Individual_assignment3

Yuhan_Xu_474154

2019/9/14

Prefix

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.

```
library("ISLR")
fix(Weekly)
attach(Weekly)
```

(a)

Q: Produce some numerical and graphical summaries of the Weekly data. Do there appear to be any patterns?

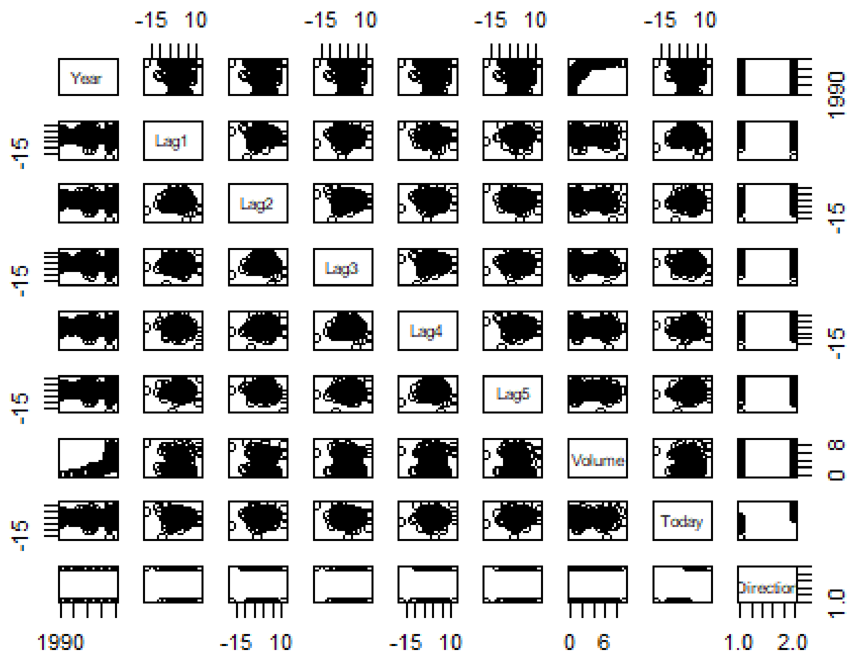
```
summary(Weekly)

##       Year           Lag1               Lag2               Lag3

##  Min.   :1990   Min.   :-18.1950   Min.   :-18.1950   Min.   :-18.19
50
##  1st Qu.:1995   1st Qu.: -1.1540   1st Qu.: -1.1540   1st Qu.: -1.15
80
##  Median :2000   Median :  0.2410   Median :  0.2410   Median :  0.24
10
##  Mean   :2000   Mean   :  0.1506   Mean   :  0.1511   Mean   :  0.14
72
##  3rd Qu.:2005   3rd Qu.:  1.4050   3rd Qu.:  1.4090   3rd Qu.:  1.40
90
##  Max.   :2010   Max.   : 12.0260   Max.   : 12.0260   Max.   : 12.02
60
##       Lag4               Lag5              Volume
##  Min.   :-18.1950   Min.   :-18.1950   Min.   :0.08747
##  1st Qu.: -1.1580   1st Qu.: -1.1660   1st Qu.:0.33202
##  Median :  0.2380   Median :  0.2340   Median :1.00268
##  Mean   :  0.1458   Mean   :  0.1399   Mean   :1.57462
##  3rd Qu.:  1.4090   3rd Qu.:  1.4050   3rd Qu.:2.05373
##  Max.   : 12.0260   Max.   : 12.0260   Max.   :9.32821
##      Today            Direction
##  Min.   :-18.1950   Down:484
##  1st Qu.: -1.1540   Up  :605
```

```
##   Median :  0.2410
##   Mean   :  0.1499
##   3rd Qu.:  1.4050
##   Max.   : 12.0260
```

```r
pairs(Weekly)
```




```r
cor(Weekly[,-9])
```

```
##               Year         Lag1        Lag2        Lag3         Lag4
## Year    1.00000000 -0.032289274 -0.03339001 -0.03000649 -0.031127923
## Lag1   -0.03228927  1.000000000 -0.07485305  0.05863568 -0.071273876
## Lag2   -0.03339001 -0.074853051  1.00000000 -0.07572091  0.058381535
## Lag3   -0.03000649  0.058635682 -0.07572091  1.00000000 -0.075395865
## Lag4   -0.03112792 -0.071273876  0.05838153 -0.07539587  1.000000000
## Lag5   -0.03051910 -0.008183096 -0.07249948  0.06065717 -0.075675027
## Volume  0.84194162 -0.064951313 -0.08551314 -0.06928771 -0.061074617
## Today  -0.03245989 -0.075031842  0.05916672 -0.07124364 -0.007825873
##                Lag5      Volume        Today
## Year   -0.030519101  0.84194162 -0.032459894
## Lag1   -0.008183096 -0.06495131 -0.075031842
## Lag2   -0.072499482 -0.08551314  0.059166717
## Lag3    0.060657175 -0.06928771 -0.071243639
## Lag4   -0.075675027 -0.06107462 -0.007825873
## Lag5    1.000000000 -0.05851741  0.011012698
## Volume -0.058517414  1.00000000 -0.033077783
## Today   0.011012698 -0.03307778  1.000000000
```

A: From the summary, we know that the data are collected from 1990 to 2010. Also, all the lag variables and today (percentage return for this week) goes from -18.1960 to 12.0260. In 484 weeks, the direction is “down”, while in 605 weeks the direction is “up”.
From the scatter plots, we didn’t see any pattern between the variables except for year and volume.
From the correlations, we know that Volume positively correlates with year, which means as time goes by, the volume of shares traded has increased.

