readme

Vladimir

19 12 2019

Test: Adult

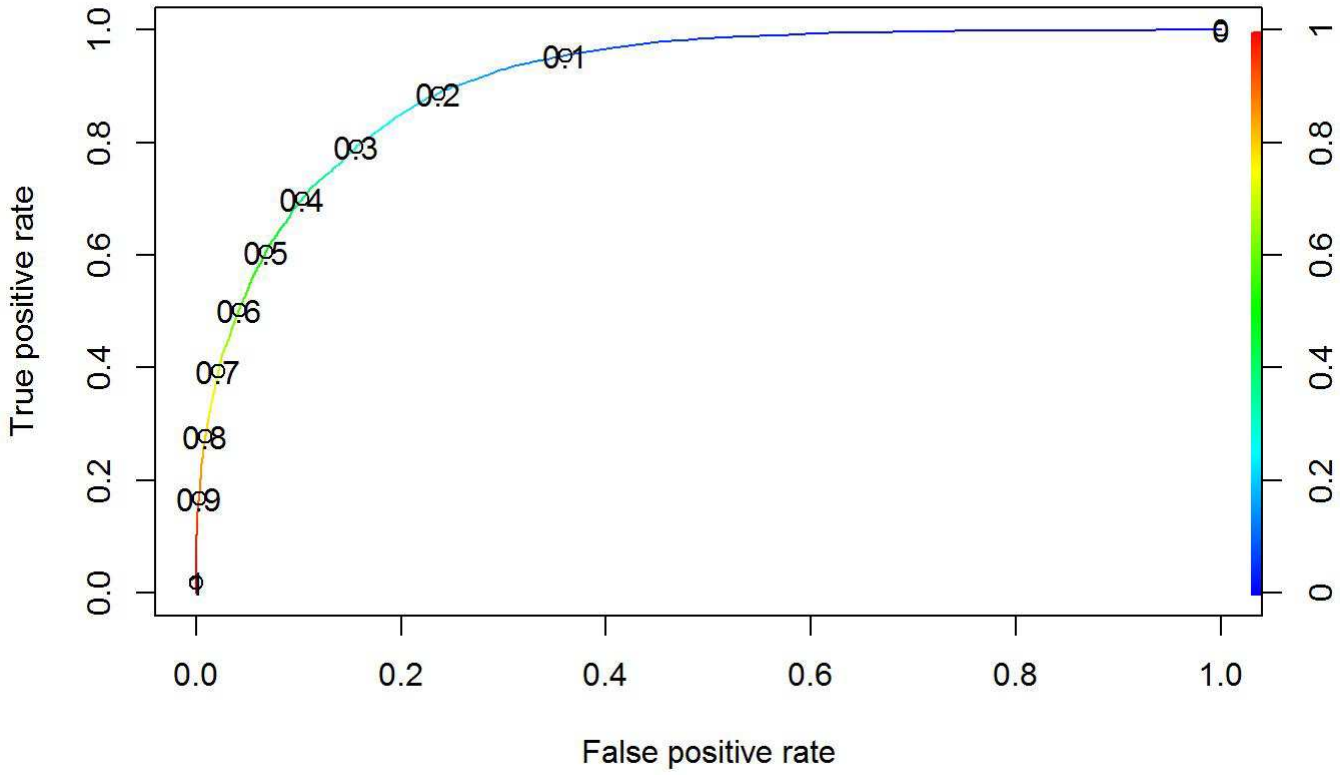
```
library(ROCR)
adult <- read.csv("D:/1_Vladimir_Fuji/6_Grade/Projects/Tests/adult.csv")

#смотрю структуру данных
#summary(adult)
str(adult)
```

```
## 'data.frame':    32561 obs. of  15 variables:
##  $ age           : int  90 82 66 54 41 34 38 74 68 41 ...
##  $ workclass     : Factor w/ 9 levels "?","Federal-gov",..: 1 5 1 5 5 5 5 8 2 5 ...
##  $ fnlwgt        : int  77053 132870 186061 140359 264663 216864 150601 88638 422013 70037
...
##  $ education     : Factor w/ 16 levels "10th","11th",..: 12 12 16 6 16 12 1 11 12 16 ...
##  $ education.num : int  9 9 10 4 10 9 6 16 9 10 ...
##  $ marital.status: Factor w/ 7 levels "Divorced","Married-AF-spouse",..: 7 7 7 1 6 1 6 5 1
5 ...
##  $ occupation    : Factor w/ 15 levels "?","Adm-clerical",..: 1 5 1 8 11 9 2 11 11 4 ...
##  $ relationship  : Factor w/ 6 levels "Husband","Not-in-family",..: 2 2 5 5 4 5 5 3 2 5
...
##  $ race          : Factor w/ 5 levels "Amer-Indian-Eskimo",..: 5 5 3 5 5 5 5 5 5 5 ...
##  $ sex           : Factor w/ 2 levels "Female","Male": 1 1 1 1 1 1 2 1 1 2 ...
##  $ capital.gain  : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ capital.loss  : int  4356 4356 4356 3900 3900 3770 3770 3683 3683 3004 ...
##  $ hours.per.week: int  40 18 40 40 40 45 40 20 40 60 ...
##  $ native.country: Factor w/ 42 levels "?","Cambodia",..: 40 40 40 40 40 40 40 40 40 1 ...
##  $ income        : Factor w/ 2 levels "<=50K",">50K": 1 1 1 1 1 1 1 2 1 2 ...
```

```
#строю модель логистической регрессии и смотрю на значимость предикторов
