Practical: Efficient Code

BaselRBootcamp 2017

Slides

Here a link to the lecture slides for this session: **LINK** (https://therbootcamp.github.io/_sessions/D4S1_EfficientCode/EfficientCode.html)

Overview

In this practical you'll learn how to write efficient code. By the end of this practical you will know how to:

1. Profile your code to identify critical parts.
2. Make code more efficient.
3. How to do parallel computing.

Benchmarking and profiling functions

Functions to profile your code are:

Function	Package	Description
`proc.time()`	`base`	Returns the time.
`system.time()`	`base`	Runs one expression once and returns elapsed CPU time
`microbenchmark()`	`microbenchmark`	Runs one or many expressions multiple times and returns statistics on elapsed time.
`lineprof(), shine()`	`lineprof`	Evaluates entire scripts. (From Hadley's Github)

Microbenchmark: Example

Small (minimal) chunks of code can conveniently be tested using `microbenchmark()`.

```
# load packages
library(microbenchmark)
library(tibble)

# get data
df <- data.frame('var_1' = rnorm(1000,1),
                 'var_2' = rnorm(1000,1))
tbl <- as.tibble(df)

# microbenchmark pt. 1
microbenchmark(df[['var_1']], df$var_1, tbl$var_1)
```

Profiling: Example

Larger code chunks or even scripts can conveniently be tested using `system.time()` and `lineprof()` from the `lineprof` package.

```
# ---- install and load package
install.packages('devtools')
devtools::install_github("hadley/lineprof")
library(lineprof)
library(readr)
library(dplyr)

# ---- define code chunk as function

my_chunkfun <- function(){

  # load data
  data <- read_csv('http://tinyurl.com/titanic-dataset-web')

  # remove first column
  data <- data[,-1]

  # mutate
  data <- data %>%
    mutate(months = Age * 12)

  # select
  test_data <- data %>%
    select(Sex, Age, Survived)

  # multiple regression
  # Survival predicted by Sex, Age, and their interaction
  model <- glm(Survived ~ Sex * Age,
               data = test_data,
               family = 'binomial')

  # evaluate model
  summary(model)

  }

# ---- profiling

# profile using system.time
system.time(my_chunkfun())

# profile using lineprof
#profile <- lineprof(my_chunkfun())
#shine(profile)
```

Tasks

Microbenchmark

1. Run the microbenchmark example from above. What do you find? Are `tibble` 's fast or slow?

```
# load packages
library(microbenchmark)
library(tibble)

# get data
df <- data.frame('var_1' = rnorm(1000, 1),
                 'var_2' = rnorm(1000, 1))
tbl <- as.tibble(df)

# microbenchmark
microbenchmark(df[['var_1']], df$var_1, tbl$var_1)
```

```
Unit: microseconds
          expr    min     lq     mean  median      uq     max neval cld
 df[["var_1"]]  3.134  4.228  5.74518  4.9390  5.9190  41.958   100  a
      df$var_1  4.072  5.404  6.60436  6.3495  7.1760  14.947   100  a
     tbl$var_1 34.547 38.098 42.09851 39.3750 40.9395 176.698   100   b
```

2. Repeat the comparison of `tibble` s and basic `data.frame` s of the first exercise and include now for both data frame types also the `.subset2()` function (don't forget the dot). The function takes two arguments: The first argument is the data frame, the second argument is the column identifyier (index or name). What do you find?

```
# load packages
library(microbenchmark)
library(tibble)

# get data
df <- data.frame('var_1' = rnorm(1000,1),
                 'var_2' = rnorm(1000,1))
tbl <- as.tibble(df)

# microbenchmark
microbenchmark(df[['var_1']], df$var_1, tbl$var_1,
               .subset(df,'var_1'), .subset(tbl,'var_1'))
```

```
Unit: nanoseconds
                  expr   min      lq     mean  median      uq    max neval
         df[["var_1"]]  3361  4194.0  5043.88  4979.5  5940.5   7789   100
              df$var_1  4193  5021.0  6354.11  5794.0  7090.5  37153   100
             tbl$var_1 34665 38137.0 40737.62 39271.0 41687.0 119972   100
  .subset(df, "var_1")   364   503.0   595.56   586.0   674.5   1162   100
 .subset(tbl, "var_1")   348   484.5   713.75   589.0   698.5  12412   100
 cld
  b
  b
   c
 a
 a
```