(b)

Q: 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
```

A: Only Lag2 is statistically significant at 95% confidence level.

(c)

Q: 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, Weekly, type = "response")
glm.pred = ifelse(glm.probs > 0.5, "Up", "Down")
table(glm.pred, Direction)

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

mean(glm.pred == Direction)

## [1] 0.5610652
```

A: The overall fraction of correct predictions is 56.1%. From the confusion matrix, we know that this model can predict correctly most of the time when the actual direction is "Up". However, when the actual direction is "Down", the predictions are false most of the time.

(d)

Q: 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 = (Year < 2009)
Weekly.train = Weekly[train,]
Weekly.test = Weekly[!train,]

glm.fit1 = glm(Direction~Lag2, data = Weekly.train, family = "binomial
")
glm.probs = predict(glm.fit1, Weekly.test, type = "response")
glm.pred = ifelse(glm.probs > 0.5, "Up", "Down")
table(glm.pred, Direction[!train])

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

mean(glm.pred == Direction[!train])

## [1] 0.625
```

(e)

Q: Repeat (d) using LDA.

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

##
##         Down Up
##    Down    9  5
##    Up     34 56

mean(lda.pred$class == Direction[!train])

## [1] 0.625
```

(g)

Q: Repeat (d) using KNN with K = 1.

```
library(class)
knn.pred = knn(as.matrix(Lag2[train]), as.matrix(Lag2[!train]), Directi
on[train], k = 1)
table(knn.pred, Direction[!train])

##
## knn.pred Down Up
##     Down   21 29
##     Up     22 32

mean(knn.pred == Direction[!train])

## [1] 0.5096154
```

(h)

Q: Which of these methods appears to provide the best results on this data?

A: LDA and logistic regression provide the same result, which is better than the results provided by KNN.

(i)

Q: 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.

A:

(1)Fit the logistic regression model using a training data period from 1990 to 2008, with Lag1, Lag2 two predictors and also considering the interaction between them:

```
glm.fit2 = glm(Direction~Lag1*Lag2, Weekly.train, family = "binomial")
glm.probs = predict(glm.fit2, Weekly.test, type = "response")
glm.pred = ifelse(glm.probs > 0.5, "Up", "Down")
table(glm.pred, Direction[!train])

##
## glm.pred Down Up
##     Down    7  8
##     Up     36 53

mean(glm.pred == Direction[!train])

## [1] 0.5769231
```

(2)Fit the logistic regression model using a training data period from 1990 to 2008, with Lag2, Lag3 two predictors and also considering the interaction between them:

```
glm.fit3 = glm(Direction~Lag2*Lag3, Weekly.train, family = "binomial")
glm.probs = predict(glm.fit3, Weekly.test, type = "response")
glm.pred = ifelse(glm.probs > 0.5, "Up", "Down")
table(glm.pred, Direction[!train])

##
## glm.pred Down Up
##     Down    8  4
##     Up     35 57

mean(glm.pred == Direction[!train])

## [1] 0.625
```

(3)Fit the logistic regression model using a training data period from 1990 to 2008, only considering the interaction between Lag2 and Lag3:

```
glm.fit4 = glm(Direction~Lag2:Lag3, Weekly.train, family = "binomial")
glm.probs = predict(glm.fit4, Weekly.test, type = "response")
glm.pred = ifelse(glm.probs > 0.5, "Up", "Down")
table(glm.pred, Direction[!train])

##
## glm.pred Down Up
##       Up   43 61

mean(glm.pred == Direction[!train])

## [1] 0.5865385
```

(4)Fit the logistic regression model using a training data period from 1990 to 2008, with Lag2^2 as the only predictor:

```
glm.fit5 = glm(Direction~Lag2^2, Weekly.train, family = "binomial")
glm.probs = predict(glm.fit5, Weekly.test, type = "response")
glm.pred = ifelse(glm.probs > 0.5, "Up", "Down")
table(glm.pred, Direction[!train])

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

mean(glm.pred == Direction[!train])

## [1] 0.625
```

(5)Fit the LDA model using a training data period from 1990 to 2008, with Lag2, Lag3 two predictors and also considering the interaction between them:

```
lda.fit1 = lda(Direction~Lag2*Lag3, data = Weekly.train)
lda.pred = predict(lda.fit1, Weekly.test)
table(lda.pred$class, Direction[!train])

##
##        Down Up
##   Down    8  4
##   Up     35 57

mean(lda.pred$class == Direction[!train])

## [1] 0.625
```

(6)Fit the KNN model using a training data period from 1990 to 2008, with Lag2 as the only predictor, using K from 1 to 100:

```
highest_rate=0
highest_trial=0
for (i in 1:100){
  knn.pred = knn(as.matrix(Lag2[train]), as.matrix(Lag2[!train]), Direc
tion[train], k = i)
  table(knn.pred, Direction[!train])
  accuracy_rate = mean(knn.pred == Direction[!train])
  if (accuracy_rate>highest_rate){
      highest_rate = accuracy_rate
      highest_trial = i
  }
}

highest_rate

## [1] 0.6153846
```

```
highest_trial

## [1] 47
```

When k = 47, the highest accuracy rate of KNN model is 61.5%.

In conclusion, fit the logistic regression model with Lag2, Lag3 two predictors and also consider their interaction has the highest accuracy rate. But the accuracy rate is the same as the model we used in previous quesiton, that is to fit the logistic regression and LDA with Lag2 as the only predictor. Therefore, these three models have the best results.