

OPENCOURSEWARE
APRENDIZAJE AUTOMÁTICO PARA EL ANÁLISIS DE DATOS
GRADO EN ESTADÍSTICA Y EMPRESA
Ricardo Aler



AjusteHiperRF

Ricardo Aler

1/8/2019

Ajuste hiper-parámetros Random Forests

```
library(mlr)
```

```
## Warning: package 'mlr' was built under R version 3.5.3
```

```
## Loading required package: ParamHelpers
```

```
## Warning: package 'ParamHelpers' was built under R version 3.5.3
```

```
library(mlbench)
```

```
library(mlrHyperopt) # Esta librería ayuda en la obtención de hiper-parámetros
```

```
library(ggplot2) # Librería para hacer plots
```

```
library(mlrMBO) # Para hacer ajuste de hiper-parámetros con model-based optimization
```

```
## Loading required package: smooof
```

```
## Loading required package: BBmisc
```

```
##
```

```
## Attaching package: 'BBmisc'
```

```
## The following object is masked from 'package:base':
```

```
##
```

```
## isFALSE
```

```
## Loading required package: checkmate
```

```
# Esto es para que no haya demasiados mensajes de información en el documento
```

```
configureMlr(show.info = FALSE, on.learner.warning = "quiet")
```

```
data(BostonHousing)
```

```
# Primero, definimos una tarea de regresión, cuyos datos están contenidos en el data.frame BostonHousing
```

```
task_bh <- makeRegrTask(data= BostonHousing, target="medv")
```

```
# Segundo, definimos el nombre del algoritmo de aprendizaje ("learner"). Hay muchos de random forests, pro
```

```
learner_name <- "regr.ranger"
```

```
# Tercero, vemos que hiper-parámetros ajustables tiene
```

```
filterParams(getParamSet(learner_name), tunable = TRUE)
```

```
##           Type len      Def
## num.trees integer -       500
## mtry       integer -         -
## min.node.size integer -         5
## replace    logical -       TRUE
## sample.fraction numeric -         -
## split.select.weights numericvector <NA>         -
## always.split.variables untyped -         -
## respect.unordered.factors discrete -   ignore
## splitrule   discrete - variance
```

```
## num.random.splits      integer -      1
## alpha                  numeric -      0.5
## minprop                numeric -      0.1
##                        Constr Req Tunable Trafo
## num.trees              1 to Inf -    TRUE -
## mtry                   1 to Inf -    TRUE -
## min.node.size          1 to Inf -    TRUE -
## replace                - -      TRUE -
## sample.fraction        0 to 1 -    TRUE -
## split.select.weights   0 to 1 -    TRUE -
## always.split.variables - -      TRUE -
## respect.unordered.factors ignore,order,partition - TRUE -
## splitrule               variance,extratrees,maxstat - TRUE -
## num.random.splits      1 to Inf Y    TRUE -
## alpha                  0 to 1 Y    TRUE -
## minprop                0 to 0.5 Y    TRUE -
```

Nuestro learner inicial va a ser una SVM con kernel gaussiano (radial). Lo definimos así:

```
learner_rf <- makeLearner(learner_name)
```

Vemos que tiene muchos hiper-parámetros, pero sabemos que los más importantes son el número de árboles **num.trees** y el **mtry**. También **min.nodesize**, ue controla la profundidad, aunque de manera indirecta. Más información aquí [<https://www.rdocumentation.org/packages/ranger/versions/0.10.1>]. Más adelante tomaremos una decisión acerca de cuales ajustar.

