|  |  |
| --- | --- |
| Big Data and Machine Learning in Logistics | 2nd semester 2021 |
| Student number: 50211#90 | Student name: Nolwenn PIGEON 놀웬 피지언 |

**R Self-work submission 3**

1. LDA, QDA, KNN with Smarket dataset

# LDA

>lda.fit=lda(**Direction~Lag1+Lag2+Lag3**,data=Smarket,subset=train) # not the same as in the RLab session

>lda.fit

> library(ISLR)

> library(MASS)

> fix(Smarket)

> names(Smarket)

[1] "Year" "Lag1" "Lag2" "Lag3" "Lag4"

[6] "Lag5" "Volume" "Today" "Direction"

> attach(Smarket)

> train=Year<2005

> lda.fit=lda(Direction~Lag1+Lag2+Lag3, data = Smarket,subset = train)

Une image contenant texte

Description générée automatiquement> lda.fit

We have 49.2% “Down” and 50.8% “Up”

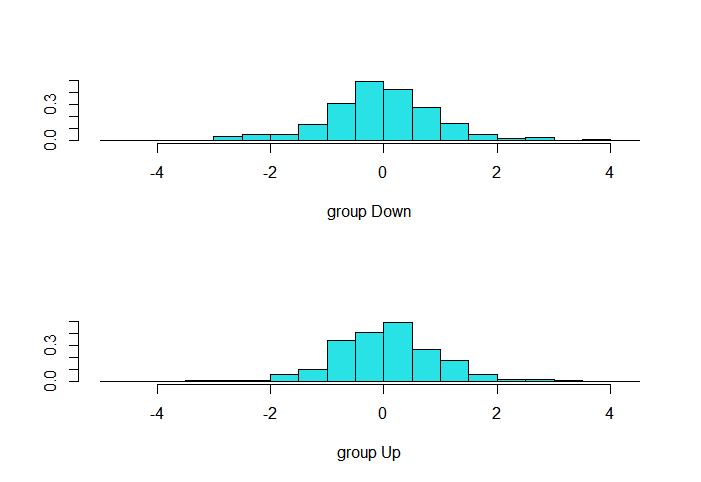
In Lag1 group means of “Down” is 0.427 & “Up” is -0.0395

In Lag2 group means of “Down” is 0.034 & “Up” is -0.0313

In Lag3 group means of “Down” is -0.010 & “Up” is 0.006

The analysis gives the coefficient of Lag1, Lag2, Lag3 to maximize the difference between the group means

> plot(lda.fit)



>lda.pred=predict(lda.fit,Smarket.2005)

> Smarket.2005=Smarket[!train,]

> lda.pred=predict(lda.fit,Smarket.2005)

> names(lda.pred)

[1] "class" "posterior" "x"

> data.frame(lda.pred)[1:5,]

class posterior.Down posterior.Up LD1

999 Up 0.4901932 0.5098068 0.08163079

1000 Up 0.4792223 0.5207777 0.58634400

1001 Up 0.4671730 0.5328270 1.14175192

1002 Up 0.4758412 0.5241588 0.74205347

1003 Up 0.4953589 0.5046411 -0.15583293

> table(lda.pred$class,Direction.2005)

> Direction.2005=Direction[!train]

> table(lda.pred$class,Direction.2005) # we take the categorical data

Direction.2005

Down Up

Down 38 31 when the model say “down” and we have “down” as a result

Up 73 110

> mean(lda.pred$class==Direction.2005) we calculate the accuracy of our model.

> mean(lda.pred$class==Direction.2005)

[1] 0.5873016 #59% accurate (38+110)/dim(Smarket.2005)

# QDA

>qda.fit=qda(Direction~Lag1+Lag2+Lag3,data=Smarket,subset=train)

> qda.fit

> qda.fit=qda(Direction~Lag1+Lag2+Lag3,data=Smarket,subset=train)

> qda.fit

Une image contenant texte

Description générée automatiquement

> qda.pred=predict(qda.fit,Smarket.2005)

> table(qda.pred$class,Direction.2005)

> qda.pred=predict(qda.fit,Smarket.2005)

> data.frame(qda.pred)[1:5,]

class posterior.Down posterior.Up

999 Up 0.4869749 0.5130251

1000 Up 0.4754586 0.5245414

1001 Up 0.4586606 0.5413394

1002 Up 0.4628359 0.5371641

1003 Up 0.4991439 0.5008561

> table(qda.pred$class,Direction.2005)

Direction.2005

Down Up

Down 26 21

Up 85 120

> mean(qda.pred$class==Direction.2005)

