hw6_c50.R

lukemcevoy

2021-11-15

```
# clear the environment
rm(list=ls())

# select the data
# filename<-file.choose()
filename<-'/Users/lukemcevoy/Develop/stevens/f21/dataMining/week10/hw6/breast-cancer-wisconsin.csv'
cancer<-read.csv(filename, na.strings = "?",
                 colClasses=c("Sample"="character",
                              "F1"="factor","F2"="factor","F3"="factor",
                              "F4"="factor","F5"="factor","F6"="factor",
                              "F7"="factor","F8"="factor","F9"="factor",
                              "Class"="factor"))

# split data
index<-sort(sample(nrow(cancer), round(.3*nrow(cancer))))
training<-cancer[-index,]
test<-cancer[index,]

# C50 classification
library('C50')
summary(cancer)
```

```
##     Sample                F1            F2            F3            F4
##  Length:699         1      :145   1      :384   1      :353   1      :407
##  Class :character   5      :130   10     : 67   2      : 59   2      : 58
##  Mode  :character   3      :108   3      : 52   10     : 58   3      : 58
##                     4      : 80   2      : 45   3      : 56   10     : 55
##                     10     : 69   4      : 40   4      : 44   4      : 33
##                     2      : 50   5      : 30   5      : 34   8      : 25
##                     (Other):117   (Other): 81   (Other): 95   (Other): 63
##        F5            F6            F7            F8            F9      Class
##  2      :386   1      :402   2      :166   1      :443   1      :579   2:458
##  3      : 72   10     :132   3      :165   10     : 61   2      : 35   4:241
##  4      : 48   2      : 30   1      :152   3      : 44   3      : 33
##  1      : 47   5      : 30   7      : 73   2      : 36   10     : 14
##  6      : 41   3      : 28   4      : 40   8      : 24   4      : 12
##  5      : 39   (Other): 61   5      : 34   6      : 22   7      :  9
##  (Other): 66   NA's   : 16   (Other): 69   (Other): 69   (Other): 17
```

```
C50_class <- C5.0(Class~., data=cancer)

summary(C50_class)
```

```
##
## Call:
## C5.0.formula(formula = Class ~ ., data = cancer)
##
##
## C5.0 [Release 2.07 GPL Edition]      Mon Nov 15 16:28:36 2021
## -------------------------------
##
## Class specified by attribute 'outcome'
##
## Read 699 cases (11 attributes) from undefined.data
##
## Decision tree:
##
## F2 in {1,2}: 2 (429/12)
## F2 in {10,3,4,5,6,7,8,9}: 4 (270/41)
##
##
## Evaluation on training data (699 cases):
##
##      Decision Tree
##    ----------------
##    Size      Errors
##
##       2   53( 7.6%)   <<
##
##
##     (a)   (b)    <-classified as
##    ----  ----
##     417    41    (a): class 2
##      12   229    (b): class 4
##
##
##  Attribute usage:
##
##  100.00% F2
##
##
## Time: 0.0 secs
```

```
C50_predict<-predict(C50_class, test, type="class")
table(actual=test[,11], C50=C50_predict)
```

```
##       C50
## actual   2   4
##      2 125  12
##      4   2  71
```

```
wrong<-(test[,11]!=C50_predict)
C50_rate<-sum(wrong)/length(test[,11])
C50_rate
```

```
## [1] 0.06666667
```