Introduction to Statistical Learning - Ch4 - Ex10

Thalles Quinaglia Liduares

23/03/2023

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.

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

```
library(ISLR)
library(MASS)
library(caret)
```

Upload data

```
data<-ISLR::Weekly
```

View structure of Weekly data

```
str(Weekly)
```

```
## 'data.frame':    1089 obs. of  9 variables:
##  $ Year     : num  1990 1990 1990 1990 1990 1990 1990 1990 1990 1990 ...
##  $ Lag1     : num  0.816 -0.27 -2.576 3.514 0.712 ...
##  $ Lag2     : num  1.572 0.816 -0.27 -2.576 3.514 ...
##  $ Lag3     : num  -3.936 1.572 0.816 -0.27 -2.576 ...
##  $ Lag4     : num  -0.229 -3.936 1.572 0.816 -0.27 ...
##  $ Lag5     : num  -3.484 -0.229 -3.936 1.572 0.816 ...
##  $ Volume   : num  0.155 0.149 0.16 0.162 0.154 ...
##  $ Today    : num  -0.27 -2.576 3.514 0.712 1.178 ...
##  $ Direction: Factor w/ 2 levels "Down","Up": 1 1 2 2 2 1 2 2 2 1 ...
```

summary statistics

```
summary(Weekly)
```

```
##       Year           Lag1                Lag2                Lag3          
##  Min.   :1990   Min.   :-18.1950   Min.   :-18.1950   Min.   :-18.1950  
##  1st Qu.:1995   1st Qu.: -1.1540   1st Qu.: -1.1540   1st Qu.: -1.1580  
##  Median :2000   Median :  0.2410   Median :  0.2410   Median :  0.2410  
##  Mean   :2000   Mean   :  0.1506   Mean   :  0.1511   Mean   :  0.1472  
##  3rd Qu.:2005   3rd Qu.:  1.4050   3rd Qu.:  1.4090   3rd Qu.:  1.4090  
##  Max.   :2010   Max.   : 12.0260   Max.   : 12.0260   Max.   : 12.0260  
##       Lag4                Lag5              Volume            Today          
##  Min.   :-18.1950   Min.   :-18.1950   Min.   :0.08747   Min.   :-18.1950  
##  1st Qu.: -1.1580   1st Qu.: -1.1660   1st Qu.:0.33202   1st Qu.: -1.1540  
##  Median :  0.2380   Median :  0.2340   Median :1.00268   Median :  0.2410  
##  Mean   :  0.1458   Mean   :  0.1399   Mean   :1.57462   Mean   :  0.1499  
##  3rd Qu.:  1.4090   3rd Qu.:  1.4050   3rd Qu.:2.05373   3rd Qu.:  1.4050  
##  Max.   : 12.0260   Max.   : 12.0260   Max.   :9.32821   Max.   : 12.0260  
##  Direction 
##  Down:484  
##  Up  :605  
##            
##            
##            
##            
```

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?

Fit a logistic regression model

```
model <- glm(Direction ~ Lag1 + Lag2 + Lag3 + Lag4 + Lag5 + Volume, data = Weekly, family = b
inomial)
```

Print the summary of the model

```
summary(model)
```

```
##
## Call:
## glm(formula = Direction ~ Lag1 + Lag2 + Lag3 + Lag4 + Lag5 +
##     Volume, 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
```

Based on the p-values in the summary output, we can see that only the intercept and the Lag1 variable have a statistically significant relationship with the response variable Direction at a significance level of 0.05. The p-values for the other predictor variables are all greater than 0.05, indicating that they are not statistically significant predictors in this model.

Therefore, we can conclude that only the intercept and the Lag1 variable are statistically significant predictors of the Direction variable in this logistic regression model.

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.

```
# Make predictions using the logistic regression model
predictions <- ifelse(predict(model, Weekly, type = "response") > 0.5, "Up", "Down")

# Create a confusion matrix
conf_mat <- table(predictions, Weekly$Direction)

# Print the confusion matrix
conf_mat
```

```
##
## predictions Down  Up
##        Down   54  48
##        Up    430 557
```

```
# Calculate the overall fraction of correct predictions
correct_frac <- mean(predictions == Weekly$Direction)
correct_frac
```

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

The overall fraction of correct predictions for this model is 56.04%, which means that the model correctly predicted the direction of the stock market for slightly more than half of the weeks in the data set.

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).

```
# Create a training data set from 1990 to 2008, with Lag2 as the only predictor
train <- subset(Weekly, Year < 2009, select=c("Direction", "Lag2"))

# Create a testing data set from 2009 to 2010, with Lag2 as the only predictor
test <- subset(Weekly, Year >= 2009, select=c("Direction", "Lag2"))

# Fit a logistic regression model to the training data set
model <- glm(Direction ~ Lag2, data = train, family = binomial)

# Make predictions on the testing data set
probabilities <- predict(model, newdata = test, type = "response")
predictions <- ifelse(probabilities > 0.5, "Up", "Down")

# Compute the confusion matrix and the overall fraction of correct predictions
table(predictions, test$Direction)
```

```
##
## predictions Down Up
##        Down    9  5
##        Up     34 56
```

```
accuracy <- mean(predictions == test$Direction)
accuracy
```

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

The accuracy of the model is equal to 62.5%.

e. Repeat (d) using LDA (Linear discriminant analysis)

Fit the LDA model on the training data

```
lda_model <- lda(Direction ~ Lag2, data = train)
```

Make predictions on the held out data using the fitted LDA model

```
lda_pred <- predict(lda_model, newdata = test)
```

confusion matrix

```
confusionMatrix(lda_pred$class, test$Direction)$table
```

```
##           Reference
## Prediction Down Up
##       Down    9  5
##       Up     34 56
```

```
confusionMatrix(lda_pred$class, test$Direction)$overall["Accuracy"]
```

```
## Accuracy
##    0.625
```