R Notebook

Practice Assignment 9

```
#import library's
library('mlr3')
library('mlr3learners')
library('mlr3pipelines')
library('mlr3tuning')
```

```
## Loading required package: paradox
```

```
library('paradox')
```

```
#import data
df<-read.csv("titanic.csv")[, c("Survived","Pclass","Sex","Age","Fare","Embarked")]
head(df)
```

```
##   Survived Pclass    Sex Age    Fare Embarked
## 1        0      3   male  22  7.2500        S
## 2        1      1 female  38 71.2833        C
## 3        1      3 female  26  7.9250        S
## 4        1      1 female  35 53.1000        S
## 5        0      3   male  35  8.0500        S
## 6        0      3   male  NA  8.4583        Q
```

Passengers gender and ticket class based on survival

```
#linear regression

df2<-read.csv("titanic.csv")[, c("Survived","Pclass","Sex")]

task <- TaskRegr$new('titanic2', backend=df2, target = 'Survived')
measure <- msr('regr.mse')

learner_lm <- lrn('regr.lm')

gr_lm <- po('imputemean') %>>%
  po(learner_lm)
glrn_lm <- GraphLearner$new(gr_lm)

set.seed(1)
train_set <- sample(task$nrow, 0.7 * task$nrow)
test_set <- setdiff(seq_len(task$nrow), train_set)
glrn_lm$train(task, row_ids = train_set)
glrn_lm$predict(task, row_ids = test_set)$score()
```

```
##  regr.mse
## 0.1399328
```

Passengers age per ticket class based on survival

```
#linear regression

df3<-read.csv("titanic.csv")[, c("Survived","Pclass","Age")]

task <- TaskRegr$new('titanic3', backend=df3, target = 'Survived')
measure <- msr('regr.mse')

learner_lm <- lrn('regr.lm')

gr_lm <- po('imputemean') %>>%
  po(learner_lm)
glrn_lm <- GraphLearner$new(gr_lm)

set.seed(1)
train_set <- sample(task$nrow, 0.7 * task$nrow)
test_set <- setdiff(seq_len(task$nrow), train_set)
glrn_lm$train(task, row_ids = train_set)
glrn_lm$predict(task, row_ids = test_set)$score()
```

```
##  regr.mse
## 0.1972198
```

