基本分类模型案例：股票市场走势预测

吴翔

2019-03-18

概述

我们通过R语言 `ISLR` 包中股票市场走势的案例来阐述如何使用如下基本分类模型：

- logistic回归
- LDA
- QDA
- KNN

数据集 `Smarket` 包含了2001-2005年1250天里S&P 500股票指数的投资回报率。Lag1 ~ Lag5是过去5个交易日的投资回报率，Volume为前一交易日的股票成交量（单位为十亿），Today为当日的投资回报率，Direction为市场走势方向（Up或者Down）。

```
# clean the work directory
rm(list = ls())

# set seeds
set.seed(123)

# read dataset
suppressMessages(library(ISLR))
suppressMessages(library(tidyverse))
data("Smarket")
# display the variables
str(Smarket)
```

```
## 'data.frame':    1250 obs. of  9 variables:
##  $ Year     : num  2001 2001 2001 2001 2001 ...
##  $ Lag1     : num  0.381 0.959 1.032 -0.623 0.614 ...
##  $ Lag2     : num  -0.192 0.381 0.959 1.032 -0.623 ...
##  $ Lag3     : num  -2.624 -0.192 0.381 0.959 1.032 ...
##  $ Lag4     : num  -1.055 -2.624 -0.192 0.381 0.959 ...
##  $ Lag5     : num  5.01 -1.055 -2.624 -0.192 0.381 ...
##  $ Volume   : num  1.19 1.3 1.41 1.28 1.21 ...
##  $ Today    : num  0.959 1.032 -0.623 0.614 0.213 ...
##  $ Direction: Factor w/ 2 levels "Down","Up": 2 2 1 2 2 2 1 2 2 2 ...
```

```
# summary of dataset
summary(Smarket)
```

```
##       Year           Lag1                Lag2                Lag3          
##  Min.   :2001   Min.   :-4.92   Min.   :-4.92   Min.   :-4.92  
##  1st Qu.:2002   1st Qu.:-0.64   1st Qu.:-0.64   1st Qu.:-0.64  
##  Median :2003   Median : 0.04   Median : 0.04   Median : 0.04  
##  Mean   :2003   Mean   : 0.00   Mean   : 0.00   Mean   : 0.00  
##  3rd Qu.:2004   3rd Qu.: 0.60   3rd Qu.: 0.60   3rd Qu.: 0.60  
##  Max.   :2005   Max.   : 5.73   Max.   : 5.73   Max.   : 5.73  
##       Lag4           Lag5              Volume          Today        
##  Min.   :-4.92   Min.   :-4.92   Min.   :0.356   Min.   :-4.92  
##  1st Qu.:-0.64   1st Qu.:-0.64   1st Qu.:1.257   1st Qu.:-0.64  
##  Median : 0.04   Median : 0.04   Median :1.423   Median : 0.04  
##  Mean   : 0.00   Mean   : 0.01   Mean   :1.478   Mean   : 0.00  
##  3rd Qu.: 0.60   3rd Qu.: 0.60   3rd Qu.:1.642   3rd Qu.: 0.60  
##  Max.   : 5.73   Max.   : 5.73   Max.   :3.152   Max.   : 5.73  
##  Direction
##  Down:602
##  Up  :648
##
##
##
##
```

logistic回归

我们用2001-2004年的数据作为训练集，2005年的数据作为测试集。

```
# training set
train <- Smarket$Year < 2005
smarket.test <- Smarket[!train, ]
smarket.train <- Smarket[train, ]
```

