

# fig-train-vs-test-error.R

carol

2021-06-17

```
# PREREQ -----  
  
library(mlr3)  
library(mlr3learners)  
library(ggplot2)  
  
## Warning: package 'ggplot2' was built under R version 4.0.5  
  
# DATA -----  
  
set.seed(123L)  
  
#n_reps = 2L #50 !!  
ss_iters = 50L  
  
#n = 5000L  
#n_2 = 50L  
#data = data.table::as.data.table(mlbench::mlbench.spirals(n, sd = 1))  
  
# EXPERIMENT -----  
task = tsk("boston_housing")  
n = task$nrow  
#mlr3::TaskClassif$new("spirals", backend = data, target = "classes")  
learner = lrn("regr.kknn", k = 6)  
  
# Increasing training set size  
  
test_size = 206L  
train_sizes = as.integer(c(seq(7, n-test_size, length.out= 8)))  
#task_sizes = test_size + train_sizes  
  
length = 2*length(train_sizes)*ss_iters  
results_train = data.frame(error = numeric(length),  
                           type = character(length),  
                           train_size = numeric(length),  
                           rep = numeric(length),  
                           stringsAsFactors = FALSE)  
  
for (j in seq_along(train_sizes)) {  
  
  #task_subset = task$clone()$filter(sample(n, task_sizes[[j]]))
```

```

for(k in seq_len(ss_iters)){

  train_set = sample(task$nrow, train_sizes[[j]])
  test_set = setdiff(seq_len(task$nrow), train_set)

  learner$train(task, row_ids = train_set)

  #calculate training error
  pred_train = learner$predict(task, row_ids = train_set)
  train_err = pred_train$score() #classif.ce

  #calculate test error
  pred_test = learner$predict(task, row_ids = test_set)
  test_err = pred_test$score()


  #fill data frame
  index = (j-1)*ss_iters+k
  results_train$train_size[index] = train_sizes[[j]]
  results_train$error[index] = train_err
  results_train$type[index] = "train error"
  results_train$rep[index] = k

  results_train$train_size[length/2+index] = train_sizes[[j]]
  results_train$error[length/2+index] = test_err
  results_train$type[length/2+index] = "test error"
  results_train$rep[length/2+index] = k
}
}

```

```

## Warning: package 'kkn' was built under R version 4.0.5

## Warning: package 'kkn' was built under R version 4.0.5

## Warning: package 'kkn' was built under R version 4.0.5

## Warning: package 'kkn' was built under R version 4.0.5

## Warning: package 'kkn' was built under R version 4.0.5

## Warning: package 'kkn' was built under R version 4.0.5

## Warning: package 'kkn' was built under R version 4.0.5

## Warning: package 'kkn' was built under R version 4.0.5

## Warning: package 'kkn' was built under R version 4.0.5

## Warning: package 'kkn' was built under R version 4.0.5

## Warning: package 'kkn' was built under R version 4.0.5

## Warning: package 'kkn' was built under R version 4.0.5

```









[illegible]









[illegible]







[illegible]

[illegible]



[illegible]





[illegible]











[illegible]





[illegible]

[illegible]

[illegible]