Family size per ticket class

```
df4<-read.csv("titanic.csv")[, c("Survived","Pclass","SibSp","Parch")]
head(df4)
```

```
##   Survived Pclass SibSp Parch
## 1        0      3     1     0
## 2        1      1     1     0
## 3        1      3     0     0
## 4        1      1     1     0
## 5        0      3     0     0
## 6        0      3     0     0
```

```
#create new column of family size
df4$Family_class <- df4$SibSp + df4$Parch
```

```
task <- TaskRegr$new('titanic4', backend=df4, target = 'Survived')
```

```
#ridge regression
learner_ridge <- lrn('regr.glmnet')
learner_ridge$param_set$values <- list(alpha = 0)
gr_ridge <- po('scale') %>>%
  po('imputemean') %>>%
  po(learner_ridge)
glrn_ridge <- GraphLearner$new(gr_ridge)
```

```r
#Tuning environment
tune_lambda <- ParamSet$new (list(
 ParamDbl$new('regr.glmnet.lambda', lower = 0.001, upper = 2)
))
tuner<-tnr('grid_search')
terminator <- trm('evals', n_evals = 20)
#Combine new learner
at_ridge <- AutoTuner$new(
  learner = glrn_ridge,
  resampling = rsmp('cv', folds = 3),
  measure = measure,
  search_space = tune_lambda,
  terminator = terminator,
  tuner = tuner
)
#Train learner
at_ridge$train(task, row_ids = train_set)
```

```
## INFO  [11:37:40.603] [bbotk] Starting to optimize 1 parameter(s) with '<TunerGridSearch>' and '<Term
## INFO  [11:37:40.623] [bbotk] Evaluating 1 configuration(s)
## INFO  [11:37:40.656] [mlr3] Running benchmark with 3 resampling iterations
## INFO  [11:37:40.689] [mlr3] Applying learner 'scale.imputemean.regr.glmnet' on task 'titanic4' (iter
## INFO  [11:37:41.437] [mlr3] Applying learner 'scale.imputemean.regr.glmnet' on task 'titanic4' (iter
## INFO  [11:37:41.531] [mlr3] Applying learner 'scale.imputemean.regr.glmnet' on task 'titanic4' (iter
## INFO  [11:37:41.627] [mlr3] Finished benchmark
## INFO  [11:37:41.648] [bbotk] Result of batch 1:
## INFO  [11:37:41.649] [bbotk]  regr.glmnet.lambda  regr.mse runtime_learners
## INFO  [11:37:41.649] [bbotk]           0.4452222 0.2207191             0.89
## INFO  [11:37:41.649] [bbotk]                                       uhash
## INFO  [11:37:41.649] [bbotk]  8d4f786e-c6f3-4b41-b0e4-bd2a3a4b0ff6
## INFO  [11:37:41.650] [bbotk] Evaluating 1 configuration(s)
## INFO  [11:37:41.675] [mlr3] Running benchmark with 3 resampling iterations
## INFO  [11:37:41.679] [mlr3] Applying learner 'scale.imputemean.regr.glmnet' on task 'titanic4' (iter
## INFO  [11:37:41.780] [mlr3] Applying learner 'scale.imputemean.regr.glmnet' on task 'titanic4' (iter
## INFO  [11:37:41.878] [mlr3] Applying learner 'scale.imputemean.regr.glmnet' on task 'titanic4' (iter
## INFO  [11:37:41.974] [mlr3] Finished benchmark
## INFO  [11:37:42.000] [bbotk] Result of batch 2:
## INFO  [11:37:42.001] [bbotk]  regr.glmnet.lambda  regr.mse runtime_learners
## INFO  [11:37:42.001] [bbotk]               0.001 0.2157293             0.28
## INFO  [11:37:42.001] [bbotk]                                       uhash
## INFO  [11:37:42.001] [bbotk]  3ba436f9-b025-4f52-9f79-91569c783d33
## INFO  [11:37:42.002] [bbotk] Evaluating 1 configuration(s)
## INFO  [11:37:42.027] [mlr3] Running benchmark with 3 resampling iterations
## INFO  [11:37:42.030] [mlr3] Applying learner 'scale.imputemean.regr.glmnet' on task 'titanic4' (iter
## INFO  [11:37:42.124] [mlr3] Applying learner 'scale.imputemean.regr.glmnet' on task 'titanic4' (iter
## INFO  [11:37:42.222] [mlr3] Applying learner 'scale.imputemean.regr.glmnet' on task 'titanic4' (iter
## INFO  [11:37:42.321] [mlr3] Finished benchmark
## INFO  [11:37:42.342] [bbotk] Result of batch 3:
## INFO  [11:37:42.343] [bbotk]  regr.glmnet.lambda  regr.mse runtime_learners
## INFO  [11:37:42.343] [bbotk]           0.6673333 0.2233411             0.26
## INFO  [11:37:42.343] [bbotk]                                       uhash
## INFO  [11:37:42.343] [bbotk]  01599b09-417f-4b28-ad74-3596f0322a7f
## INFO  [11:37:42.343] [bbotk] Evaluating 1 configuration(s)
```

```
## INFO  [11:37:42.368] [mlr3] Running benchmark with 3 resampling iterations
## INFO  [11:37:42.372] [mlr3] Applying learner 'scale.imputemean.regr.glmnet' on task 'titanic4' (iter
## INFO  [11:37:42.469] [mlr3] Applying learner 'scale.imputemean.regr.glmnet' on task 'titanic4' (iter
## INFO  [11:37:42.564] [mlr3] Applying learner 'scale.imputemean.regr.glmnet' on task 'titanic4' (iter
## INFO  [11:37:42.662] [mlr3] Finished benchmark
## INFO  [11:37:42.683] [bbotk] Result of batch 4:
## INFO  [11:37:42.684] [bbotk]  regr.glmnet.lambda  regr.mse runtime_learners
## INFO  [11:37:42.684] [bbotk]                   2 0.2312042            0.26
## INFO  [11:37:42.684] [bbotk]                                       uhash
## INFO  [11:37:42.684] [bbotk]  5ac63b8c-9b63-4ef1-a2a9-bc312929c51d
## INFO  [11:37:42.685] [bbotk] Evaluating 1 configuration(s)
## INFO  [11:37:42.710] [mlr3] Running benchmark with 3 resampling iterations
## INFO  [11:37:42.714] [mlr3] Applying learner 'scale.imputemean.regr.glmnet' on task 'titanic4' (iter
## INFO  [11:37:42.812] [mlr3] Applying learner 'scale.imputemean.regr.glmnet' on task 'titanic4' (iter
## INFO  [11:37:42.907] [mlr3] Applying