3. Compare the the function `mean()` to the operation composed of its basic ingredients `sum()` and `length()` , i.e., `sum(my_vec) / length(my_vec)` . To do this first create a vector consisting of random numbers using `runif()` (see `?runif`). Then test both ways with `microbenchmark()` What do you find?

```
# define vector
my_vec <- runif(10000)

# microbenchmark
microbenchmark(mean(my_vec), sum(my_vec)/length(my_vec))
```

```
Unit: microseconds
                      expr    min      lq     mean median      uq    max
              mean(my_vec) 17.526 24.3470 23.83141 24.418 24.5330 37.777
 sum(my_vec)/length(my_vec)  8.034 11.2105 11.16223 11.288 11.3335 18.872
 neval cld
   100   b
   100  a
```

4. Test the type of each of `mean()`, `sum()`, `length()`, and `.subset2()` using `typeof()`. What's the fast type?

```
# test type
typeof(mean); typeof(sum); typeof(length); typeof(.subset2)
```

```
[1] "closure"
```

```
[1] "builtin"
```

```
[1] "builtin"
```

```
[1] "builtin"
```

Profiling

5. Copy the profiling example into a new script file. After installing and loading the `devtools` and `lineprof` packages, run the code under *'define code chunk as function'* and then test the function by runnning it, i.e., execute `my_chunkfun()`. You just defined and executed your first self-created function. Continue by profiling the function using `system.time()` and `lineprof()`. What do you find? What parts of the code are most computationally expensive? Repeat the analysis. Remember R compiles functions after first use.

```
# ---- install and load package
install.packages('devtools', repos = "https://stat.ethz.ch/CRAN/")
```

```


The downloaded binary packages are in
    /var/folders/4j/gkx0z2kn1b5djq50kwgl2wdc0000gp/T//RtmpLf9kur/downloaded_packages
```

```
devtools::install_github("hadley/lineprof")
library(lineprof)
library(readr)
library(dplyr)

# ---- define code chunk as function

my_chunkfun <- function(){

  # load data
  data <- read_csv('http://tinyurl.com/titanic-dataset-web')

  # remove first column
  data <- data[,-1]

  # mutate
  data <- data %>%
    mutate(months = Age * 12)

  # select
  test_data <- data %>%
    select(Sex, months, Survived)

  # multiple regression
  # Survival predicted by Sex, months, and their interaction
  model <- glm(Survived ~ Sex * months,
               data = test_data,
               family = 'binomial')

  # evaluate model
  summary(model)

  }

# ---- profiling

# profile using system.time
system.time(my_chunkfun())
```

```
   user  system elapsed
  0.119   0.008   1.212
```

```
# profile using lineprof
#profile <- lineprof(my_chunkfun())
#shine(profile)
```

Speeding up code

6. When speeding up code, the first question should always be whether faster solutions are already out there. In this case there are. Check out the `data.table` package (means: install and load) and use the `fread()` function. Try defining a new function using this function rather than `read_csv()`. Then compare the performance of the two. How much faster is the new relative to the old function (use `system.time()`)?

```
# ---- install and load package
install.packages('data.table', repos = "https://stat.ethz.ch/CRAN/")
```

```
The downloaded binary packages are in
    /var/folders/4j/gkx0z2kn1b5djq50kwgl2wdc0000gp/T//RtmpLf9kur/downloaded_packages
```

```
library(data.table)

# ---- define code chunk as function

my_chunkfun_fast <- function(){

  # load data
  data <- fread('http://tinyurl.com/titanic-dataset-web')

  # remove first column
  data <- data[,-1]

  # mutate
  data <- data %>%
    mutate(months = Age * 12)

  # select
  test_data <- data %>%
    select(Sex, months, Survived)

  # multiple regression
  # Survival predicted by Sex, months, and their interaction
  model <- glm(Survived ~ Sex * months,
               data = test_data,
               family = 'binomial')

  # evaluate model
  summary(model)

  }

# ---- profiling

# profile using system.time
system.time(my_chunkfun())
```

```
   user  system elapsed
  0.148   0.006   1.204
```

```
system.time(my_chunkfun_fast())
```

```
   user  system elapsed
  0.027   0.004   1.560
```

```
# profile using lineprof
#profile <- lineprof(my_chunkfun_fast())
#shine(profile)
```