[illegible]

```
## Warning: package 'kkn' was built under R version 4.0.5
```

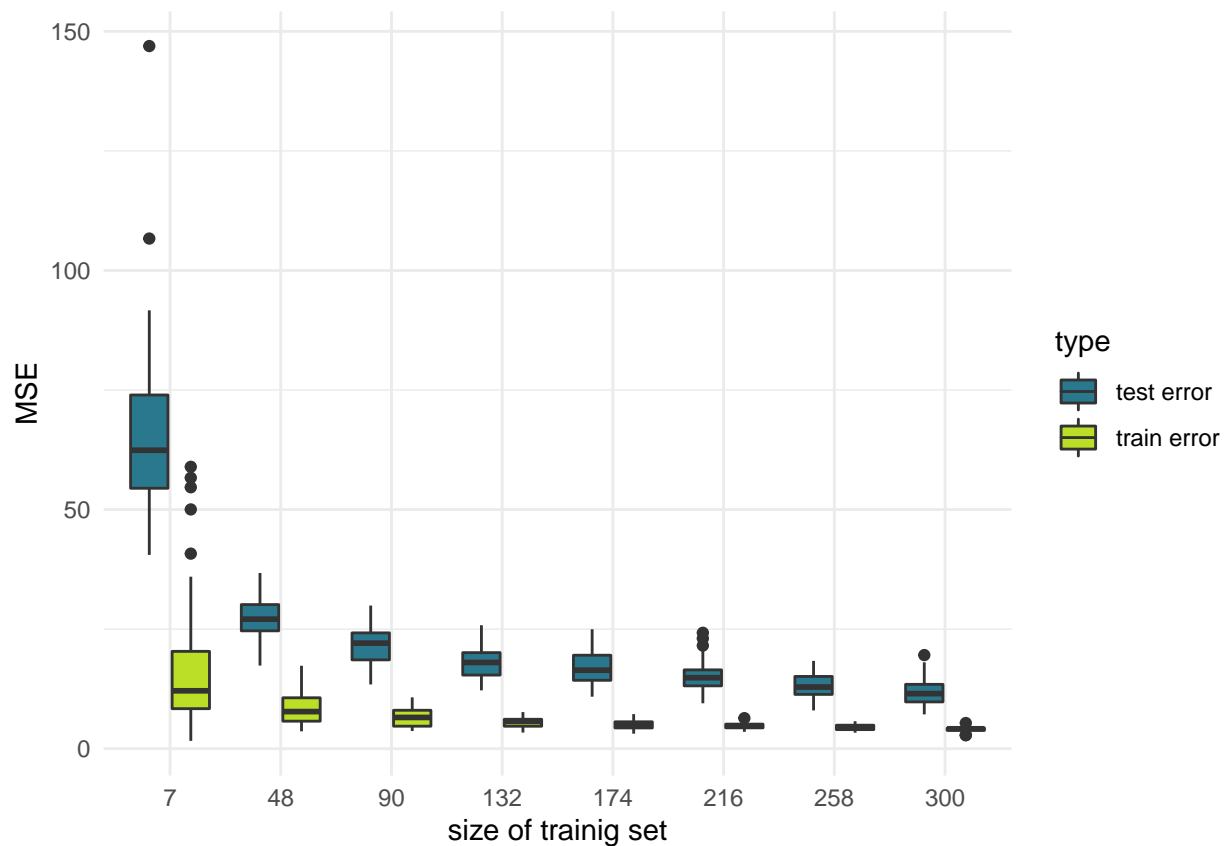
```
## Warning: package 'kkn' was built under R version 4.0.5
```

```
## Warning: package 'kkn' was built under R version 4.0.5
```

```
## Warning: package 'kkn' was built under R version 4.0.5
```

```
## Warning: package 'kkn' was built under R version 4.0.5
```

```
p1 <- ggplot(data= results_train, aes(x = factor(train_size), y= error, fill= type)) +  
  geom_boxplot() +  
  theme_minimal() +  
  scale_fill_viridis_d(begin= 0.4, end = 0.9) +  
  xlab("size of trainig set") +  
  ylab("MSE")  
p1
```



```
ggsave("../figure/fig-train-vs-test-error-1.pdf", p1, width = 8, height = 3.5)
```

```
# Increasing test set size
```

```
train_size = 50
```

```
tests_sizes = as.integer(c(seq(10L, n=train_size, length.out = 8)))
```

```
#task_sizes = tests_sizes + train_size
```



```

length = 2*length(tests_sizes)*ss_iters
results_test = data.frame(error = numeric(length),
                          type = character(length),
                          test_size = numeric(length),
                          rep = numeric(length),
                          stringsAsFactors = FALSE)

for (j in seq_along(tests_sizes)) {

  #task_subset = task$clone()$filter(sample(n, task_sizes[[j]]))

  for(k in seq_len(ss_iters)){

    train_set = sample(task$nrow, train_size)
    test_set = setdiff(seq_len(task$nrow), train_set)

    learner$train(task, row_ids = train_set)

    #calculate training error
    pred_train = learner$predict(task, row_ids = train_set)
    train_err = pred_train$score()

    #calculate test error
    pred_test = learner$predict(task, row_ids = test_set)
    test_err = pred_test$score()

    #fill data frame
    index = (j-1)*ss_iters+k
    results_test$test_size[index] = tests_sizes[[j]]
    results_test$error[index] = train_err
    results_test$type[index] = "train error"
    results_test$rep[index] = k

    results_test$test_size[length/2+index] = tests_sizes[[j]]
    results_test$error[length/2+index] = test_err
    results_test$type[length/2+index] = "test error"
    results_test$rep[length/2+index] = k
  }
}