fit  <- glm(income ~ ., adult, family = "binomial")
#summary(fit)

#строю альтернативные модели с уменьшением количества предикторов
fit1 = glm(income ~ age + occupation + education + marital.status + relationship
           + race + sex + native.country + capital.gain + capital.loss +
             hours.per.week, adult, family = "binomial")
fit2 = glm(income ~ age + education + marital.status + relationship + race + sex
           + native.country + capital.gain + capital.loss + hours.per.week,
           adult, family = "binomial")
fit3 = glm(income ~ age + occupation + education + marital.status + relationship
           + race + sex + capital.gain + capital.loss + hours.per.week,
           adult, family = "binomial")

#сравниваю альтернативные модели с исходной
anova(fit, fit3, test = 'Chisq')
```

```
## Analysis of Deviance Table
##
## Model 1: income ~ age + workclass + fnlwgt + education + education.num +
##     marital.status + occupation + relationship + race + sex +
##     capital.gain + capital.loss + hours.per.week + native.country
## Model 2: income ~ age + occupation + education + marital.status + relationship +
##     race + sex + capital.gain + capital.loss + hours.per.week
##   Resid. Df Resid. Dev  Df Deviance  Pr(>Chi)
## 1     32462      20565
## 2     32511      20801 -49  -236.22 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

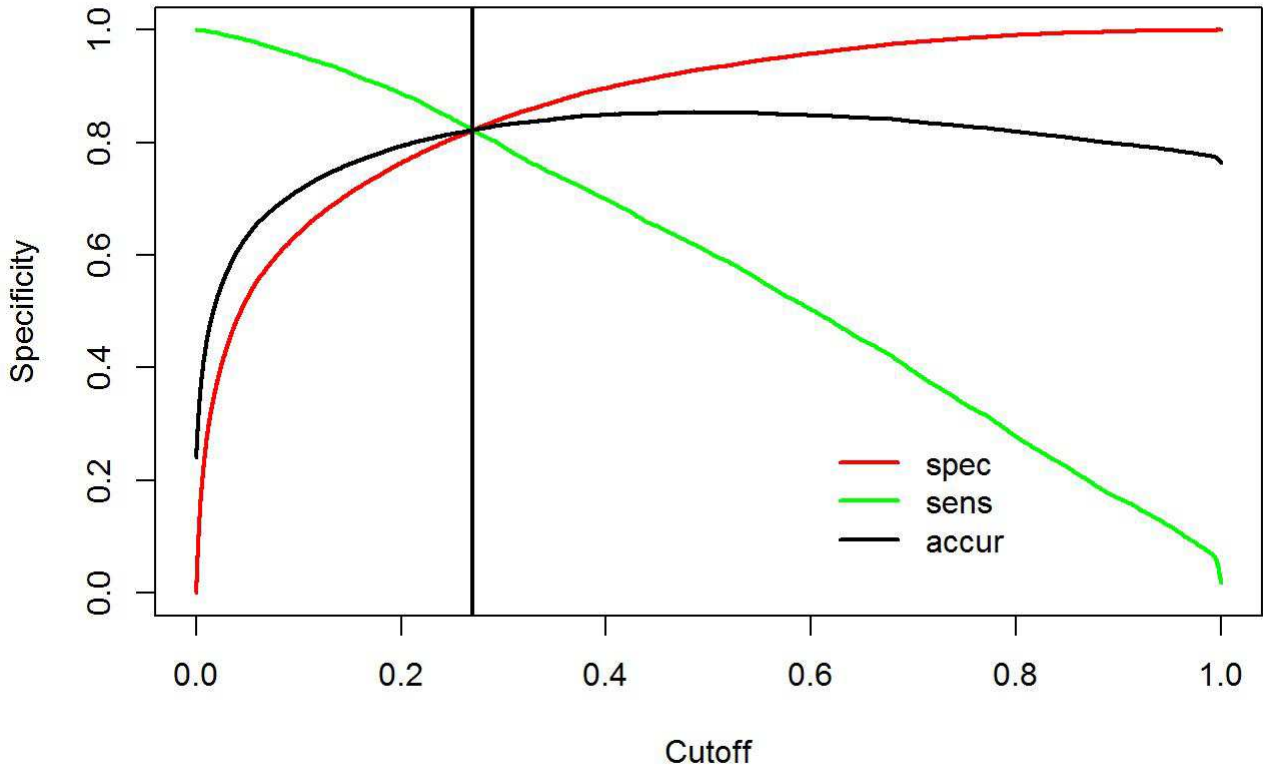
```
#добавляю колонку с предсказанием вероятности значений колонки "income"
adult$prob  <- predict(object = fit, type = "response")