训练集包含998个样本，测试集包含252个样本。

```
# logistic regression
glm.fit <- glm(Direction ~ Lag1 + Lag2 + Lag3 + Lag4 + Lag5 + Volume, data = smarket.train, family = b
inomial)
summary(glm.fit)
```

```
##
## Call:
## glm(formula = Direction ~ Lag1 + Lag2 + Lag3 + Lag4 + Lag5 +
##     Volume, family = binomial, data = smarket.train)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
##  -1.30   -1.19    1.08    1.16    1.35
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.19121    0.33369    0.57     0.57
## Lag1        -0.05418    0.05179   -1.05     0.30
## Lag2        -0.04581    0.05180   -0.88     0.38
## Lag3         0.00720    0.05164    0.14     0.89
## Lag4         0.00644    0.05171    0.12     0.90
## Lag5        -0.00422    0.05114   -0.08     0.93
## Volume      -0.11626    0.23962   -0.49     0.63
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1383.3  on 997  degrees of freedom
## Residual deviance: 1381.1  on 991  degrees of freedom
## AIC: 1395
##
## Number of Fisher Scoring iterations: 3
```

```
# predictions
glm.pred <- ifelse(predict(glm.fit, smarket.test, type = "response") > 0.5, "Up", "Down")
# compare predictions with true values
table(glm.pred, smarket.test$Direction)
```

```
##
## glm.pred Down Up
##     Down   77 97
##     Up     34 44
```

```
# performance
mean(glm.pred == smarket.test$Direction)
```

```
## [1] 0.48
```

可以看到，logistic回归预测的准确率为0.48，小于随机猜测。

检视模型，发现纳入了过多无关变量，因而出现了过度拟合的问题。因此，仅纳入Lag1和Lag2，重新运行logistic回归模型。

```
# logistic regression
glm.fit <- glm(Direction ~ Lag1 + Lag2, data = smarket.train, family = binomial)
summary(glm.fit)
```

```
##
## Call:
## glm(formula = Direction ~ Lag1 + Lag2, family = binomial, data = smarket.train)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
##  -1.35   -1.19    1.07    1.16    1.33
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   0.0322     0.0634    0.51     0.61
## Lag1         -0.0556     0.0517   -1.08     0.28
## Lag2         -0.0445     0.0517   -0.86     0.39
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1383.3  on 997  degrees of freedom
## Residual deviance: 1381.4  on 995  degrees of freedom
## AIC: 1387
##
## Number of Fisher Scoring iterations: 3
```

```
# predictions
glm.pred <- ifelse(predict(glm.fit, smarket.test, type = "response") > 0.5, "Up", "Down")
# compare predictions with true values
table(glm.pred, smarket.test$Direction)
```

```
##
## glm.pred Down  Up
##     Down   35  35
##     Up     76 106
```

```
# performance
mean(glm.pred == smarket.test$Direction)
```

```
## [1] 0.56
```