Es interesante saber que podemos conseguir ayuda sobre el método y de sus parámetros así:

```
helpLearner(learner_name)
helpLearnerParam(learner_name, "num.random.splits")
```

*# Ahora habría que definir un espacio de búsqueda para los hiper-parámetros
Una opción es ir a la página web # <http://mlrhyperopt.jakob-r.de/parconfigs>*

Una segunda opción es descargar directamente el espacio de búsqueda desde la web así:

```
pc = downloadParConfigs(learner.name=learner_name) # http://mlrhyperopt.jakob-r.de/parconfigs
print(pc)
```

```
## list()
if(length(pc)>0) {
  ps = getParConfigParSet(pc[[1]], task=task_bh)
  print(ps)
}
```

*# Desgraciadamente, en este caso no hay ninguno exactamente para ese learner
Tendríamos que ir a la página web <http://mlrhyperopt.jakob-r.de/parconfigs>
y buscarlo a mano, para randomForest.*

En este caso, usaremos una tercera opción, más sencilla: la función **generateParConfig** nos da un espacio de búsqueda básico para un learner concreto.

```
pc_rf <- generateParConfig(learner=learner_name, task=task_bh)
ps_rf <- getParConfigParSet(pc_rf, task=task_bh)
print(ps_rf)
```

```
##                               Type len Def  Constr Req Tunable Trafo
```

```
## mtry          integer - 4 1 to 13 - TRUE -
## min.node.size integer - 5 1 to 10 - TRUE -
```

Vemos que aparecen sólo 2 hiper-parámetro. La razón de que no aparezca **num.trees** es que su valor por defecto es de 500 y el autor ha considerado que son más que suficientes. También hay que recordar RandomForest no entra en sobreaprendizaje por más árboles que introduzcamos en el ensemble (a diferencia de Boosting).

Ahora definimos cómo se busca en el espacio de hiper-parámetros (con **randomsearch** en este caso y un **budget** de 50 evaluaciones), y cómo se evalúan los distintos hiper-parámetros (validación cruzada de tres folds).

```
# Ojo, he puesto 1000 evaluaciones para hacer una buena visualización después, pero puede tardar mucho.
# Posiblemente con budget = 50 sea suficiente

control_grid <- makeTuneControlRandom(budget=1000)
inner_desc <- makeResampleDesc("CV", iter=3)

# Aquí construimos una secuencia de ajuste seguida de construcción de modelo
learner_tune_rf <- makeTuneWrapper(learner_rf, resampling = inner_desc, par.set = ps_rf, control = control_grid)

# También tenemos que definir cómo se va a evaluar el modelo final. En este caso con train/test (holdout)
outer_desc <- makeResampleDesc("Holdout")
set.seed(0)
outer_inst <- makeResampleInstance(outer_desc, task_bh)

# Por último, usamos resample para entrenar y evaluar learner_tune_rf
set.seed(0)
error_tune_rf <- resample(learner_tune_rf, task_bh, outer_inst, measures = list(rmse), extract = getTuneResults)
```

Vemos que los hiper-parámetros que se eligieron en la partición de entrenamiento son los siguientes (también se muestra el error que producen dichos hiper-parámetros en la validación cruzada de 3 folds que se usó para seleccionarlos).

```
error_tune_rf$extract
```

```
## [[1]]
## Tune result:
## Op. pars: mtry=6; min.node.size=1
## rmse.test.rmse=3.6497790
```

y el error del modelo final es:

```
error_tune_rf$aggr
```

```
## rmse.test.rmse
## 3.182439
```

