hw6_randomForest.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,
                 colClasses=c("Sample"="character",
                              "F1"="factor","F2"="factor","F3"="factor",
                              "F4"="factor","F5"="factor","F6"="factor",
                              "F7"="factor","F8"="factor","F9"="factor",
                              "Class"="factor"))
cancer<-na.omit(cancer)

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

# random forest
library('randomForest')
```

```
## randomForest 4.6-14

## Type rfNews() to see new features/changes/bug fixes.
```

```
summary(cancer)
```

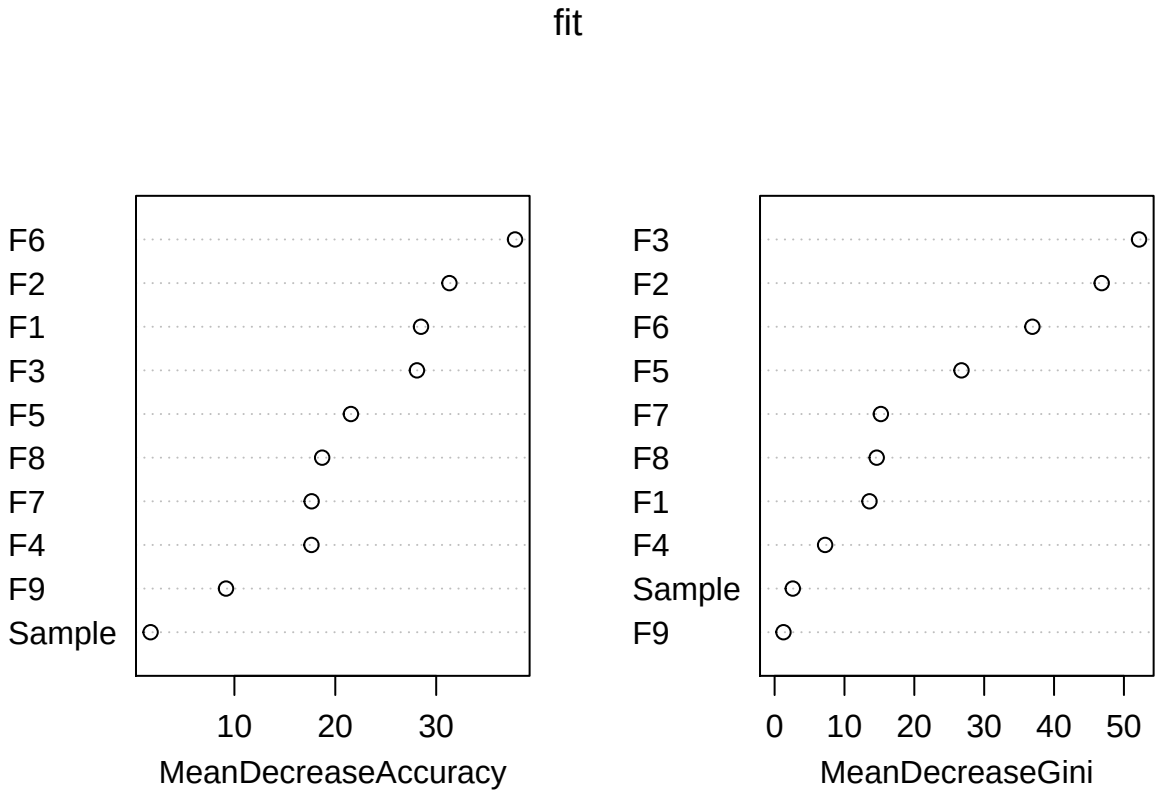
```
##     Sample                F1            F2            F3            F4
##  Length:699         1      :145   1      :384   1      :353   1      :407
##  Class :character   5      :130   10     : 67   2      : 59   2      : 58
##  Mode  :character   3      :108   3      : 52   10     : 58   3      : 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   8       : 21   5       : 34   6       : 22   7       :  9
##  (Other): 66   (Other): 56   (Other): 69   (Other): 69   (Other): 17
```

```
fit<-randomForest(Class~., data=training, importance=TRUE, ntree=1000)
importance(fit)
```

```
##                2         4 MeanDecreaseAccuracy MeanDecreaseGini
## Sample  0.439978  1.694696             1.692628         2.606063
## F1     25.778504 19.882201            28.493850        13.581469
## F2     24.411617 20.095107            31.310313        46.821958
## F3     21.337726 20.137170            28.092947        52.165965
## F4     13.871573 14.939995            17.635893         7.229786
## F5     19.224496 10.221705            21.547108        26.750229
## F6     31.497759 25.582871            37.814153        36.895518
## F7     14.148013 12.512120            17.654893        15.210175
## F8     18.140469  7.050049            18.697535        14.614127
## F9      8.180003  5.205873             9.168024         1.269559
```

```
varImpPlot(fit)
```




```
Prediction<-predict(fit, test)
table(actual=test[,11],Prediction)
```

```
##       Prediction
```

```
## actual   2   4
##      2 130   3
##      4   2  75
```

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

```
## [1] 0.02380952
```