> mean(qda.pred$class==Direction.2005)

[1] 0.5793651 #Accuracy is 58% lower than LDA

# KNN

> library(class)

> train.X=cbind(Lag1,Lag2,Lag3)[train ,]

> test.X=cbind (Lag1,Lag2,Lag3)[!train ,]

> train.Direction =Direction [train]

> knn.pred=knn (train .X,test.X,train .Direction ,k=3)

> table(knn .pred ,Direction .2005) # to obtain the confusion matrix

> library(class)

> train.X=cbind(Lag1,Lag2,Lag3)[train,]

> test.X=cbind(Lag1,Lag2,Lag3)[!train,]

> train.Direction=Direction[train]

> set.seed(1)

> knn.pred=knn(train.X,test.X,train.Direction,k=3)

> table(knn.pred,Direction.2005)

Direction.2005

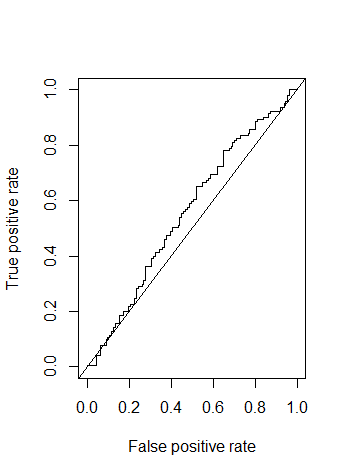
knn.pred Down Up

Down 48 58

Up 63 83

> mean(knn.pred==Direction.2005)

[1] 0

2. Model comparison

Provide ROC curves and AUC for the LDA, QDA and KNN models of Question 1

LDA

> library(ROCR)

> pred\_lda=prediction(lda.pred$posterior[,2],Direction.2005)

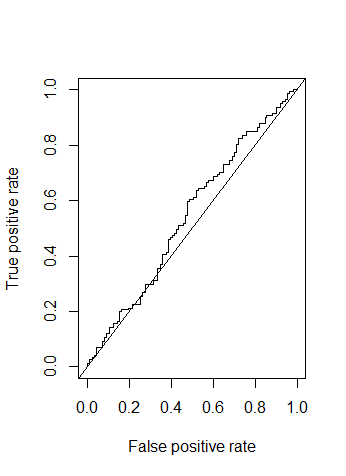
> plot(performance(pred\_lda,'tpr','fpr'))

> abline(0,1)

> auc\_lda=performance(pred\_lda,'auc')

> auc\_lda

A performance instance

 'Area under the ROC curve'

> auc\_lda=auc\_lda@y.values

> auc\_lda

[[1]]

[1] 0.5568334

QDA

> pred\_qda=prediction(qda.pred$posterior[,2],Direction.2005)

> plot(performance(pred\_qda,'tpr','fpr'))

> abline(0,1)

> auc\_qda=performance(pred\_qda,'auc')

> auc\_qda

A performance instance

'Area under the ROC curve'

> auc\_qda=auc\_qda@y.values

> auc\_qda

[[1]]

[1] 0.5430963

KNN

knn.pred

> knn.pred<-as.vector(Direction.2005,mode='numeric')

> Direction.2005<-as.vector(Direction.2005,mode='numeric')

> pred\_knn=prediction(knn.pred,Direction.2005)

> plot(performance(pred\_knn,'tpr', 'fpr'))

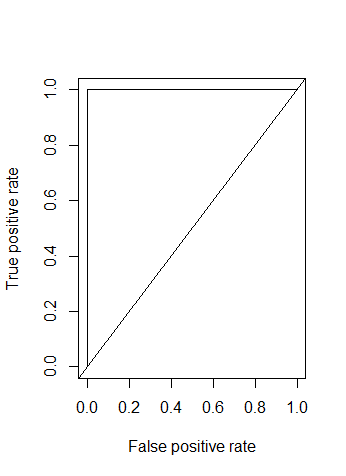
> abline(0,1)

> pred\_knn<-prediction(knn.pred,Direction.2005)

Error: Format of predictions is invalid. It couldn't be coerced to a list.

> knn.pred<-as.vector(Direction.2005,mode='numeric')

> Direction.2005<-as.vector(Direction.2005,mode='numeric')

> pred\_knn=prediction(knn.pred,Direction.2005)

> plot(performance(pred\_knn,'tpr','fpr'))

> abline(0,1)

> auc\_knn=performance(pred\_knn,measure='auc')

> auc\_knn

A performance instance

'Area under the ROC curve'

> auc\_knn=auc\_knn@y.values

> auc\_knn

[[1]]

[1] 1