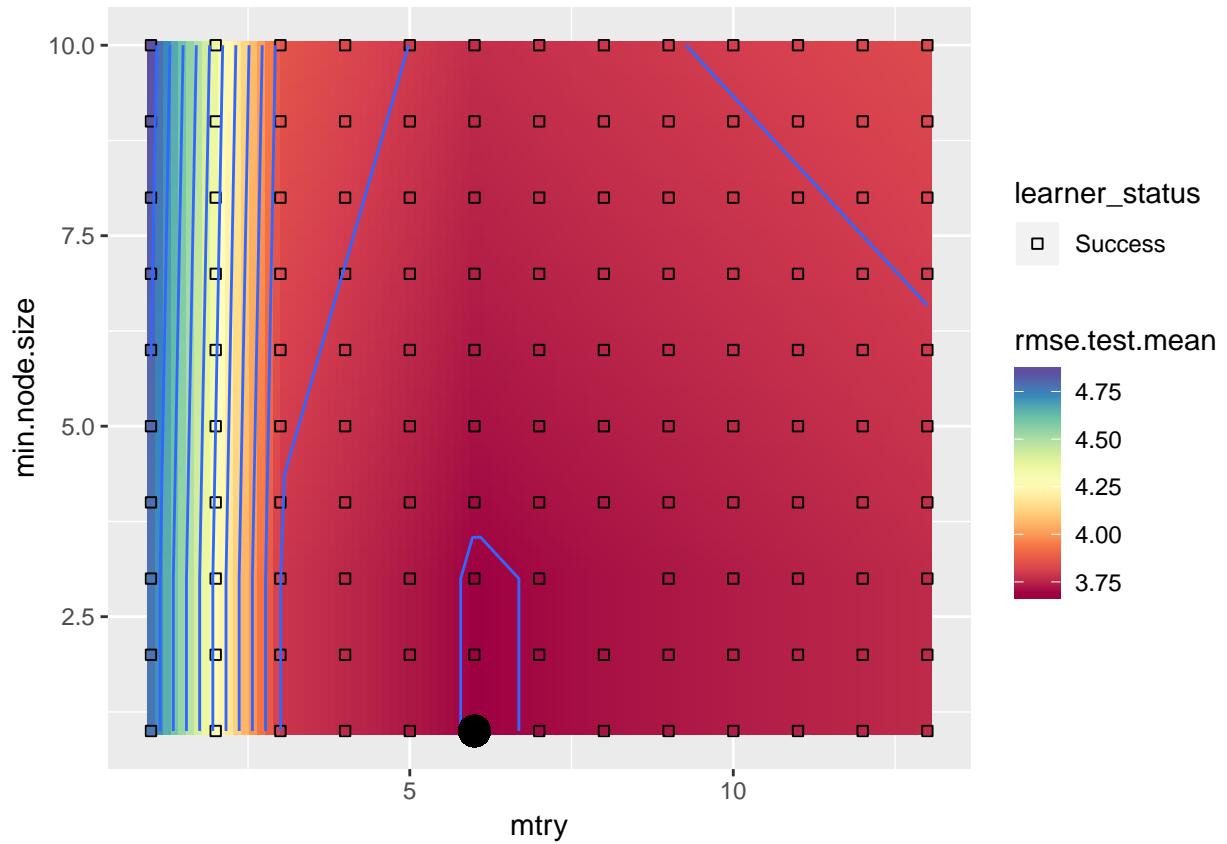
```
data <- generateHyperParsEffectData(error_tune_rf, include.diagnostics = FALSE, trafo=TRUE)
```

```
# Las siguientes dos líneas son para solventar un bug de MLR
names(data$data)[names(data$data)=="rmse.test.rmse"] <- "rmse.test.mean"
data$measures[data$measures=="rmse.test.rmse"] <- "rmse.test.mean"
```

Aquí vemos el espacio de búsqueda que hemos explorado:

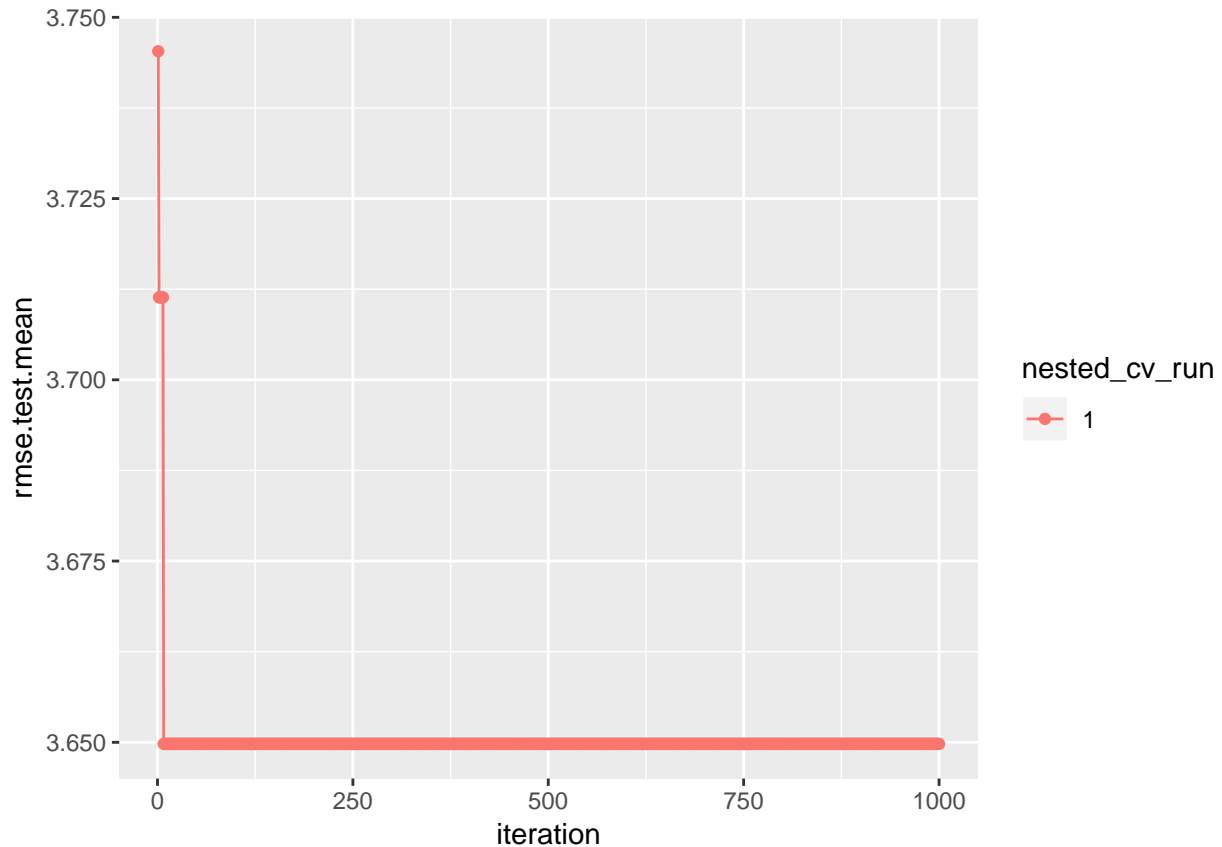
```
plt = plotHyperParsEffect(data, x = "mtry", y = "min.node.size", z = "rmse.test.mean",
  plot.type = "contour", interpolate = "regr.earth", show.experiments = TRUE)
```

```
plot(plt + geom_point(x=error_tune_rf$extract[[1]]$x$mtry, y=error_tune_rf$extract[[1]]$x$min.node.size
```



A continuación podemos ver como se ha ido reduciendo el error con las iteraciones.

```
plt = plotHyperParsEffect(data, x = "iteration", y = "rmse.test.mean", plot.type = "line")  
plt
```