7. The next step of optimising code is to identify bits that are not necessary. Try to identify a bit that is not entirely necessary, remove it, and evaluate the function's performance again.

```
# ---- define code chunnk as function

my_chunkfun_fast2 <- function(){

  # load data
  data <- fread('http://tinyurl.com/titanic-dataset-web')

  # remove first column
  data <- data[,-1]

  # mutate
  data <- data %>%
    mutate(months = Age * 12)

  # multiple regression
  # Survival predicted by Sex, months, and their interaction
  model <- glm(Survived ~ Sex * months,
               data = data,
               family = 'binomial')

  # evaluate model
  summary(model)

  }

# ---- profiling

# profile using lineprof
#profile <- lineprof(my_chunkfun_fast2())
#shine(profile)

# profile using system.time
system.time(my_chunkfun())
```

```
   user  system elapsed
  0.025   0.002   1.153
```

```
system.time(my_chunkfun_fast())
```

```
   user  system elapsed
  0.028   0.005   1.364
```

```
system.time(my_chunkfun_fast2())
```

```
   user  system elapsed
  0.014   0.004   1.296
```

8. Next think about whether vectorization may make sense. Find a code chunk that may be written using a vector and vector multiplication and try to implement it. Is there any improvement?

```
# ---- define code chunk as function

my_chunkfun_fast3 <- function(){

  # load data
  data <- fread('http://tinyurl.com/titanic-dataset-web')

  # remove first column
  data <- data[,-1]

  # mutate
  data[['months']] <- data$Age * 12

  # multiple regression
  # Survival predicted by Sex, months, and their interaction
  model <- glm(Survived ~ Sex * months,
               data = data,
               family = 'binomial')

  # evaluate model
  summary(model)

  }

# ---- profiling

# profile using lineprof
#profile <- lineprof(my_chunkfun_fast3())
#shine(profile)

# profile using system.time
system.time(my_chunkfun())
```

```
   user  system elapsed
  0.026   0.002   1.135
```

```
system.time(my_chunkfun_fast())
```

```
   user  system elapsed
  0.021   0.004   1.597
```

```
system.time(my_chunkfun_fast2())
```

```
   user  system elapsed
  0.018   0.004   1.535
```

```
system.time(my_chunkfun_fast3())
```

```
   user  system elapsed
  0.010   0.004   1.090
```

Speeding up code pt 2 (advanced)

9. In 95% of all cases the above steps will produce efficient-enough code. Sometimes, however, one is interested in even faster code execution. This is particularly the case when dealing with large data sets. One reason for this is that run time can be a super-linear function of data size, i.e., a twice as large data sets requires more than twice as much computation time. To see this, program a simple function that identifies the smallest value of a vector (passed on as the argument) using `sort()` and selecting the first element of the sorted vector, i.e., `sort(my_vector)[1]`. Then feed it random vectors (using `runif()`) of length either 1e5+, 1e+6, 1e+7, 1e+8, and 1e+9 and evalute the computation time (using `system.time()`). Does it increase by more or less than 10 times each step? You want to repeat this a couple of times.

