Performance Classification

Correction

2025-10-07

```r
rm(list=objects())      # supprime les objets existant en session
graphics.off()          # supprime les graphiques existant en session
 # définit le répertoire en cours, à compléter

library(MASS)
data(birthwt) # charge les données dans la session R

birthwt <- within(birthwt,{
  race <- factor(race, labels=c("white", "black", "other"))
  smoke <- factor(smoke,labels=c("No", "Yes"))
  ptl = factor(ptl > 0)
  ht= factor(ht>0)
  ui <- factor(ui,labels=c("No", "Yes"))
  ftv = factor(ftv)
  levels(ftv)[-(1:2)] = "2+"
})
birthwt=birthwt[,-10]
birthwt$low=factor(birthwt$low)
summary(birthwt)
```

```
##  low          age             lwt            race     smoke        ptl
##  0:130   Min.   :14.00   Min.   : 80.0   white:96   No :115   FALSE:
159
##  1: 59   1st Qu.:19.00   1st Qu.:110.0   black:26   Yes: 74   TRUE :
 30
##          Median :23.00   Median :121.0   other:67
##          Mean   :23.24   Mean   :129.8
##          3rd Qu.:26.00   3rd Qu.:140.0
##          Max.   :45.00   Max.   :250.0
##     ht          ui        ftv
##  FALSE:177   No :161   0 :100
##  TRUE : 12   Yes: 28   1 : 47
##                        2+: 42
##
##
##
```

Prédiction par régression logistique

```
t.glm=glm(low~.,data=birthwt, family=binomial)
summary(t.glm)

##
## Call:
## glm(formula = low ~ ., family = binomial, data = birthwt)
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.82302    1.24471   0.661  0.50848
## age         -0.03723    0.03870  -0.962  0.33602
## lwt         -0.01565    0.00708  -2.211  0.02705 *
## raceblack    1.19241    0.53597   2.225  0.02609 *
## raceother    0.74069    0.46174   1.604  0.10869
## smokeYes     0.75553    0.42502   1.778  0.07546 .
## ptlTRUE      1.34376    0.48062   2.796  0.00518 **
## htTRUE       1.91317    0.72074   2.654  0.00794 **
## uiYes        0.68019    0.46434   1.465  0.14296
## ftv1        -0.43638    0.47939  -0.910  0.36268
## ftv2+        0.17901    0.45638   0.392  0.69488
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 234.67  on 188  degrees of freedom
## Residual deviance: 195.48  on 178  degrees of freedom
## AIC: 217.48
##
## Number of Fisher Scoring iterations: 4

# La variable race est-elle significative?
# test de deviance (ou rapport de vraisemblance) à savoir construire
t0.glm=glm(low~.-race,data=birthwt, family=binomial)
TRV=deviance(t0.glm)-deviance(t.glm)
TRV

## [1] 5.751273

# [1] 5.751273
1-pchisq(TRV,2)

## [1] 0.05638025

# [1] 0.05638025 race non significative dans le modèle complet



Y0=data.frame(age=20, lwt=105, race="white", smoke="Yes",ptl="FALSE", h
t="FALSE",ui="No",  ftv="1")
```

```r
Y1=data.frame(age=25,lwt=130,race="black",smoke="No",ptl="TRUE",ht="FALSE",ui="Yes",ftv="0")

predl=predict(t.glm,newdata=Y0) # par défaut type="link"
# -1.24608
sum(coef(t.glm)*c(1,20,105,0,0,1,0,0,0,1,0)) # idem, Combinaison linéaire sum(beta*Y0)
```

```
## [1] -1.246084
```

```r
prob=predict(t.glm,newdata=Y0,type="response") # probabilité P(low=1) estimée
# 0.2233787
# exp(predl)/(1+exp(predl))  idem
# plogis(predl)  # idem

## IC de la prédiction

predl <-predict(t.glm, newdata = Y0, type = "link",se=TRUE)
# IC à savoir justifier
cbind(plogis(predl$fit-qnorm(0.975)*predl$se.fit),
      plogis(predl$fit),
      plogis(predl$fit+qnorm(0.975)*predl$se.fit))
```

```
##          [,1]      [,2]      [,3]
## 1 0.09152853 0.2233787 0.4508949
```

```r
#  0.09152853 0.2233787 0.4508949
# l'IC est < 0.5 on prédit low = 0

# remarque:
binomial()$linkinv(predl$fit) # 0.2233787  idem qu'avec plogis
```

```
##         1
## 0.2233787
```

```r
predl_Y1=predict(t.glm, newdata = Y1, type = "link",se=TRUE)
cbind(plogis(predl_Y1$fit-qnorm(0.975)*predl_Y1$se.fit),
      plogis(predl_Y1$fit),
      plogis(predl_Y1$fit+qnorm(0.975)*predl_Y1$se.fit))
```

```
##         [,1]     [,2]      [,3]
## 1 0.3963216 0.745289 0.9287802
```

```r
#  0.3963216 0.745289 0.9287802
# l'IC n'est pas entièrement >0.5, un peu d'incertitude sur la prédiction low=1
# En ML, règle de classification prédit low=1 car 0.745289>0.5
```

