Chapter_4

Henry Bourne

2022-10-24

Functional and Object-Oriented Programming

In this chapter we will go over how to write in both the functional and object oriented programming paradigms in R. First we will look at how to programme functionally.

Functional Programming

R supports many functional programming features, perhaps the most important is that is has **first-class functions**: a programming language has first-class functions if you can define a function, functions can be arguments to other functions, functions can be returned by functions and functions can be stored in data structures.

When functional programming it is good to avoid writing functions that modify the global programme state. A **pure function** is a function which given the same arguments returns the same output and has no side affects. Although it is unrealistic that all your functions will be pure, for example functions that generate pseudo-random numbers will not be pure.

In R we have that functions only have access to variables defined in their environment. If in the environment that a function is declared there are free variables (variables used in the function that haven't been assigned a value), then the environment binds the free variables and the function is a **closure**. So if there are free variables we get back a function and we can then pass to the function the arguments which we would like it to use for the free variables. For example:

```
f <- function(x){
  add_y <- function(y){
    x + y
  }
}
add_10 <- f(10)
add_10(5)
```

```
## [1] 15
```

One thing to take especial note of is that R uses lazy evaluation. **Lazy evaluation** means that R doesn't evaluate an expression until it's value is needed. So, in the example above R doesn't evaluate the expression until we've given *add_10()* its argument and expect a value returned, ie. it doesn't evaluate the expression when we call f. This saves allot of computation, however, we must be careful as if we change variables along the way we must keep in mind that R will only evaluate when a value is needed (and therefore will use the most recently defined (current) values for variables).

Object-Oriented Programming (OOP)

Now we will look at how to do OOP in R. Note that in R this isn't very straightforward and there are actually multiple different ways in which we can programme in an object-oriented way in R. The key idea

of OOP is **polymorphism** which essentially means that a single symbol may refer to different types. In OOP we have that data and methods that operate on the data are bundled or **encapsulated** together in an object. This means that we can hide values or the state of a structured data object inside a class preventing direct access to them. The state of an object is defined by its **fields**, we can then perform operations on the state by defining **methods**. Many OOP implementations allow a class to **inherit** all the fields and methods from another class. We can use the **Unified Modelling Language (UML)** to graphically display the hierarchical structures of our classes (this is called a class diagram). We can also use OOP to create general purpose solutions to potential problems, allowing us to quickly perform a desired function without having to rewrite code, these general, reusable solutions are called design **patterns**.

In base R there are three ways of carrying out OOP. The first is S3 which is the least rigorous and can be thought of as functional OOP. Second is S4 which requires formal definitions. And lastly we have Reference Classes which implement properly encapsulated OOP, Reference Classes are also **mutable** (can be modified in-place). We start by describing S3...

S3

An **S3 object** is a base object with attribute class set to the class name. Everything in R is an object, however not all of these objects are object-oriented (OO) objects, **base objects** are objects that don't have a class attribute (and hence aren't OO objects). An **S3 object** is simply a base object but with a class attribute (set to the class name). We use the *class()* function to set the class attribute of an object and hence create an S3 object, we can also use the *unclass()* function to strip the class name attribute and turn it back into a non OO object. It is good practice to have a **constructor** function that a user can call to initialize an object correctly and then return it, like how you would create an object in other languages such as python. It's also recommended to include a **helper** function which is there to correct common mistakes on the fly.

We can add a method to an S3 object by assigning a function to the following name ., note that this is only for implementing methods for which there exists a generic function of the same name such as plot or print. Also note that the method must take the same arguments as the generic. If we want to we can create a new generic function using *<- function( ){ UseMethod("") }* then assign a method with the new generic method name to our S3 class. We can also inherit in S3 by setting the class attribute to a vector of class names, to do this the child class will have a vector containing the names of it and its parent classes.

