Individual Assignment 3

Charlie Ling

9/15/2021

4.7 Exercises Problem 10

This is an individual assignment. You are allowed to collaborate with other students. However, you are not allowed to copy others code and/or report. You are expected to write your own code and produce your own report as a pdf file. Use R Markdown to produce this report. If you encounter trouble with R Markdown, transfer your code, output, and comments to a Word document and convert that document to a PDF file.

10. This question should be answered using the Weekly data set, which is part of the ISLR package. This data is similar in nature to the Smarket data from this chapter's lab, except that it contains 1, 089 weekly returns for 21 years, from the beginning of 1990 to the end of 2010.

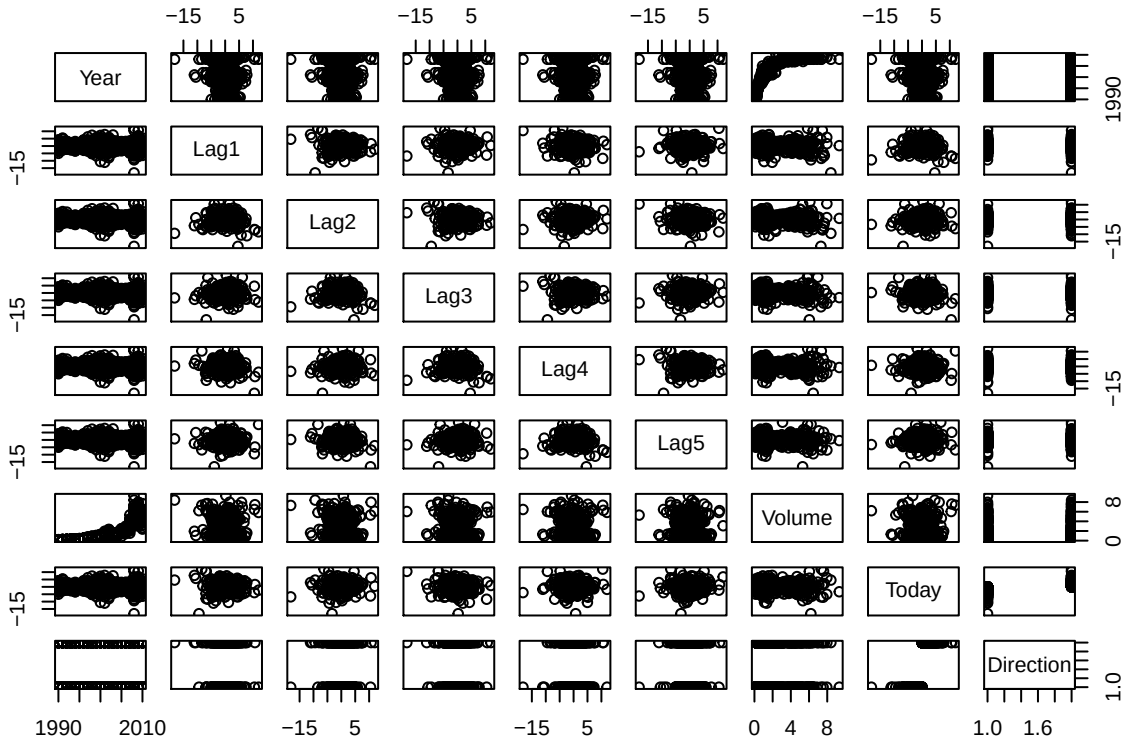
```
rm(list=ls())#release memory
library(ISLR)
```

```
## Warning: package 'ISLR' was built under R version 4.0.5
```

```
Weekly=na.omit(Weekly)
View(Weekly)
```

(a) Produce some numerical and graphical summaries of the Weekly data. Do there appear to be any patterns?

```
pairs(Weekly)
```




```
lm.fit=lm(Volume~Year,data = Weekly)
summary(lm.fit)
```

```
##
## Call:
## lm(formula = Volume ~ Year, data = Weekly)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.8505 -0.5862 -0.2402  0.4935  5.8821
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -4.692e+02  9.151e+00  -51.27   <2e-16 ***
## Year         2.354e-01  4.575e-03   51.45   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9105 on 1087 degrees of freedom
## Multiple R-squared:  0.7089, Adjusted R-squared:  0.7086
## F-statistic:  2647 on 1 and 1087 DF,  p-value: < 2.2e-16
```

```
# Volumn increses as year goes by
```

(b) Use the full data set to perform a logistic regression with Direction as the response and the five lag variables plus Volume as predictors. Use the summary function to print the results. Do any of the predictors appear to be statistically significant? If so, which ones?

```
glm.fit=glm(Direction~.-(Today+Year),data=Weekly,family = binomial)
summary(glm.fit)
```

```
##
## Call:
## glm(formula = Direction ~ . - (Today + Year), family = binomial,
##     data = Weekly)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6949  -1.2565   0.9913   1.0849   1.4579
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.26686    0.08593   3.106   0.0019 **
## Lag1        -0.04127    0.02641  -1.563   0.1181
## Lag2         0.05844    0.02686   2.175   0.0296 *
## Lag3        -0.01606    0.02666  -0.602   0.5469
## Lag4        -0.02779    0.02646  -1.050   0.2937
## Lag5        -0.01447    0.02638  -0.549   0.5833
## Volume      -0.02274    0.03690  -0.616   0.5377
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1496.2  on 1088  degrees of freedom
## Residual deviance: 1486.4  on 1082  degrees of freedom
## AIC: 1500.4
##
## Number of Fisher Scoring iterations: 4
```

```
# Lag2 appears to be statistically significant
```