#строю ROC - кривую и смотрю на AUC (довольно не плохая)
pred_fit <- prediction(adult$prob, adult$income)
perf_fit <- performance(pred_fit,"tpr","fpr")
plot(perf_fit, colorize=T , print.cutoffs.at = seq(0,1,by=0.1))
```




```
auc  <- performance(pred_fit, measure = "auc")
str(auc)
```

```
## Formal class 'performance' [package "ROCR"] with 6 slots
##   ..@ x.name      : chr "None"
##   ..@ y.name      : chr "Area under the ROC curve"
##   ..@ alpha.name  : chr "none"
##   ..@ x.values    : list()
##   ..@ y.values    :List of 1
##   .. ..$ : num 0.909
##   ..@ alpha.values: list()
```

```
#специфичность (TNR)
perf3  <- performance(pred_fit, x.measure = "cutoff", measure = "spec")
#чувствительность (TPR)
perf4  <- performance(pred_fit, x.measure = "cutoff", measure = "sens")
#точность правильных классификаций
perf5  <- performance(pred_fit, x.measure = "cutoff", measure = "acc")

#строю графики
plot(perf3, col = "red", lwd =2)
plot(add=T, perf4 , col = "green", lwd =2)
plot(add=T, perf5, lwd =2)
legend(x = 0.6,y = 0.3, c("spec", "sens", "accur"),
       lty = 1, col =c('red', 'green', 'black'), bty = 'n', cex = 1, lwd = 2)

#и подбираю точку пересечения
abline(v= 0.27, lwd = 2)
```



```r
#добавляю новую переменную с предсказанным значением
adult$pred_resp  <- factor(ifelse(adult$prob > 0.27, 1, 0), labels = c("<=50K", ">50K"))

#сравниваю предсказанное значение с истинным в выборке
adult$correct  <- ifelse(adult$pred_resp == adult$income, 1, 0)

#% правильно предсказанных значений
mean(adult$correct)
```

```
## [1] 0.8218421
```

```r
#количество правильно предсказанных значение
sum(adult$correct)
```

```
## [1] 26760
```