learner 'scale.imputemean.regr.glmnet' on task 'titanic4' (iter
## INFO  [11:37:43.003] [mlr3] Finished benchmark
## INFO  [11:37:43.026] [bbotk] Result of batch 5:
## INFO  [11:37:43.027] [bbotk]  regr.glmnet.lambda regr.mse runtime_learners
## INFO  [11:37:43.027] [bbotk]            1.333667 0.228418            0.27
## INFO  [11:37:43.027] [bbotk]                                       uhash
## INFO  [11:37:43.027] [bbotk]  f273ff6d-07e1-41af-a554-23c19d3e3049
## INFO  [11:37:43.028] [bbotk] Evaluating 1 configuration(s)
## INFO  [11:37:43.053] [mlr3] Running benchmark with 3 resampling iterations
## INFO  [11:37:43.057] [mlr3] Applying learner 'scale.imputemean.regr.glmnet' on task 'titanic4' (iter
## INFO  [11:37:43.154] [mlr3] Applying learner 'scale.imputemean.regr.glmnet' on task 'titanic4' (iter
## INFO  [11:37:43.252] [mlr3] Applying learner 'scale.imputemean.regr.glmnet' on task 'titanic4' (iter
## INFO  [11:37:43.351] [mlr3] Finished benchmark
## INFO  [11:37:43.372] [bbotk] Result of batch 6:
## INFO  [11:37:43.373] [bbotk]  regr.glmnet.lambda regr.mse runtime_learners
## INFO  [11:37:43.373] [bbotk]            1.555778 0.229516            0.27
## INFO  [11:37:43.373] [bbotk]                                       uhash
## INFO  [11:37:43.373] [bbotk]  bfc85c4a-fc05-4876-9cb7-473b73f60de9
## INFO  [11:37:43.373] [bbotk] Evaluating 1 configuration(s)
## INFO  [11:37:43.398] [mlr3] Running benchmark with 3 resampling iterations
## INFO  [11:37:43.401] [mlr3] Applying learner 'scale.imputemean.regr.glmnet' on task 'titanic4' (iter
## INFO  [11:37:43.497] [mlr3] Applying learner 'scale.imputemean.regr.glmnet' on task 'titanic4' (iter
## INFO  [11:37:43.597] [mlr3] Applying learner 'scale.imputemean.regr.glmnet' on task 'titanic4' (iter
## INFO  [11:37:43.696] [mlr3] Finished benchmark
## INFO  [11:37:43.718] [bbotk] Result of batch 7:
## INFO  [11:37:43.719] [bbotk]  regr.glmnet.lambda  regr.mse runtime_learners
## INFO  [11:37:43.719] [bbotk]            1.111556 0.2270802            0.29
## INFO  [11:37:43.719] [bbotk]                                       uhash
## INFO  [11:37:43.719] [bbotk]  4b413899-cc6c-4c9a-864b-11da3eaab104
## INFO  [11:37:43.719] [bbotk] Evaluating 1 configuration(s)
## INFO  [11:37:43.744] [mlr3] Running benchmark with 3 resampling iterations
## INFO  [11:37:43.748] [mlr3] Applying learner 'scale.imputemean.regr.glmnet' on task 'titanic4' (iter
## INFO  [11:37:43.846] [mlr3] Applying learner 'scale.imputemean.regr.glmnet' on task 'titanic4' (iter
## INFO  [11:37:43.946] [mlr3] Applying learner 'scale.imputemean.regr.glmnet' on task 'titanic4' (iter
## INFO  [11:37:44.045] [mlr3] Finished benchmark
## INFO  [11:37:44.067] [bbotk] Result of batch 8:
## INFO  [11:37:44.068] [bbotk]  regr.glmnet.lambda regr.mse runtime_learners
## INFO  [11:37:44.068] [bbotk]            1.777889 0.230431            0.28
## INFO  [11:37:44.068] [bbotk]                                       uhash
## INFO  [11:37:44.068] [bbotk]  fa0386ac-ac44-4284-a574-f1600688766c
```

```
## INFO  [11:37:44.068] [bbotk] Evaluating 1 configuration(s)
## INFO  [11:37:44.093] [mlr3] Running benchmark with 3 resampling iterations
## INFO  [11:37:44.097] [mlr3] Applying learner 'scale.imputemean.regr.glmnet' on task 'titanic4' (iter
## INFO  [11:37:44.194] [mlr3] Applying learner 'scale.imputemean.regr.glmnet' on task 'titanic4' (iter
## INFO  [11:37:44.327] [mlr3] Applying learner 'scale.imputemean.regr.glmnet' on task 'titanic4' (iter
## INFO  [11:37:44.421] [mlr3] Finished benchmark
## INFO  [11:37:44.448] [bbotk] Result of batch 9:
## INFO  [11:37:44.450] [bbotk]  regr.glmnet.lambda  regr.mse runtime_learners
## INFO  [11:37:44.450] [bbotk]           0.2231111 0.2176152              0.3
## INFO  [11:37:44.450] [bbotk]                                      uhash
## INFO  [11:37:44.450] [bbotk]  48916d7f-26c6-4d1f-9286-cc74fd0114c3
## INFO  [11:37:44.451] [bbotk] Evaluating 1 configuration(s)
## INFO  [11:37:44.476] [mlr3] Running benchmark with 3 resampling iterations
## INFO  [11:37:44.480] [mlr3] Applying learner 'scale.imputemean.regr.glmnet' on task 'titanic4' (iter
## INFO  [11:37:44.578] [mlr3] Applying learner 'scale.imputemean.regr.glmnet' on task 'titanic4' (iter
## INFO  [11:37:44.683] [mlr3] Applying learner 'scale.imputemean.regr.glmnet' on task 'titanic4' (iter
## INFO  [11:37:44.783] [mlr3] Finished benchmark
## INFO  [11:37:44.810] [bbotk] Result of batch 10:
## INFO  [11:37:44.811] [bbotk]  regr.glmnet.lambda  regr.mse runtime_learners
## INFO  [11:37:44.811] [bbotk]           0.8894444 0.2254233             0.25
## INFO  [11:37:44.811] [bbotk]                                      uhash
## INFO  [11:37:44.811] [bbotk]  9d38eeaa-45f7-43a3-8926-a369c06a5d4c
## INFO  [11:37:44.815] [bbotk] Finished optimizing after 10 evaluation(s)
## INFO  [11:37:44.815] [bbotk] Result:
## INFO  [11:37:44.816] [bbotk]  regr.glmnet.lambda learner_param_vals  x_domain  regr.mse
## INFO  [11:37:44.816] [bbotk]              0.001          <list[3]> <list[1]> 0.2157293
```

```
at_ridge$predict(task, row_ids = test_set)$score()
```

```
##  regr.mse
## 0.1985571
```