Excursion: How to program functions? Functions are always defined as this `my_fun <- function(){}`. Within the parentheses you define the names of the arguments, e.g., `function(variable_1, variable_2)`. Within the curly brackets you define the function's expression, i.e., what it's supposed to do. This could be for instance `variable_1 + variable_2`, when the goal is to compute the element-wise sum of two vectors. By calling (executing) the function, the argument names inside the functions expression, i.e., `variable_1` and `variable_2` will then be replaced by the objects that were provided (passed on) as arguments. That is, if the function is provided with two vectors `my_vec_1` and `my_vec_2`, i.e., `my_fun(my_vec_1, my_vec_2)`, then the function will compute the sum of these two vectors. This requires, of course, that the provided arguments fit to whatever is done with them inside the function. In this case, the objects are thus required to be `numerical` and cannot be, e.g., of type `character`. The full function definition in this case is `my_fun <- function(variable_1, variable_2) {variable_1 + variable_2}`. After its been defined you would could call it using `my_fun(object_1, object_2)`.

```
# define function
my_fun <- function(x) sort(x)[1]

# profile using system.time
system.time(my_fun(runif(1e+5)))
```

```
   user  system elapsed
  0.008   0.000   0.008
```

```
system.time(my_fun(runif(1e+6)))
```

```
   user  system elapsed
  0.101   0.012   0.114
```

```
system.time(my_fun(runif(1e+7)))
```

```
   user  system elapsed
  1.628   0.150   1.781
```

```
system.time(my_fun(runif(1e+8)))
```

```
   user  system elapsed
 13.990   1.668  15.738
```

9. When large datasets need to be processed and speed is of the essence, it can be extremely useful to rely on multi-threaded, parallel computation. That is to run a task on multiple processors in parallel. To do this, R has relatively convenient packages, in particular, the `paralell` package, which has recently been included in the standard R library. To use parallel execution four things need to be done. (1) The data need to be split into separate jobs. For instance, a vector may be split into a list containing 100 separate pieces. (2) A function needs to be defined that performs the desired operations on a single job (one piece of the vector). (3) A cluster of workers needs to be created using, e.g., `makeCluster` . (3) The jobs and the function need to be combined in one of `parallel` 's functions. Those functions manage the passing on of jobs to individual workers in the cluster and the retrieval of the results of their computations. This particular style of programming is also known as functional programming. Now try to run the code below, which implements the function from above in this 'divide and conquer'-manner. If it runs and you understood what it does, compare its execution time to its non-parallel twin above.

```
library(parallel)

# define data
my_vec <- runif(1e+8)

# create jobs
# matrix splits data in 100 columns
# as.data.frame and as.list transform it to a list with 100 vectors
jobs <- as.list(as.data.frame(matrix(my_vec, ncol = 100)))

# define function
my_fun <- function(x) sort(x)[1]

# create a cluster with as many workers as cores
cl <- makeCluster(detectCores())

# apply function to jobs using a load balanced (LB) handler
result <- clusterApplyLB(cl, jobs, my_fun)

# flatten result and apply my_fun one last time
my_fun(unlist(result))
```

```
# load package
library(parallel)

# define parallel fun
my_parallel <- expression({

  # create jobs
  # matrix splits data in 100 columns
  # as.data.frame and as.list transform it to a list with 100 vectors
  jobs <- as.list(as.data.frame(matrix(my_vec, ncol = 100)))

  # define function
  my_fun <- function(x) sort(x)[1]

  # create a cluster with as many workers as cores
  cl <- makeCluster(detectCores())

  # apply function to jobs using a load balanced (LB) handler
  result <- clusterApplyLB(cl, jobs, my_fun)

  # stop cluster
  stopCluster(cl)

  # flatten result and apply my_fun one last time
  my_fun(unlist(result))

})

# define data
my_vec <- runif(1e+8)

# test timing
system.time(sort(my_vec)[1])
```

```
   user  system elapsed
 10.111   1.284  11.416
```

```
system.time(eval(my_parallel))
```

```
   user  system elapsed
  2.273   0.830   5.864
```

Additional reading

- For more details check out check out Hadley Wickham's Advanced R (http://adv-r.had.co.nz/).
- For more on parallel computing see Parallel R (https://www.amazon.com/dp/B005Z29QT4/ref=cm_sw_su_dp) by McCallum and Weston.