Erreur test

```r
set.seed(2025)

test = sample(1:length(birthwt$low),60)
train = -test
train = birthwt[train, ] #129 observations
test = birthwt[test,]


fit.train=glm(low~.,data=train,family="binomial")
probs=predict(fit.train, newdata=test,type="response")

# classifieur de Bayes
y.pred=rep(0,dim(test)[1])
y.pred[probs>0.5]=1

# ou as.numeric(predict(fit.train,newdata=test,type="response")>0.5)

test.error=mean(y.pred!=test$low)
test.error
```

```
## [1] 0.2666667
```

```r
# [1] 0.2666667

table(y.pred,test$low) # matrice de confusion
```

```
##
## y.pred  0  1
##      0 37 12
##      1  4  7
```

```r
# y.pred  0  1
#      0 37 12
#      1  4  7

sum( probs>=0.5 & test$low==0) # 4 faux positifs
```

```
## [1] 4
```

courbe ROC

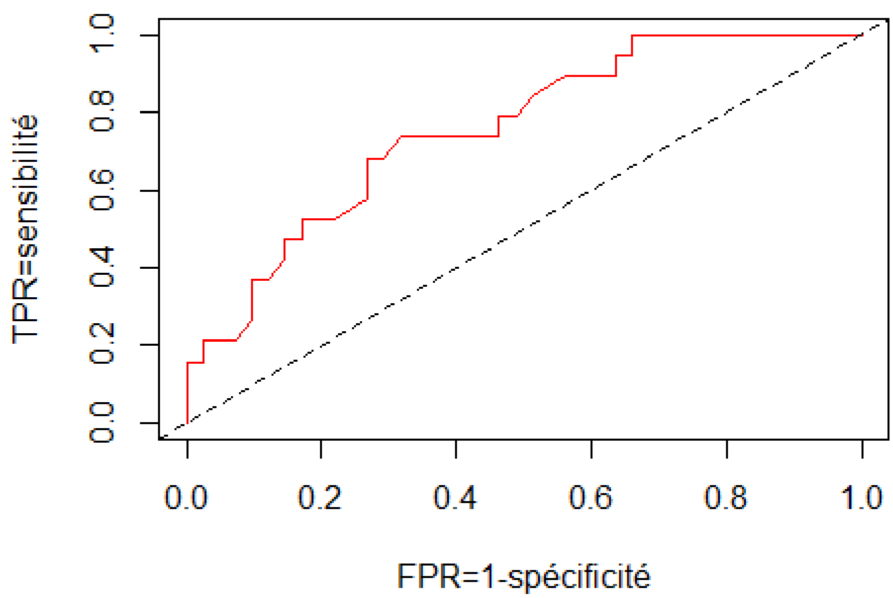
```r
fit.train=glm(low~.,data=train,family="binomial")
probs=predict(fit.train,newdata=test,type="response")

#Initialisation :
s=seq(0,1,.01)
absc=numeric(length(s));ordo=numeric(length(s))

# Courbe Roc reg logistique
```

```
for (i in 1:length(s)){
  ordo[i]=sum( probs>=s[i] & test$low==1)/sum(test$low==1)
  absc[i]=sum( probs>=s[i] & test$low==0)/sum(test$low==0)
}

plot(absc,ordo,col="red",type="l",xlab="FPR=1-spécificité",ylab="TPR=se
nsibilité")
abline(0,1,lty=2)
```



Erreur validation croisée

```
K=5
set.seed(2025)
ind_fold=sample(1:K,nrow(birthwt),replace=TRUE)
error=numeric()
s=0.5

  for (j in 1:K)
  {
    fit.glm=glm(low~.,data=birthwt[ind_fold!=j,],family="binomial")
    probs=predict(fit.glm, newdata=birthwt[ind_fold==j,],type="response
")
    y.pred=rep(0,dim(birthwt[ind_fold==j,])[1])
    y.pred[probs>s]=1
    error[j]=mean(y.pred!=birthwt[ind_fold==j,]$low)
  }
```

```r
error
```

```
## [1] 0.4242424 0.3947368 0.3684211 0.2800000 0.4000000
```

```r
# [1] 0.4242424 0.3947368 0.3684211 0.2800000 0.4000000
cv.error=mean(error)
cv.error
```

```
## [1] 0.3734801
```

```r
# [1] 0.3734801
```