Performance Estimation

Solutions to Hands On Exercises

L. Torgo

ltorgo@knoyda.com

KNOYDA, Know Your Data!

Jul, 2019

Hands on Performance Estimation

the Algae data set

Load in the data set `algae` and answer the following questions:

1. Estimate the MSE of a regression tree for forecasting alga *a1* using 10-fold Cross validation. solution
2. Repeat the previous exercise this time trying some variants of random forests. Check what are the characteristics of the best performing variant. solution
3. Compare the results in terms of mean absolute error of the default variants of a regression tree, a linear regression model and a random forest, in the task of predicting alga a3. Use 2 repetitions of a 5-fold Cross Validation experiment. solution
4. Carry out an experiment designed to select what are the best models for each of the seven harmful algae. Use 10-fold Cross Validation. For illustrative purposes consider only the default variants of regression trees, linear regression and random forests. solution

Solutions to Exercise 1

- Estimate the MSE of a regression tree for forecasting alga *a1* using 10-fold Cross validation.

```r
library(DMwR)
library(performanceEstimation)
data(algae)
algae <- algae[-c(62,199),]
res.a1 <- performanceEstimation(
    PredTask(a1 ~ .,algae[,1:12],"algaA1"),
    Workflow("standardWF",learner="rpartXse",pre="knnImp"),
    EstimationTask("mse",method=CV())
    )
```

Solutions to Exercise 1 (cont.)

- Estimate the MSE of a regression tree for forecasting alga *a1* using 10-fold Cross validation.

```
summary(res.a1)


##
## == Summary of a  Cross Validation Performance Estimation Experiment ==
##
## Task for estimating  mse  using
##  1 x 10 - Fold Cross Validation
##   Run with seed =  1234
##
## * Predictive Tasks ::  algaA1
## * Workflows  ::  standardWF
##
## -> Task:  algaA1
##   *Workflow: standardWF
##            mse
## avg     321.10
## std     200.70
## med     302.53
## iqr     302.49
## min      96.22
## max     637.87
## invalid   0.00
```

Go back

Solutions to Exercise 2

- Repeat the previous exercise this time trying some variants of random forests. Check what are the characteristics of the best performing variant.

```
library(randomForest)
resrf.a1 <- performanceEstimation(
    PredTask(a1 ~ .,algae[,1:12],"algaA1"),
    workflowVariants("standardWF",
                     learner="randomForest",
                     learner.pars=list(ntree=c(500,750,1000)),
                     pre="knnImp"),
    EstimationTask("mse",method=CV())
    )
```

Solutions to Exercise 2 (cont.)

```
summary(resrf.a1)


##
## == Summary of a  Cross Validation Performance Estimation Experiment ==
##
## Task for estimating  mse  using
##  1 x 10 - Fold Cross Validation
##   Run with seed =  1234
##
## * Predictive Tasks ::  algaA1
## * Workflows  ::  randomForest.v1, randomForest.v2, randomForest.v3
##
## -> Task:  algaA1
##   *Workflow: randomForest.v1
##             mse
## avg     255.79
## std     167.89
## med     200.92
## iqr     178.99
## min      73.26
## max     640.69
## invalid   0.00
##
##   *Workflow: randomForest.v2
##             mse
## avg     256.09
## std     166.75
## med     203.92
## iqr     172.09
## min      74.05
```

Solutions to Exercise 2 (cont.)

- Repeat the previous exercise this time trying some variants of random forests. Check what are the characteristics of the best performing variant.

```
topPerformer(resrf.a1,"mse","algaA1")

## Workflow Object:
##  Workflow ID       ::  randomForest.v1
##  Workflow Function ::  standardWF
##       Parameter values:
##   learner.pars  ->  ntree=500
##   learner  ->  randomForest
```

Go back

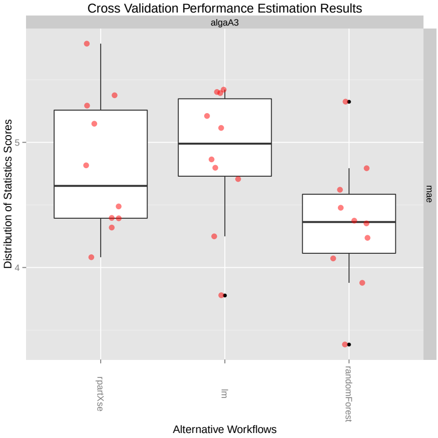
Solutions to Exercise 3

- Compare the results in terms of mean absolute error of the default variants of a regression tree, a linear regression model and a random forest, in the task of predicting alga a3. Use 2 repetitions of a 5-fold Cross Validation experiment. Plot the results

```
res.a3 <- performanceEstimation(
    PredTask(a3 ~ ., algae[,c(1:11,14)], "algaA3"),
    workflowVariants("standardWF",
                     learner=c("rpartXse","lm","randomForest"),
                     pre="knnImp"),
    EstimationTask("mae",method=CV(nReps=2,nFolds=5))
    )
```

Solutions to Exercise 3 (cont.)

```
plot(res.a3)
```



Go back

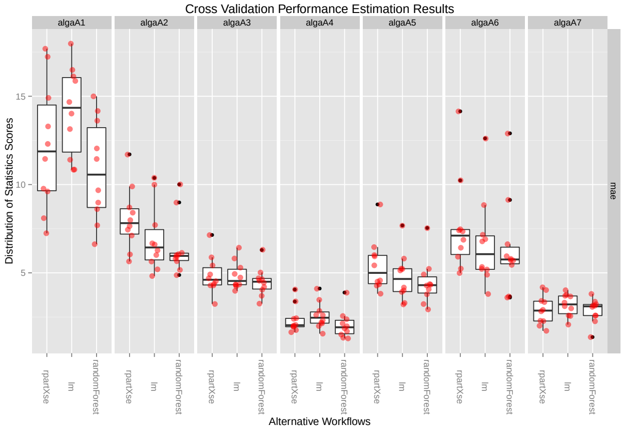
Solutions to Exercise 4

- Carry out an experiment designed to select what are the best models for each of the seven harmful algae. Use 10-fold Cross Validation. For illustrative purposes consider only the default variants of regression trees, linear regression and random forests.

```
tgts <- 12:18
tasks <- c()
for(t in tgts)
    tasks <- c(tasks,
               PredTask(as.formula(paste(colnames(algae)[t],'~ .')),
                        algae[,c(1:11,t)],
                        paste0("algaA",t-11),
                        copy=TRUE))
res.algae <- performanceEstimation(
    tasks,
    workflowVariants(learner=c("rpartXse","lm","randomForest"),
                     pre="knnImp"),
    EstimationTask("mae",method=CV())
    )
```

Solutions to Exercise 4 (cont.)

```r
plot(res.algae)
```



Solutions to Exercise 4 (cont.)

```
topPerformers(res.algae)
```

```
## $algaA1
##          Workflow Estimate
## mae randomForest   10.785
##
## $algaA2
##          Workflow Estimate
## mae randomForest    6.461
##
## $algaA3
##          Workflow Estimate
## mae randomForest    4.486
##
## $algaA4
##          Workflow Estimate
## mae randomForest    2.059
##
## $algaA5
##          Workflow Estimate
## mae randomForest    4.466
##
## $algaA6
##          Workflow Estimate
## mae randomForest    6.453
##
## $algaA7
##          Workflow Estimate
## mae randomForest    2.855
```