Vamos a hacer el ajuste de hiper-parámetros con Model Based Optimization, del paquete mbo

```
# Dedicar 80 iteraciones al ajuste
control = makeMBOControl()
control = setMBOControlTermination(control, iters = 80)
control = setMBOControlInfill(control, crit = makeMBOInfillCritEI())

control_grid <- makeTuneControlMBO(mbo.control = control)

inner_desc <- makeResampleDesc("CV", iter=3)

# Aquí construimos una secuencia de ajuste seguida de construcción de modelo
learner_tune_rf <- makeTuneWrapper(learner_rf, resampling = inner_desc, par.set = ps_rf, control = control)

# También tenemos que definir cómo se va a evaluar el modelo final. En este caso con train/test (holdout)
outer_desc <- makeResampleDesc("Holdout")
set.seed(0)
outer_inst <- makeResampleInstance(outer_desc, task_bh)

# Por último, usamos resample para entrenar y evaluar learner_tune_rf
set.seed(0)
error_tune_rf <- resample(learner_tune_rf, task_bh, outer_inst, measures = list(rmse), extract = getTuneResults)
```



```
## Warning in generateDesign(control$infill.opt.focusearch.points,
## ps.local, : generateDesign could only produce 130 points instead of 1000!

## Warning in generateDesign(control$infill.opt.focusearch.points,
## ps.local, : generateDesign could only produce 28 points instead of 1000!

## Warning in generateDesign(control$infill.opt.focusearch.points,
## ps.local, : generateDesign could only produce 10 points instead of 1000!

## Warning in generateDesign(control$infill.opt.focusearch.points,
## ps.local, : generateDesign could only produce 6 points instead of 1000!

## Warning in generateDesign(control$infill.opt.focusearch.points,
## ps.local, : generateDesign could only produce 6 points instead of 1000!
```

Vemos que los hiper-parámetros que se eligieron en la partición de entrenamiento son los siguientes (también se muestra el error que producen dichos hiper-parámetros en la validación cruzada de 3 folds que se usó para seleccionarlos).

```
error_tune_rf$extract
```

```
## [[1]]
## Tune result:
## Op. pars: mtry=8; min.node.size=2
## rmse.test.rmse=3.6510000
```

y el error del modelo final es muy parecido también al que obtuvimos con Random Search:

```
error_tune_rf$aggr
```

```
## rmse.test.rmse
##      3.244562
```

Lo que si parece claro que este tipo de búsqueda se centra mucho mejor en los valores adecuados de n.trees y Por último, podemos ver un plot que nos muestra cómo cambia el error con los distintos valores de los hiper-parámetros.