```

```
## Warning: package 'kkn' was built under R version 4.0.5
```

```
## Warning: package 'kkn' was built under R version 4.0.5
```

```
## Warning: package 'kkn' was built under R version 4.0.5
```

```
## Warning: package 'kkn' was built under R version 4.0.5
```

```
## Warning: package 'kkn' was built under R version 4.0.5
```











[illegible]





























[illegible]

















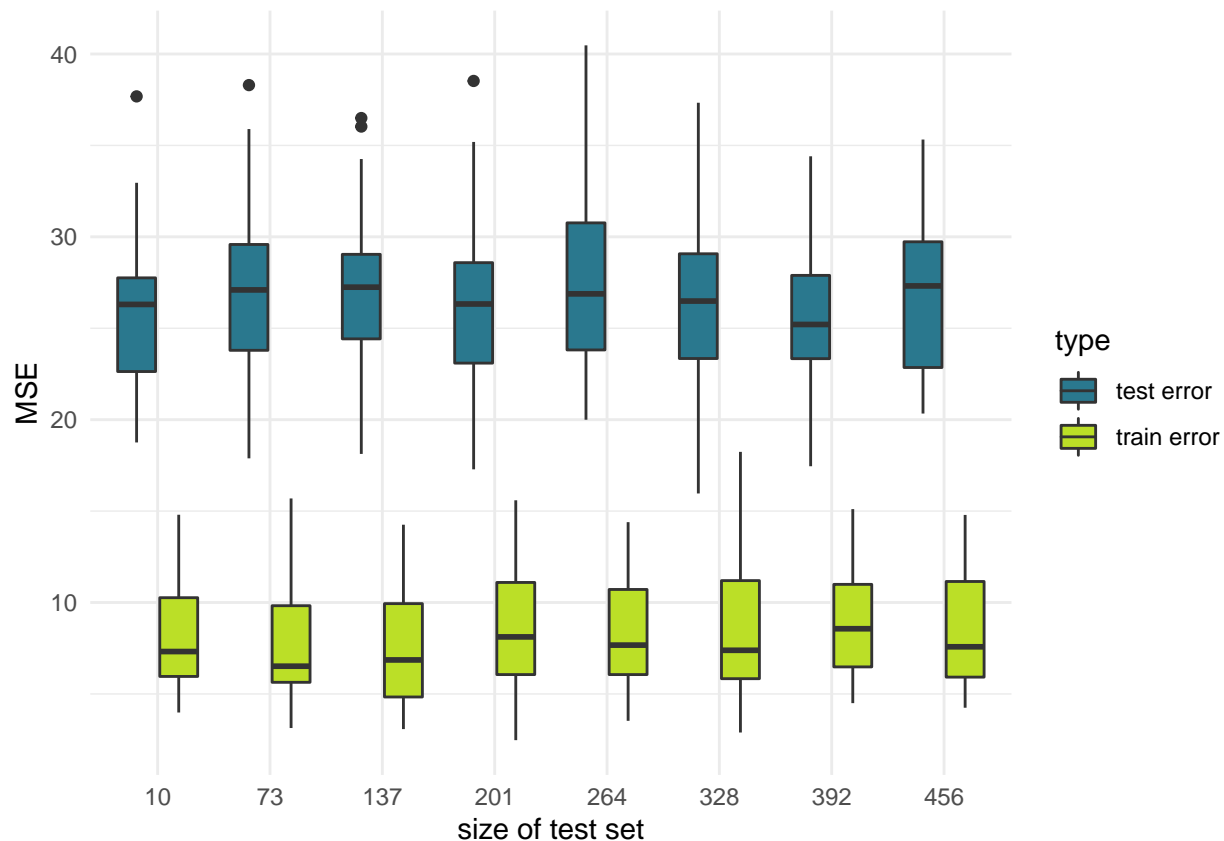






```
## Warning: package 'kknn' was built under R version 4.0.5
## Warning: package 'kknn' was built under R version 4.0.5
## Warning: package 'kknn' was built under R version 4.0.5
## Warning: package 'kknn' was built under R version 4.0.5
## Warning: package 'kknn' was built under R version 4.0.5
## Warning: package 'kknn' was built under R version 4.0.5
## Warning: package 'kknn' was built under R version 4.0.5
## Warning: package 'kknn' was built under R version 4.0.5
## Warning: package 'kknn' was built under R version 4.0.5
## Warning: package 'kknn' was built under R version 4.0.5
## Warning: package 'kknn' was built under R version 4.0.5
## Warning: package 'kknn' was built under R version 4.0.5
## Warning: package 'kknn' was built under R version 4.0.5
```

```
p2 <- ggplot(data= results_test, aes(x = factor(test_size), y= error, fill= type)) +
  geom_boxplot() +
  theme_minimal() +
  scale_fill_viridis_d(begin= 0.4, end = 0.9) +
  xlab("size of test set") +
  ylab("MSE")
p2
```



```
ggsave("../figure/fig-train-vs-test-error-2.pdf", p2, width = 8, height = 3.5)