Family size and survival rate

```
#create new column of family size
df5<-read.csv("titanic.csv")[, c("Survived","Pclass","SibSp","Parch")]
df5$Family_class <- df4$SibSp + df4$Parch+1
```

```
task <- TaskRegr$new('titanic5', backend=df5, target = 'Survived')
```

```
#Ridge Regression
learner_ridge <- lrn('regr.glmnet')
learner_ridge$param_set$values <- list(alpha = 0, lambda = 0.01)
gr_ridge <- po('scale') %>>%
  po('imputemean') %>>%
  po(learner_ridge)
glrn_ridge<- GraphLearner$new(gr_ridge)
glrn_ridge$train(task, row_ids = train_set)
glrn_ridge$predict(task, row_ids = test_set)$score()
```

```
##  regr.mse
## 0.1987019
```

```
#Lasso Regression
learner_ridge <- lrn('regr.glmnet')
learner_ridge$param_set$values <- list(alpha = 1, lambda = 0.01)
gr_ridge <- po('scale') %>>%
  po('imputemean') %>>%
  po(learner_ridge)
glrn_ridge<- GraphLearner$new(gr_ridge)
glrn_ridge$train(task, row_ids = train_set)
glrn_ridge$predict(task, row_ids = test_set)$score()
```

```
##  regr.mse
## 0.2004331
```

