TD4 - Détection de nouveauté par One-class SVM et Kernel PCA

Romain Dudoit, Franck Doronzo, Marie Vachet

12/14/2021

Lecture et description des données

```
setwd("D:/OneDrive/1_Universite/Master 2 Sise/S1/Data Mining - Apprentissage statistique/TP4")
```

```
D = read.table("breast-cancer-wisconsin.data",sep = ",", na.strings = "?")
```

```
print(class(D))
```

```
## [1] "data.frame"
```

```
print(str(D))
```

```
## 'data.frame':    699 obs. of  11 variables:
##  $ V1 : int  1000025 1002945 1015425 1016277 1017023 1017122 1018099 1018561 1033078 1033078 ...
##  $ V2 : int  5 5 3 6 4 8 1 2 2 4 ...
##  $ V3 : int  1 4 1 8 1 10 1 1 1 2 ...
##  $ V4 : int  1 4 1 8 1 10 1 2 1 1 ...
##  $ V5 : int  1 5 1 1 3 8 1 1 1 1 ...
##  $ V6 : int  2 7 2 3 2 7 2 2 2 2 ...
##  $ V7 : int  1 10 2 4 1 10 10 1 1 1 ...
##  $ V8 : int  3 3 3 3 3 9 3 3 1 2 ...
##  $ V9 : int  1 2 1 7 1 7 1 1 1 1 ...
##  $ V10: int  1 1 1 1 1 1 1 1 5 1 ...
##  $ V11: int  2 2 2 2 2 4 2 2 2 2 ...
## NULL
```

```
print(head(D))
```

```
##        V1 V2 V3 V4 V5 V6 V7 V8 V9 V10 V11
## 1 1000025  5  1  1  1  2  1  3  1   1   2
## 2 1002945  5  4  4  5  7 10  3  2   1   2
## 3 1015425  3  1  1  1  2  2  3  1   1   2
## 4 1016277  6  8  8  1  3  4  3  7   1   2
## 5 1017023  4  1  1  3  2  1  3  1   1   2
## 6 1017122  8 10 10  8  7 10  9  7   1   4
```

```
summary(D)
```

```
##        V1                 V2               V3               V4
##  Min.   :   61634   Min.   : 1.000   Min.   : 1.000   Min.   : 1.000
##  1st Qu.:  870688   1st Qu.: 2.000   1st Qu.: 1.000   1st Qu.: 1.000
##  Median : 1171710   Median : 4.000   Median : 1.000   Median : 1.000
##  Mean   : 1071704   Mean   : 4.418   Mean   : 3.134   Mean   : 3.207
##  3rd Qu.: 1238298   3rd Qu.: 6.000   3rd Qu.: 5.000   3rd Qu.: 5.000
##  Max.   :13454352   Max.   :10.000   Max.   :10.000   Max.   :10.000
##
##        V5               V6               V7               V8
##  Min.   : 1.000   Min.   : 1.000   Min.   : 1.000   Min.   : 1.000
##  1st Qu.: 1.000   1st Qu.: 2.000   1st Qu.: 1.000   1st Qu.: 2.000
##  Median : 1.000   Median : 2.000   Median : 1.000   Median : 3.000
##  Mean   : 2.807   Mean   : 3.216   Mean   : 3.545   Mean   : 3.438
##  3rd Qu.: 4.000   3rd Qu.: 4.000   3rd Qu.: 6.000   3rd Qu.: 5.000
##  Max.   :10.000   Max.   :10.000   Max.   :10.000   Max.   :10.000
##                                    NA's   :16
##        V9              V10              V11
##  Min.   : 1.000   Min.   : 1.000   Min.   :2.00
##  1st Qu.: 1.000   1st Qu.: 1.000   1st Qu.:2.00
##  Median : 1.000   Median : 1.000   Median :2.00
##  Mean   : 2.867   Mean   : 1.589   Mean   :2.69
##  3rd Qu.: 4.000   3rd Qu.: 1.000   3rd Qu.:4.00
##  Max.   :10.000   Max.   :10.000   Max.   :4.00
##
```

Séparation des donneés en “train” et “test”

Nous allons pré-traiter les données et séparer celles-ci en deux sous-ensembles disjoints. L’un sera utilisé pour l’apprentissage, l’autre pour le test.