(c) Compute the confusion matrix and overall fraction of correct predictions. Explain what the confusion matrix is telling you about the types of mistakes made by logistic regression.

```
glm.probs = predict(glm.fit, type = "response")
glm.pred = rep("Down",length(glm.probs))
glm.pred[glm.probs>0.5]="Up"
table(glm.pred,Weekly$Direction)
```

```
##
## glm.pred Down  Up
##     Down   54  48
##     Up    430 557
```

```
mean(glm.pred==Weekly$Direction)
```

```
## [1] 0.5610652
```

```
# There are 430 actual "Down" is mistakenly classified as "UP",
# while 48 actual "UP" is is mistakenly classified as "Down"
```

(d) Now fit the logistic regression model using a training data period from 1990 to 2008, with Lag2 as the only predictor. Compute the confusion matrix and the overall fraction of correct predictions for the held out data (that is, the data from 2009 and 2010).

```
train=(Weekly$Year>=1990)&(Weekly$Year<=2008)
glm.fit=glm(Direction~Lag2,data=Weekly[train,],family = binomial)
glm.probs = predict(glm.fit,Weekly[!train,], type = "response")
glm.pred = rep("Down",length(glm.probs))
glm.pred[glm.probs>0.5]="Up"
table(glm.pred,Weekly[!train,]$Direction)
```

```
##
## glm.pred Down Up
##     Down    9  5
##     Up     34 56
```

```
mean(glm.pred==Weekly[!train,]$Direction)
```

```
## [1] 0.625
```

(e) Repeat (d) using LDA.

```
library(MASS)
lda.fit = lda(Direction~Lag2,data=Weekly,subset = train)
lda.pred = predict(lda.fit,Weekly[!train,])
lda.class = lda.pred$class
table(lda.class,Weekly[!train,]$Direction)
```

```
##
## lda.class Down Up
##      Down    9  5
##      Up     34 56
```

```
mean(lda.class==Weekly[!train,]$Direction)
```

```
## [1] 0.625
```

(g) Repeat (d) using KNN with K = 1.

```
train.X = cbind(Weekly[train,]$Lag2)
test.X = cbind(Weekly[!train,]$Lag2)
train.Direction = Weekly[train,]$Direction
test.Direction = Weekly[!train,]$Direction
library(class)
set.seed(1)
knn.pred = knn(train.X,test.X,train.Direction,k=1)
table(knn.pred,test.Direction)
```

```
##         test.Direction
## knn.pred Down Up
##     Down   21 30
##     Up     22 31
```

```
mean(knn.pred == test.Direction)
```

```
## [1] 0.5
```

(h) Which of these methods appears to provide the best results on this data?

```
# logistic regression and LDA both provide the best results on this data
```

(i) Experiment with different combinations of predictors, including possible transformations and interactions, for each of the methods. Report the variables, method, and associated confusion matrix that appears

to provide the best results on the held out data. Note that you should also experiment with values for K in the KNN classifier.

```
glm.fit=glm(Direction~Lag2:Volume,data=Weekly[train,],family = binomial)
glm.probs = predict(glm.fit,Weekly[!train,], type = "response")
glm.pred = rep("Down",length(glm.probs))
glm.pred[glm.probs>0.5]="Up"
table(glm.pred,Weekly[!train,]$Direction)
```

```
##
## glm.pred Down Up
##     Down    9  6
##     Up     34 55
```

```
mean(glm.pred==Weekly[!train,]$Direction)
```

```
## [1] 0.6153846
```

```
lda.fit = lda(Direction~Lag2:Volume,data=Weekly,subset = train)
lda.pred = predict(lda.fit,Weekly[!train,])
lda.class = lda.pred$class
table(lda.class,Weekly[!train,]$Direction)
```

```
##
## lda.class Down Up
##      Down    8  6
##      Up     35 55
```

```
mean(lda.class==Weekly[!train,]$Direction)
```

```
## [1] 0.6057692
```

```
train.X = cbind(Weekly[train,]$Lag2*Weekly[train,]$Volume)
test.X = cbind(Weekly[!train,]$Lag2*Weekly[!train,]$Volume)
train.Direction = Weekly[train,]$Direction
test.Direction = Weekly[!train,]$Direction


k=rep(0,10)
test_accuracy=rep(0,10)
for(i in 1:10){
    knn.pred=knn(train.X,test.X,train.Direction,k=i)
    accuracy=mean(knn.pred == test.Direction)
    test_accuracy[i]=accuracy
    k[i]=i
}
cbind(k,test_accuracy)
```

```
##        k test_accuracy
##  [1,]  1     0.5000000
##  [2,]  2     0.4423077
##  [3,]  3     0.4807692
##  [4,]  4     0.5096154
##  [5,]  5     0.5096154
##  [6,]  6     0.5192308
##  [7,]  7     0.4711538
##  [8,]  8     0.4903846
##  [9,]  9     0.5000000
```

```
## [10,] 10     0.5096154
```

```
# logistic regression provides the best results on this data,
# with the predictor of Lag2:Volume
```