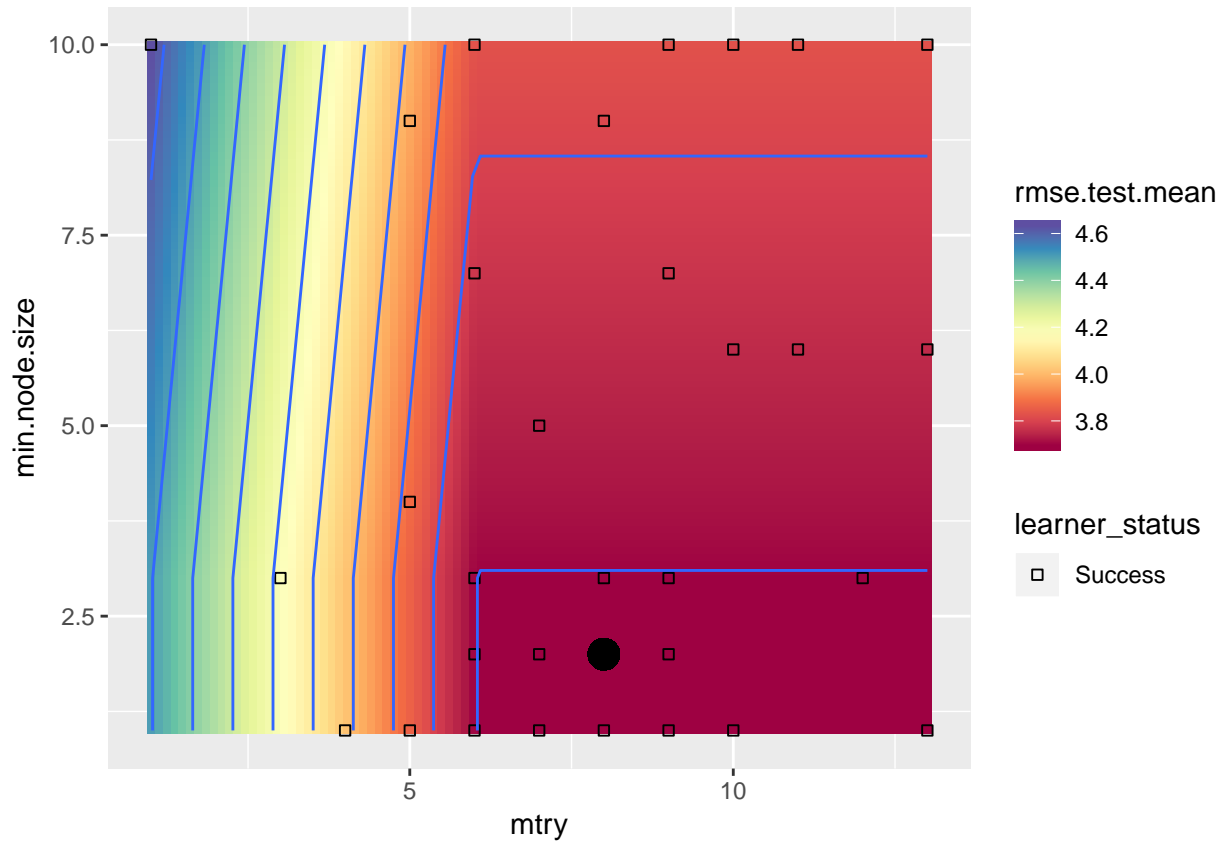
```
data <- generateHyperParsEffectData(error_tune_rf, include.diagnostics = FALSE)
```

```
# Las siguientes dos líneas son para solventar un bug de MLR
names(data$data)[names(data$data)=="rmse.test.rmse"] <- "rmse.test.mean"
data$measures[data$measures=="rmse.test.rmse"] <- "rmse.test.mean"
```