4. La variable D comporte des données manquantes. Identifiez les observations comportant au moins une donnée manquante à l’aide de la commande complete.cases. Vous devez identifier 16 cas.

```
missval<-which(complete.cases(D)== F)
# Retourne la liste des numéros d'observation comportant des valeurs manquantes

print(missval)
```

```
##  [1]  24  41 140 146 159 165 236 250 276 293 295 298 316 322 412 618
```

```
print(length(missval))
```

```
## [1] 16
```

5. Modifiez D de sorte à ce qu'il ne possède que des données complètes

```
D<-D[which(complete.cases(D)),]
```

6. Stockez dans la variable X les variables explicatives qui concernent les colonnes 2 à 10 (inclus) de D.

La variable cible sera stockée dans la variable y qui est donnée par la colonne 11 de D.

```
X<-D[,c(2:10)]
y<-D[,11]
```

7. Recodez y de sorte à ce que les valeurs 2 deviennent des 0 (bégnine) et les valeurs 4 deviennent des 1 (maligne).

```
y<-factor(y)
levels(y) <- c(0,1)
```

8. Stockez dans la variable benin (resp. malin) les indices des observations correspondant à des

tumeurs bégnines (resp. maligne). Vous pourrez utiliser pour cela la commande which.

```
benin<-which(y==0)
malin<-which(y==1)
```

9. Nous garderons dans l'ensemble d'entrainement uniquement les 200 premières observations bégnines.

Stockez dans la variable train_set ces 200 observations. Dans l'ensemble de test vous garderez les observations bégnines qui ne sont pas dans l'ensemble d'entrainement et toutes les observations malignes. Vous stockerez les indices des observations de test dans la variable test set.

```
train_set <- head(benin,200)
Xtrain <- X[train_set,]
ytrain<- y[train_set]

test_set <- c(benin[201:444],malin)
Xtest <- X[test_set,]
ytest <-y[test_set]
```

4 One-class SVM

10. Chargez la librairie e1071.

```
library(e1071)
```

```
## Warning: package 'e1071' was built under R version 4.1.2
```

11 Stockez dans la variable oc_svm_fit les résultats de l'estimation du modèle à partir de l'ensemble d'entraînement.

Vous utiliserez pour cela la commande svm. Vous utiliserez un noyau gaussien de paramètre gamma = 1/2, vous indiquerez que le type de modèle est one-classification.

```
oc_svm_fit <- svm(as.matrix(X[train_set,]),y=NULL,kernel = "radial", type ="one-classification" , gamma
oc_svm_fit
```

```
##
## Call:
## svm.default(x = as.matrix(X[train_set, ]), y = NULL, type = "one-classification",
##     kernel = "radial", gamma = 1/2)
##
##
## Parameters:
##    SVM-Type:  one-classification
##  SVM-Kernel:  radial
##       gamma:  0.5
##          nu:  0.5
##
## Number of Support Vectors:  106
```

12 A l'aide du modèle estimé stocké dans oc svm fit, vous prédirez les scores des observations de test.

Pour cela, utilisez la commande predict et vous indiquerez de façcon adéquate le paramètre decision.values.

```
oc_svm_pred_test  <- predict(oc_svm_fit, X[test_set,],decision.values = TRUE)
```

13. Entrez, exécutez et commentez les commandes suivantes :

attr (oc_svm_pred_test ," decision . values ") oc_svm_score_test = - as . numeric (attr (oc_svm_pred_test ," decision .values "))

```
attr (oc_svm_pred_test ," decision . values ")
```

```
## NULL
```

```
oc_svm_score_test = -as.numeric(attr(oc_svm_pred_test," decision.values "))
```

5 Courbe ROC