.cases(trainMeaningFull),]
stargazer(head(train), type = "pdf")
```

```
##
## % Error: 'style' must be either 'latex' (default), 'html' or 'text.'
```

```
stargazer(rbind(summary(train), sapply(train, sd)))
```

```
##
## % Table created by stargazer v.5.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvar
## % Date and time: Wed, Jul 22, 2015 - 16:33:39
## \begin{table}[!htbp] \centering
##   \caption{}
##   \label{}
## \begin{tabular}{@{\extracolsep{5pt}} ccccccccccc}
## \\[-1.8ex]\hline
## \hline \\[-1.8ex]
##  &       id &   member\_id &  funded\_amnt &     term\_m &    int\_rate &  installment &   emp\_leng
## \hline \\[-1.8ex]
##  & Min.   : 54734   & Min.   :  70699   & Min.   :  500   & Min.   :36.00   & Min.   :0.0542   & Min
##  & 1st Qu.:496560   & 1st Qu.: 635543   & 1st Qu.: 5000   & 1st Qu.:36.00   & 1st Qu.:0.0925   & 1st
##  & Median :623258   & Median : 798198   & Median : 9250   & Median :36.00   & Median :0.1171   & Medi
##  & Mean   :626300   & Mean   : 783617   & Mean   :10654   & Mean   :42.22   & Mean   :0.1191   & Mean
##  & 3rd Qu.:767345   & 3rd Qu.: 967828   & 3rd Qu.:15000   & 3rd Qu.:60.00   & 3rd Qu.:0.1435   & 3rd
##  & Max.   :975902   & Max.   :1198245   & Max.   :35000   & Max.   :60.00   & Max.   :0.2459   & Max
##  &  &  &  &  &  &  & NA's   :868   &  &  & NA's   :44   \\
##  & 170023.183351294 & 225109.217961916 & 6974.75731097762 & 10.5166250234748 & 0.0362426792330949 & 1
## \hline \\[-1.8ex]
## \end{tabular}
## \end{table}
```

2.

a)

```
ols <- lm(int_rate ~ ., trainMeaningFull)
olsSum <- summary(ols)
olsSum
```

Based on the OLS regression, all of the columns which are not IDs appear to be significant

b)

```
olsSum$r.squared
```

The baseline model has an R^2 of about .51. It is hard to judge if the baseline model is successful because, we do not have anything to compare it to. But using all of the available information we can explain 39% of the variance which is not terrible.

c)

```
plot(ols$fitted.values, ols$residuals)
```

d)

```
rmse <- mean((ols$residuals)^2)
rmse
```

e)

```
test <- read.csv("test.csv")
test <- subset(test, select=-c(id, member_id))
test <- test[complete.cases(test),]
```

```
predOLS <- predict(ols,test, interval = "none")
rmseOLS <- mean((test$int_rate - predOLS)^2)
cat("RMSE OLS: ", rmseOLS)
```

3.

```
interactDF <- as.data.frame(model.matrix(~(.)^2,subset(trainMeaningFull, select=-c(int_rate))))
interactTest <- as.data.frame(model.matrix(~(.)^2,subset(test, select=-c(int_rate))))
```

a)

```
smallest = lm(trainMeaningFull$int_rate ~ 1, interactDF)
biggest = as.formula(lm(trainMeaningFull$int_rate ~ ., interactDF))
stepForwardAIC <- step(smallest, scope = biggest, k = 2, direction ="forward", trace = F)
stepForwardBIC <- step(smallest, scope = biggest, k = log(length(trainMeaningFull$int_rate)), direction
cat("R^2 AIC: ", summary(stepForwardAIC)$r.squared, "R^2 BIC: ", summary(stepForwardBIC)$r.squared)
```

b)

```
cat("AIC: ", AIC(stepForwardAIC), "BIC: ", BIC(stepForwardBIC))
```

c)

```
rmseAIC <- mean((stepForwardAIC$residuals)^2)
rmseBIC <- mean((stepForwardBIC$residuals)^2)
cat("RMSE AIC: ", rmseAIC, "RMSE BIC: ", rmseBIC)
```

d)

```
predAIC <- predict(stepForwardAIC,interactTest, interval = "none")
predBIC <- predict(stepForwardBIC,interactTest, interval = "none")
rmseAIC <- mean((test$int_rate - predAIC)^2)
rmseBIC <- mean((test$int_rate - predBIC)^2)
cat("RMSE AIC: ", rmseAIC, "RMSE BIC: ", rmseBIC)
```

4. a)