此时logistic回归预测的准确率为0.56，略大于随机猜测。

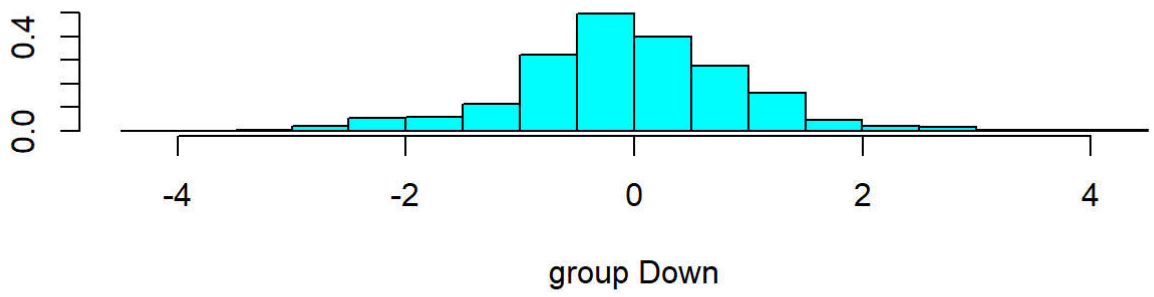
LDA

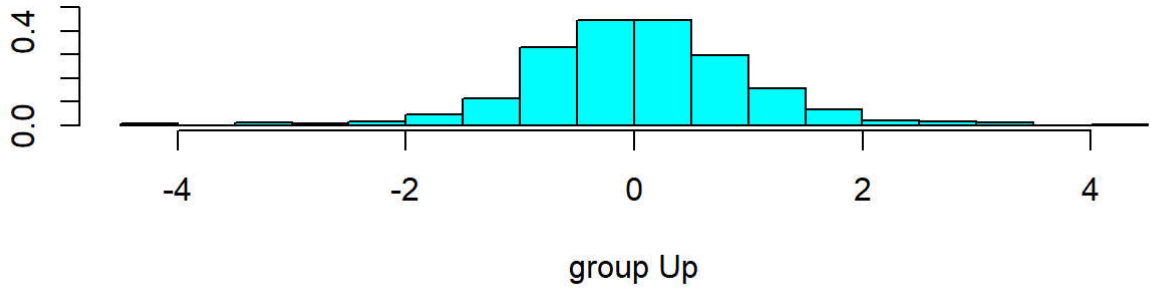
采用LDA预测股票市场走势。

```
suppressMessages(library(MASS))
# LDA
lda.fit <- lda(Direction ~ Lag1 + Lag2, data = smarket.train)
lda.fit
```

```
## Call:
## lda(Direction ~ Lag1 + Lag2, data = smarket.train)
##
## Prior probabilities of groups:
##  Down    Up
## 0.492 0.508
##
## Group means:
##         Lag1    Lag2
## Down  0.0428  0.0339
## Up   -0.0395 -0.0313
##
## Coefficients of linear discriminants:
##         LD1
## Lag1 -0.642
## Lag2 -0.514
```

```
# plot
plot(lda.fit)
```





类似地，评估预测效果。

```
# predictions
lda.pred <- predict(lda.fit, smarket.test)
# compare predictions with true values
table(lda.pred$class, smarket.test$Direction)
```

```
##
##         Down  Up
##   Down    35  35
##   Up      76 106
```

```
# performance
mean(lda.pred$class == smarket.test$Direction)
```

```
## [1] 0.56
```

LDA预测的准确率为0.56，略大于随机猜测，与logistic回归相当。

QDA

采用QDA预测股票市场走势。

```
# QDA
qda.fit <- qda(Direction ~ Lag1 + Lag2, data = smarket.train)
qda.fit
```

```
## Call:
## qda(Direction ~ Lag1 + Lag2, data = smarket.train)
##
## Prior probabilities of groups:
##  Down    Up
## 0.492 0.508
##
## Group means:
##         Lag1    Lag2
## Down  0.0428  0.0339
## Up   -0.0395 -0.0313
```

类似地，评估预测效果。

```
# predictions
qda.pred <- predict(qda.fit, smarket.test)
# compare predictions with true values
table(qda.pred$class, smarket.test$Direction)
```

```
##
##         Down  Up
##   Down    30  20
##   Up      81 121
```

```
# performance
mean(qda.pred$class == smarket.test$Direction)
```

```
## [1] 0.599
```

QDA预测的准确率为0.599，高于logistic回归和LDA。

KNN

采用KNN预测股票市场走势。

```
suppressMessages(library(class))
train.X <- smarket.train[, c("Lag1", "Lag2")]
test.X <- smarket.test[, c("Lag1", "Lag2")]
train.Direction <- smarket.train$Direction
# KNN
accuracy <- NULL
for (kn in 1:6){
  knn.pred <- knn(train = train.X, test = test.X, cl = train.Direction, k = kn)
  accuracy <- c(accuracy, mean(knn.pred == smarket.test$Direction))
}
# accuracy for k = 1, ..., 6
accuracy
```

```
## [1] 0.500 0.512 0.532 0.512 0.484 0.504
```

可以看到，KNN预测的最高准确率为0.532，此时$K = 3$。

```
# K = 3
knn.pred <- knn(train = train.X, test = test.X, cl = train.Direction, k = 3)
# compare predictions with true values
table(knn.pred, smarket.test$Direction)
```

```
##
## knn.pred Down Up
##     Down   48 55
##     Up     63 86
```

```
# performance
mean(knn.pred == smarket.test$Direction)
```

```
## [1] 0.532
```

总结

最后，我们给出各个分类模型的效果。

```
# performance comparison
performance <- c(mean(glm.pred == smarket.test$Direction), mean(lda.pred$class == smarket.test$Directi
on), mean(qda.pred$class == smarket.test$Direction), mean(knn.pred == smarket.test$Direction))
names(performance) <- c("logistic", "LDA", "QDA", "KNN")
performance
```

```
## logistic      LDA      QDA      KNN
##    0.560    0.560    0.599    0.532
```