S4

Compared to S3, S4 has a more formal approach to OOP. In S4 we have to define classes in a more formal way, the advantage of this is that it makes working with S3 more clear and allows for build-in integrity checks. To use S4 we need to make sure we have the "methods" package installed, after we've done this we can then define a class as follows:

```
library(methods)
setClass("class_name",
   slots = c(
     slot_1  = "numeric",
     slot_2 = "logical",
     slot_3  = "character"
     )
)
```

Here we have defined a class with the name "class_name" and three attributes labelled slot_1, slot_2, slot_3 with types numeric, logical and character, ie. to create a class we call *setClass()*, passing it two arguments, the first being the class name and the second a vector of **slots** (same as fields or methods in other OOP languages). After defining the slots we need to define the methods, we will now show an example of a special method for a class which is the initialization method (the method we call to create an object from the class):

```
setMethod("initialize", "class_name",
  function(.Object, slot_1, slot_2) {
    .Object@slot_1  <- slot_1
    .Object@slot_2 <- slot_2

    if(slot_2){
      x <- as.character(slot_1)
    }
    else{
      x <- "no number"
    }
    .Object@slot_3 <- x
    return(.Object)
  }
)
```

To create a method therefore, we call *setMethod*. The first argument we pass to it is the name of the generic function we want to implement a method for, here we use "initialize" which is a non-standard generic function which is called within the function *new()* (which constructs a new object of the class corresponding to the class name for which it is a method). The second argument is the *signature*, which is the classes that the arguments of the function need for this implementation of the function to be able to run. The third argument is the function for the method, here the function has .Object as the first argument, this is because to create a new object we call *new()* with the first argument set to the class name, this then calls initialize, passing the object prototype for the class to the .Object argument. The next two arguments specify that only two more arguments are to be given to *new()* when we are creating an object, here we say that two arguments slot_1 and slot_2 should be given. An example of how we would create an object is:

```
new("class_name", 1, TRUE)
```

```
## An object of class "class_name"
## Slot "slot_1":
## [1] 1
##
## Slot "slot_2":
## [1] TRUE
##
## Slot "slot_3":
## [1] "1"
```

In the function itself we can set the values of slots by writing *.Object@* <- , to achieve proper encapsulation the @ operator should only be used within method definitions. To allow the user to change or fetch slot values we should create getter and setter methods. An example of how we would define a getter and setter method for slot_1 is:

```
setGeneric("slot_1", function(x) standardGeneric("slot_1"))
```

```
## [1] "slot_1"
```

```
setGeneric("slot_1<-", function(x, value) standardGeneric("slot_1<-"))
```

```
## [1] "slot_1<-"
```

```
setMethod("slot_1", "class_name", function(x) x@regressor)
setMethod("slot_1<-", "class_name",
  function(x, value) {
    x <- initialize(x, slot_1 = value, slot_2 = x@slot_2)
```

```
    validObject(x)
    return(x)
  }
)
```

In the final line of the function we return the object, hence when we call *new()* we get given back an instantitation of the class with all its slots set appropriately. Note that again in S4 methods must be implementations of a generic function, to create a new generic function we can use the command *setGeneric()*. It is often useful to check in our methods that we have a valid object, to do this we can use the command *validObject()*, this is one of the build-in mechanisms we mentioned earlier. By default it checks that all slots are present and of the correct type, we can also add validity check by using the function *setValidity()* and passing it the class name for which we want to add a validity check and a method which carries out the validity check.

We can also have relationships between objects in S4. For example two individual objects can be related to each other, to model this relation we create a slot in each of the classes for which their objects will be related and make the type of the slot the class of the related object. When all the objects of one class are related to all of the objects of another class (ie. one class inherits or is a more general version of another class) then we can model this by adding the *contains* argument to the call of setClass and setting it to a vector of the classes from which it should inherit. This then means that the child class will have all the slots the parent class has and methods. In regards to methods if for a particular method the child class has an implementation of this method then this will be called, if it doesn't then the parent class will be checked for this method and used if it has it, if not then the grandparent class will be checked and so on.

Reference Classes

The final way of doing OOP in R that we will discuss is reference classes. Reference classes provide a way of conducting OOP in R with a higher degree of encapsulation than with S3 and S4, it also allows for modify-in-place (memory changed directly) as opposed to copy-on-modify (creates a copy in memory with modification) which is used most of the time in R. We will explain how reference classes work within the context of an example, what we will do is create a class that can carry out cross-validation for the "stattools" package. First we install the stattools package:

```
library(stattools)
# Can use "devtools::install_github("h-aze/compass_yr1", subdir = "/labs/stattools")" if
↪  not already downloaded
```