```
biggest = lm(trainMeaningFull$int_rate ~ ., interactDF)
smallest = as.formula(lm(trainMeaningFull$int_rate ~ 1, interactDF))
stepBackwardAIC <- step(biggest, scope = smallest, k = 2, direction ="backward", trace = F)
stepBackwardBIC <- step(biggest, scope = smallest, k = log(length(trainMeaningFull$int_rate)), directio
cat("R^2 AIC: ", summary(stepBackwardAIC)$r.squared, "R^2 BIC: ", summary(stepBackwardBIC)$r.squared)
```

b)

```
cat("AIC: ", AIC(stepBackwardAIC), "BIC: ", BIC(stepBackwardBIC))
```

c)

```
rmseBackAIC <- mean((stepBackwardAIC$residuals)^2)
rmseBackBIC <- mean((stepBackwardBIC$residuals)^2)
cat("RMSE AIC: ", rmseBackAIC, "RMSE BIC: ", rmseBackBIC)
```

d)

```
predBackAIC <- predict(stepBackwardAIC,interactTest, interval = "none")
predBackBIC <- predict(stepBackwardBIC,interactTest, interval = "none")
rmseBackAIC <- mean((test$int_rate - predBackAIC)^2)
rmseBackBIC <- mean((test$int_rate - predBackBIC)^2)
cat("RMSE AIC: ", rmseBackAIC, "RMSE BIC: ", rmseBackBIC)
```

5.

a)

```
model=as.formula(paste("~", paste(names(interactDF)[-1], collapse= "+")))
x=model.matrix(model,interactDF);
lassoFit <- glmnet(x,trainMeaningFull$int_rate)
outLasso <- predict(lassoFit, newx = as.matrix(interactDF), s=.0001)
cat("R^2: ", var(outLasso)/var(ols$residuals+ols$fitted.values))
coef(lassoFit, s=.0001)
```

The R^2 is calculated based on the R^2 calculation in lasso_class; however, prof. Russell did mention that there is no R^2 for the lasso, so I printed the coefficients of the lambda model instead.

b)

```
insamplePred <- predict(lassoFit, newx = as.matrix(interactDF), s=.0001)
lassoRMSE <- mean((insamplePred - trainMeaningFull$int_rate)^2)
cat("RMSE Lasso: ", lassoRMSE)
```

c)

```
predLasso <- predict(lassoFit,newx = as.matrix(interactTest), s=.0001)
lassoRMSE <- mean((test$int_rate - predLasso)^2)
cat("RMSE Lasso: ", lassoRMSE)
```

6.

a)

```
lassoCVFit <- cv.glmnet(x,trainMeaningFull$int_rate, nfolds = 10)
lassoCVFit$lambda.min
```

b)

```
insampleCVPred <- predict(lassoCVFit, newx = as.matrix(interactDF), s="lambda.min")
lassoCVRMSE <- mean((insampleCVPred - trainMeaningFull$int_rate)^2)
cat("RMSE CVLasso: ", lassoCVRMSE)
```

c)

```
predCVLasso <- predict(lassoCVFit,newx = as.matrix(interactTest), s="lambda.min")
lassoCVRMSE <- mean((test$int_rate - predCVLasso)^2)
cat("RMSE CVLasso: ", lassoCVRMSE)
```

Logit Homework

```
trades <- read.csv("detailed_trades_est.csv")
trades <- subset(trades, trades$PCHANGE0==0| trades$PCHANGE0==2)
```

1)

```
logit <- glm(PCHANGE~LASK+LBID+RETURN+SIGN_VOL, family=binomial, data=trades)
cat("Lasso? AIC: ", AIC(logit), "Lasso BIC: ", BIC(logit))
```

2)

```
interactions <- subset(trades,select = -c(PCHANGE,PCHANGE0))
interactions <- as.data.frame(model.matrix(~(.)^2, interactions))
```

3)

```
trades.pca <- prcomp(interactions)
myModel <- lm(trades$PCHANGE~trades.pca$x[,1]+trades.pca$x[,2]+trades.pca$x[,3])
```

4)

```
error <- trades$PCHANGE-(as.numeric(myModel$fitted.values > .5))
mean(abs(error))
```