Aquí vemos el espacio de búsqueda que hemos explorado:

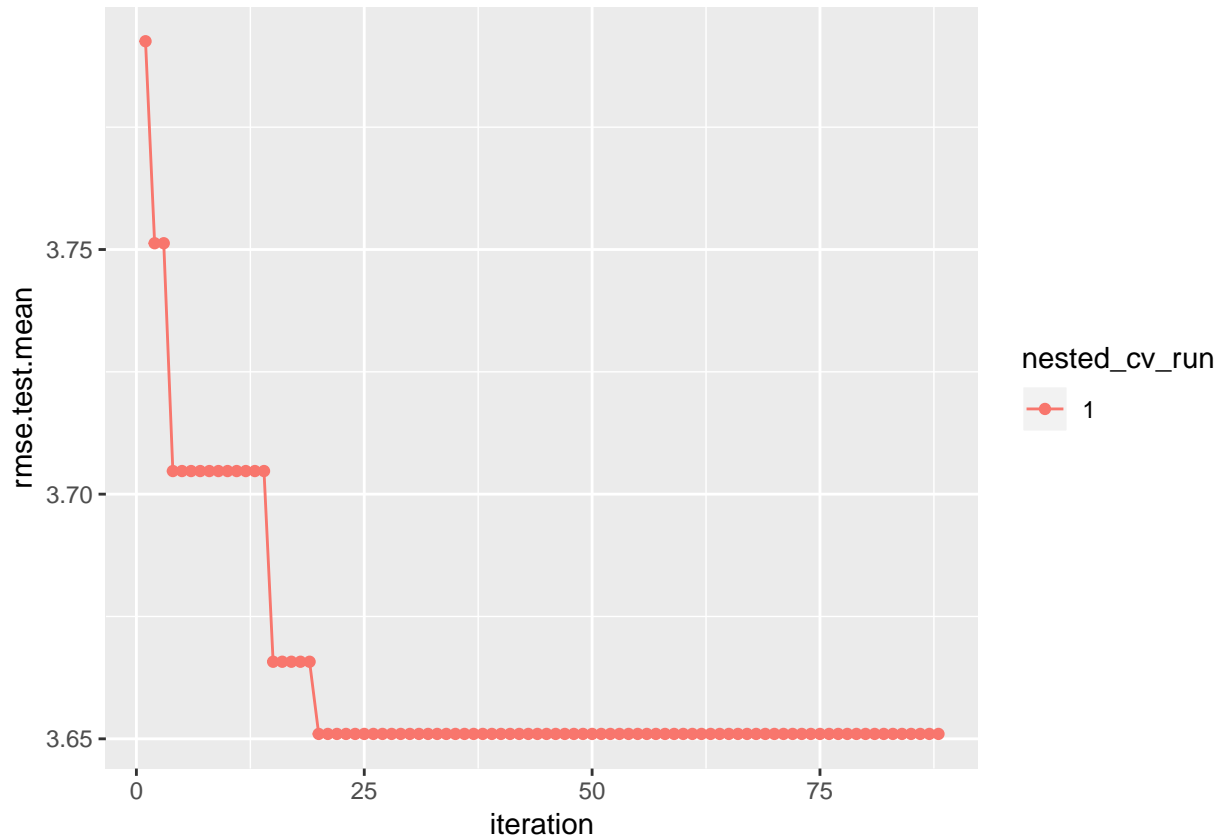
```
plt = plotHyperParsEffect(data, x = "mtry", y = "min.node.size", z = "rmse.test.mean",
  plot.type = "contour", interpolate = "regr.earth", show.experiments = TRUE)
```

```
plot(plt + geom_point(x=error_tune_rf$extract[[1]]$x$mtry, y=error_tune_rf$extract[[1]]$x$min.node.size
```



