HW1-problem1

Xiang Gao

2018年1月30日

Part A.

```
library(caret)
```

```
## Loading required package: lattice
```

```
## Loading required package: ggplot2
```

```
library(klaR)
```

```
## Warning: package 'klaR' was built under R version 3.4.3
```

```
## Loading required package: MASS
```

```
rm(list = ls())


data <- read.csv('diabetes.csv')
X <- data[,-c(9)]
y <- data[,9]

index <- createDataPartition(y=y,p=.8,list=F)

train.x <- X[index,]
train.y <- y[index]
flag <- train.y>0
ptrx <- train.x[flag,]
ntrx <- train.x[!flag,]
tex <- X[-index,]
tey <- y[-index]


ptrmean <- sapply(ptrx,mean,na.rm=T)
ntrmean <- sapply(ntrx,mean,na.rm=T)

ptrsd <- sapply(ptrx,sd,na.rm=T)
ntrsd <- sapply(ntrx,sd,na.rm=T)

pteoffsets <- t(t(tex)-ptrmean)
ptescales <- t(t(pteoffsets)/ptrsd)

ptelogs <- -(1/2)*rowSums(apply(ptescales,c(1,
2), function(x)x^2),na.rm=T)-sum(log(ptrsd))+log(nrow(ptrx)/nrow(train.x))

nteoffsets <- t(t(tex)-ntrmean)
ntescales <- t(t(nteoffsets)/ntrsd)

ntelogs <- -(1/2)*rowSums(apply(ntescales,c(1, 2),function(x)x^2),na.rm=T)-sum(log(ntrsd))+log(nrow(ntrx)
/nrow(train.x))

label <- ptelogs>ntelogs
label1<- label==tey
trscore<- sum(label1)/(sum(label1)+sum(!label1))
```

```
trscore
```

```
## [1] 0.7581699
```

It is plausible

Part B

treat 0 as NA

```
library(caret)
library(klaR)
rm(list = ls())



data <- read.csv('diabetes.csv')
X <- data[,-c(9)]
y <- data[,9]

for (i in c(3,4,6,8)){
        lb <- X[,i]==0
        X[lb,i]=NA
}

index <- createDataPartition(y=y,p=.8,list=F)

train.x <- X[index,]
train.y <- y[index]
flag <- train.y>0
ptrx <- train.x[flag,]
ntrx <- train.x[!flag,]
tex <- X[-index,]
tey <- y[-index]


ptrmean <- sapply(ptrx,mean,na.rm=T)
ntrmean <- sapply(ntrx,mean,na.rm=T)

ptrsd <- sapply(ptrx,sd,na.rm=T)
ntrsd <- sapply(ntrx,sd,na.rm=T)

pteoffsets <- t(t(tex)-ptrmean)
ptescales <- t(t(pteoffsets)/ptrsd)

ptelogs <- -(1/2)*rowSums(apply(ptescales,c(1, 2),
function(x)x^2),na.rm=T)-sum(log(ptrsd))+log(nrow(ptrx)/nrow(train.x))

nteoffsets <- t(t(tex)-ntrmean)
ntescales <- t(t(nteoffsets)/ntrsd)

ntelogs <- -(1/2)*rowSums(apply(ntescales,c(1, 2),
function(x)x^2),na.rm=T)-sum(log(ntrsd))+log(nrow(ntrx)/nrow(train.x))

label <- ptelogs>ntelogs
label1<- label==tey
trscore<- sum(label1)/(sum(label1)+sum(!label1))
```

```
trscore
```

```
## [1] 0.7254902
```

It is plausible

Part C

```
library(klaR)
library(caret)
rm(list = ls())


data<-read.csv('diabetes.csv', header=FALSE)
bigx <- data[,-c(9)]
bigy <- as.factor(data[,9])

for (i in c(3,4,6,8)){
        lb <- bigx[,i]==0
        bigx[lb,i]=NA
}
wtd <- createDataPartition(y=bigy,p=.8,list=F)

trax <- bigx[wtd,]
tray <- bigy[wtd]
model <- train(trax,tray,'nb',trControl =
                      trainControl(method='cv',
                                   number = 10))
teclasses <- predict(model,newdata=bigx[-wtd,])
```

```
sum(teclasses ==bigy[-wtd])/length(bigy[-wtd])
```

```
## [1] 0.6405229
```

It is plausible

Part D use svmlight

```
rm(list=ls())
df<-read.csv('diabetes.csv', header=FALSE)
library(klaR)
library(caret)
bigx<-df[,-c(9)]
bigy<-as.factor(df[,9])
wtd<-createDataPartition(y=bigy, p=.8, list=FALSE)

data <- df[wtd,]
svm<-svmlight(V9~.,data = data)

labels<-predict(svm, bigx[-wtd,])
foo<-labels$class
```

```
sum(foo==bigy[-wtd])/(sum(foo==bigy[-wtd])+sum(!(foo==bigy[-wtd])))
```

```
## [1] 0.6993464
```

It is plausible

reference:http://luthuli.cs.uiuc.edu/~daf/courses/AML-18/aml-home.html