Random Forests

```
learner_rf <- lrn('regr.ranger')
learner_rf$param_set$values <- list(min.node.size = 4)
gr_rf <- po('scale') %>>%
  po('imputemean') %>>%
  po(learner_rf)
glrn_rf <- GraphLearner$new(gr_rf)
tune_ntrees <- ParamSet$new (list(
 ParamInt$new('regr.ranger.num.trees', lower = 50, upper = 600)
))
at_rf <- AutoTuner$new(
  learner = glrn_rf,
  resampling = rsmp('cv', folds = 3),
  measure = measure,
  search_space = tune_ntrees,
  terminator = terminator,
  tuner = tuner
)
at_rf$train(task, row_ids = train_set)
```

```
at_rf$predict(task, row_ids = test_set)$score()
```

```
##  regr.mse
## 0.1969017
```

K-nearest neighbor

```
df6<-read.csv("titanic.csv")[, c("Survived","Pclass","Sex","Age","Fare")]
df6$Sex <- as.numeric(as.character(df6$Sex)) # converts rate into numerical
```

```
## Warning: NAs introduced by coercion
```

```
head(df6)
```

```
##   Survived Pclass Sex Age    Fare
## 1        0      3  NA  22  7.2500
```

```
## 2        1      1  NA  38 71.2833
## 3        1      3  NA  26  7.9250
## 4        1      1  NA  35 53.1000
## 5        0      3  NA  35  8.0500
## 6        0      3  NA  NA  8.4583
```

```
library(kknn)

task <- TaskRegr$new('titanic6', backend=df6, target = 'Survived')
measure <- msr('regr.mse')

mlr_learners$get("regr.kknn")
```

```
## <LearnerRegrKKNN:regr.kknn>
## * Model: -
## * Parameters: k=7
## * Packages: mlr3, mlr3learners, kknn
## * Predict Type: response
## * Feature types: logical, integer, numeric, factor, ordered
## * Properties: -
```

```
learner_kknn = LearnerRegrKKNN$new()

gr_kknn <- po('imputemean') %>>%
  po(learner_kknn)
glrn_kknn <- GraphLearner$new(gr_kknn)

set.seed(1)
train_set <- sample(task$nrow, 0.7 * task$nrow)
test_set <- setdiff(seq_len(task$nrow), train_set)
glrn_kknn$train(task, row_ids = train_set)
glrn_kknn$predict(task, row_ids = test_set)$score()
```

```
##  regr.mse
## 0.1991261
```

Rpart

```
df6<-read.csv("titanic.csv")[, c("Survived","Pclass","Sex","Age","Fare")]
df6$Sex[df$Sex == 'male'] <- 0
df6$Sex[df$Sex == 'female'] <- 1
df6$Sex <- as.integer(df6$Sex)
head(df6)
```

```
##   Survived Pclass Sex Age    Fare
## 1        0      3   0  22  7.2500
## 2        1      1   1  38 71.2833
## 3        1      3   1  26  7.9250
## 4        1      1   1  35 53.1000
## 5        0      3   0  35  8.0500
## 6        0      3   0  NA  8.4583
```

Xgboost

```
library("xgboost")

task <- TaskRegr$new('titanic6', backend=df6, target = 'Survived')
measure <- msr('regr.mse')

mlr_learners$get("regr.xgboost")
```

```
## <LearnerRegrXgboost:regr.xgboost>
## * Model: -
## * Parameters: nrounds=1, nthread=1, verbose=0
## * Packages: mlr3, mlr3learners, xgboost
## * Predict Type: response
## * Feature types: logical, integer, numeric
## * Properties: hotstart_forward, importance, missings, weights
```

```
learner_xgboost = mlr3::lrn("regr.rpart")

gr_xgboost <- po('imputemean') %>>%
  po(learner_xgboost)
glrn_xgboost <- GraphLearner$new(gr_xgboost)

set.seed(1)
train_set <- sample(task$nrow, 0.7 * task$nrow)
test_set <- setdiff(seq_len(task$nrow), train_set)
glrn_xgboost$train(task, row_ids = train_set)
glrn_xgboost$predict(task, row_ids = test_set)$score()
```

```
##  regr.mse
## 0.1410698
```