A continuación podemos ver como se ha ido reduciendo el error con las iteraciones. Parece que con pocas iteraciones se alcanzan buenos mínimos.

```
plt = plotHyperParsEffect(data, x = "iteration", y = "rmse.test.mean",
  plot.type = "line")
plt
```



Por último, puede ser interesante saber qué hubiera pasado de haber utilizado un número de árboles diferente a 500. Definimos un learner con los mejores hiper-parámetros encontrados hasta el momento y un nuevo espacio de búsqueda donde sólo cambie `num.trees`. Ahora utilizaremos la búsqueda sistemática de grid-search para hacerlo.

```

learner_rf <- setHyperPars(makeLearner(learner_name), par.vals = error_tune_rf$extract[[1]]$x)

ps_rf <- makeParamSet(
  makeDiscreteParam("num.trees", values = c(10, seq(50, 2000, by=100)))
)

control_grid <- makeTuneControlGrid()
inner_desc <- makeResampleDesc("CV", iter=3)

# Aquí construimos una secuencia de ajuste seguida de construcción de modelo
# Nótese que inner_Desc, control_grid, outer_inst, etc no cambián, con lo que usamos los mismos que ant
learner_tune_rf <- makeTuneWrapper(learner_rf, resampling = inner_desc, par.set = ps_rf, control = control_grid)

# Por último, usamos resample para entrenar y evaluar learner_tune_rf
set.seed(0)
error_tune_rf <- resample(learner_tune_rf, task_bh, outer_inst, measures = list(rmse), extract = getTuneResults)

data <- generateHyperParsEffectData(error_tune_rf, include.diagnostics = FALSE)

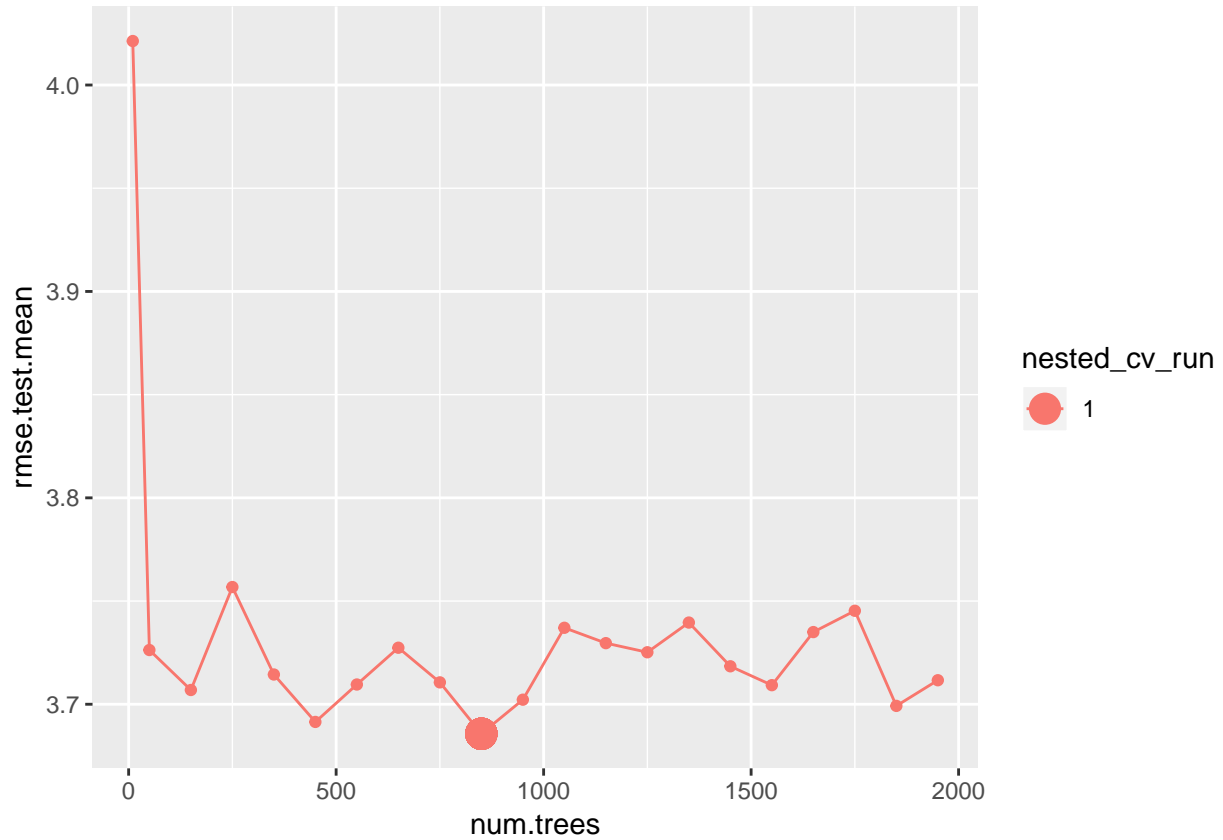
# Las siguientes dos líneas son para solventar un bug de MLR
names(data$data)[names(data$data)=="rmse.test.rmse"] <- "rmse.test.mean"
data$measures[data$measures=="rmse.test.rmse"] <- "rmse.test.mean"

```

Aquí vemos el espacio de búsqueda que hemos explorado. Aunque la dependencia del error respecto al número de árboles es algo ruidosa, parece que 500 árboles era un valor razonable, aunque tal vez nos hubieramos podido beneficiar con algo mas de 500 árboles.

```
plt = plotHyperParsEffect(data, x = "num.trees", y = "rmse.test.mean",  
  plot.type = "line", show.experiments = TRUE)
```

```
plot(plt + geom_point(x=error_tune_rf$extract[[1]]$x$num.trees, y=error_tune_rf$extract[[1]]$y, size=5))
```



```
plt = plotHyperParsEffect(data, x = "iteration", y = "rmse.test.mean",  
  plot.type = "line")  
plt
```