```

```
# Variation of model complexity
```

```
complexity_k = c(1, 5, 10, 50, 100, 150)
```

```
learners_complexity = lapply(
  complexity_k,
  function(i) mlr3::lrn("regr.kknn", k = i))
```

```
train_size = as.integer(0.7*n)
```

```
length = 2*length(learners_complexity)*ss_iters
results_complexity = data.frame(error = numeric(length),
                                type = character(length),
                                complexity = factor(length, levels = complexity_k),
                                rep = numeric(length),
                                stringsAsFactors = FALSE)
```

```
for (j in seq_along(learners_complexity)) {
  for(k in seq_len(ss_iters)){
    train_set = sample(task$nrow, train_size)
    test_set = setdiff(seq_len(task$nrow), train_set)
```



```

learners_complexity[[j]]$train(task, row_ids = train_set)

#calculate training error
pred_train = learners_complexity[[j]]$predict(task, row_ids = train_set)
train_err = pred_train$score()

#calculate test error
pred_test = learners_complexity[[j]]$predict(task, row_ids = test_set)
test_err = pred_test$score()

#fill data frame
index = (j-1)*ss_iters+k
results_complexity$complexity[index] = complexity_k[[j]]
results_complexity$error[index] = train_err
results_complexity$type[index] = "train error"
results_complexity$rep[index] = k

results_complexity$complexity[length/2+index] = complexity_k[[j]]
results_complexity$error[length/2+index] = test_err
results_complexity$type[length/2+index] = "test error"
results_complexity$rep[length/2+index] = k
}
}