Benchmark

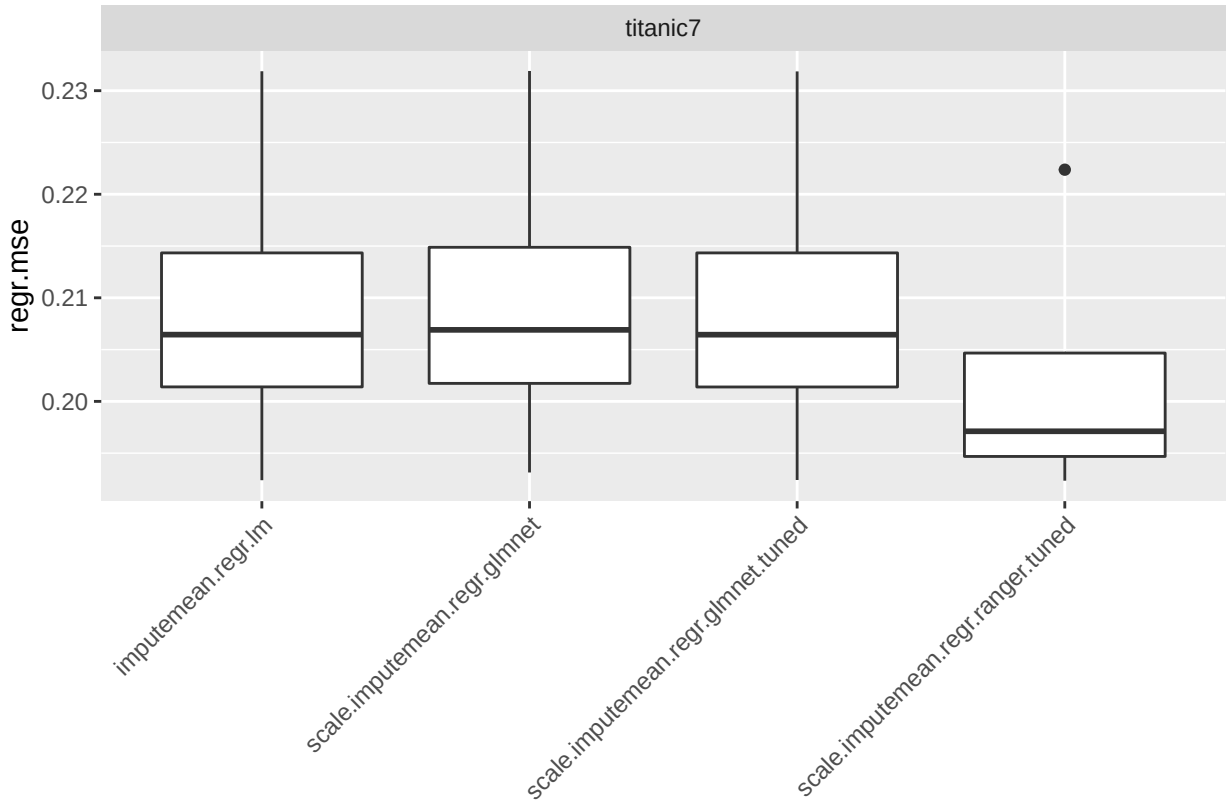
```
task <- TaskRegr$new('titanic7', backend=df4, target = 'Survived')
```

```
set.seed(100)
lrn_list <- list(
  glrn_lm,
  glrn_ridge,
  at_ridge,
  at_rf
)

bm_design <- benchmark_grid(task = task, resamplings = rsmp('cv', folds = 4), learners = lrn_list)
bmr <- benchmark(bm_design, store_models = TRUE)
```

plot

```
library('mlr3viz')
library('ggplot2')
autoplot(bmr) + theme(axis.text.x = element_text(angle = 45, hjust = 1))
```


titanic7
regr.mse
0.23
0.22
0.21
0.20
imputemean.regr.lm
scale.imputemean.regr.glmnet
scale.imputemean.regr.glmnet.tuned
scale.imputemean.regr.ranger.tuned