```

```

## Warning: package 'kkn' was built under R version 4.0.5

## Warning: package 'kkn' was built under R version 4.0.5

## Warning: package 'kkn' was built under R version 4.0.5

## Warning: package 'kkn' was built under R version 4.0.5

## Warning: package 'kkn' was built under R version 4.0.5

## Warning: package 'kkn' was built under R version 4.0.5

## Warning: package 'kkn' was built under R version 4.0.5

## Warning: package 'kkn' was built under R version 4.0.5

## Warning: package 'kkn' was built under R version 4.0.5

## Warning: package 'kkn' was built under R version 4.0.5

## Warning: package 'kkn' was built under R version 4.0.5

## Warning: package 'kkn' was built under R version 4.0.5

## Warning: package 'kkn' was built under R version 4.0.5

```

[illegible]

[illegible]



[illegible]

[illegible]

[illegible]

[illegible]



[illegible]

[illegible]



[illegible]



[illegible]

[illegible]





[illegible]

[illegible]

[illegible]

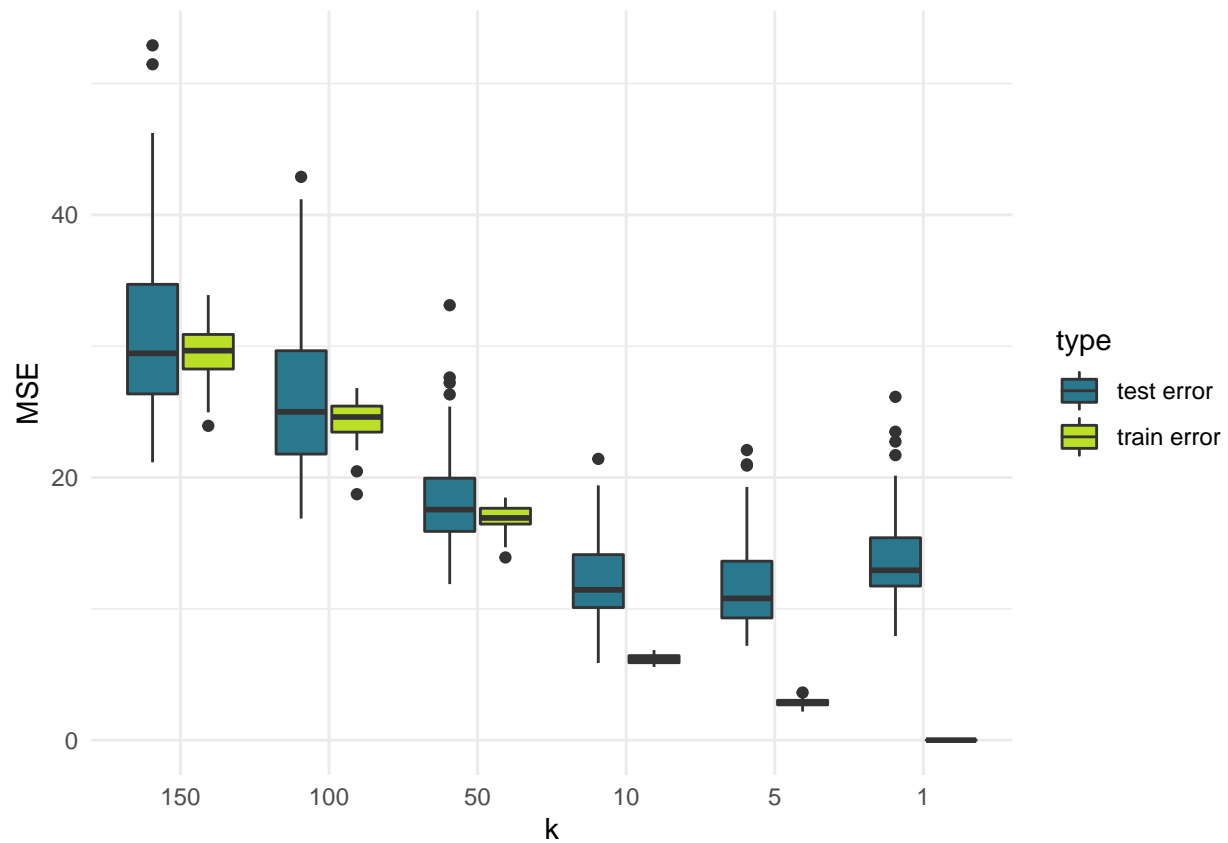
[illegible]

[illegible]

[illegible]

[illegible]

```
p3 <- ggplot(data= results_complexity, aes(x = complexity, y= error, fill= type)) +
  geom_boxplot() +
  theme_minimal() +
  scale_fill_viridis_d(begin= 0.4, end = 0.9) +
  xlab("k") +
  scale_x_discrete(limits = rev(unique(sort(results_complexity$complexity)))) +
  ylab("MSE")
p3
```



```
ggsave("../figure/fig-train-vs-test-error-3.pdf", p3, width = 8, height = 3.5)
```

```
# #-----
# # True Performance
# #-----
# resampling = mlr3::rsmp(
#   "subsampling",
#   ratio = n_2 / n,
#   repeats = ss_iters * 10L)
#
# resampling_result = mlr3::resample(
#   task,
#   learner,
#   resampling,
#   store_models = FALSE)
#
# true_performance = resampling_result$aggregate(mlr3::msr("classif.ce"))
#
#
```