Currently the cross_validation function in "stattools" is implemented as so:

```
regr_cross_val <- function(D, y, RM=LLS, k=10, ...){

  # We randomly shuffle the indices of the rows and using these randomized indices
  ↪  shuffle the rows of the dataset and the target variable (in the same order)
  if(is.vector(D)){
    ind <- sample(length(D))
    D_dash <- D[ind]
    y_dash <- y[ind]
  }
  else if(is.array(D)){
    ind <- sample(nrow(D))
    D_dash <- D[ind,]
    y_dash <- y[ind]
  }
  else{
    ind <- sample(nrow(D))
```

```
  D_dash <- D[paste(ind),]
  y_dash <- y[ind]
}

# We now create a list which indexes the rows in our dataset, we will use this to
↪  select groups of certain rows from the dataset.
#   Hence what we have effectively done here is partition the dataset (randomly - as we
↪  shuffled the dataset previously) into groups.
#   We split the dataset into 10 groups as we will be performing 10-fold cross
↪  validation.
if(is.vector(D)){
  subsets <- cut(seq(1,length(D)), breaks = k, labels = FALSE)
}
else{
  subsets <- cut(seq(1,nrow(D)), breaks = k, labels = FALSE)
}

# We create a vector to store our cross-val errors
errors <- c()

# Loop for carrying out 10-fold cross-val
#TODO: be able to adjust the number of folds
for(i in 1:k){

  # We segment our data into testing, D.test, and training, D.train, datasets
  testIndexes <- which(subsets==i,arr.ind=TRUE)
  if(is.vector(D)){
    D.test <- D_dash[testIndexes]
    D.train <- D_dash[-testIndexes]
  }
  else{
    D.test <- D_dash[testIndexes, ,drop=FALSE]
    D.train <- D_dash[-testIndexes, ,drop=FALSE]
  }

  # We get the model matrix and predictor variables for the training data
  # and then we find the LS estimator
  X.train <- model_matrix(D.train)
  y.train <- y_dash[-testIndexes]
  w <- RM(X.train, y.train,...)

  # We get the model matrix and predictor variables for the testing data
  # and then we calculate the least squares testing error
  X.test <- model_matrix(D.test)
  y.test <- y_dash[testIndexes]
  test.error <- norm(y.test - X.test %*% w, type="2")**2
  # We add the testing error to our error vector
  errors[i] <- test.error

}

# We find and return the cross-val error
error.CV <- sum(errors)/k
```

```
  error.CV
}
```

The problem with this function is that it can be a hastle to use. If we want to carry out cross-validation multiple times with a specific regression method we must specify it each time and if the regression method has hyper-parameters then we must first specify the regression method, RM, we want to use and the k we want to use, even if we are happy with the default values. Further we may sometimes want more information from this function, for example we may want a list of all the estimators found and their corresponding errors so we can find the estimator that produced the minimal error. We may also want to add further functionality such as the ability to change the cost function. By writing a class for this function we will be able to solve all of the above problems and add extra functionality, whilst making it easier for the end user.

We have already made sure that we have the methods package in this RMarkdown script, so now we can create the class:

```
CrossValidation <- setRefClass("CrossValidation",
                               fields=c( data="numeric",
                                         target="numeric",
                                         k="integer",
                                         Regr_method="ANY",
                                         Regr_method.name="character",
                                         E_fun = "ANY",
                                         E_fun.name = "character",
                                         Feat_trans = "ANY",
                                         Feat_trans.name = "character",
                                         k_test_errors = "numeric",
                                         estimators = "list",
                                         cv_error = "numeric",
                                         flag = "logical"
                                         )
                               )
```

Here we have created a class called CrossValidation and defined the fields it will have along with their names. Note that in reference classes they are called fields and not slots like in S4. Now we would like to define some methods for the class:

```
CrossValidation$methods(
  initialize = function(data, target, k=as.integer(10), Regr_method=LLS, E_fun=E_l2 ) {
    .self$data <- data
    .self$target <- target
    .self$k <- k

    .self$setRegr_method(Regr_method)
    .self$Regr_method.name <- as.character(substitute(Regr_method)) # We set the name
↪   again when initializing otherwise will use argument name
    .self$setE_fun(E_fun)
    .self$E_fun.name <-as.character(substitute(E_fun)) # We set the name again when
↪   initializing otherwise will use argument name
    .self$Feat_trans <- NULL
    .self$Feat_trans.name <- ""
    .self$k_test_errors <- numeric(0)
    .self$estimators <- vector(mode = "list", length = 0)
    .self$cv_error <- numeric(0)
    .self$flag <- FALSE
  },
```

```
# Carries out k-fold cross-validation for a regression problem
regr_cv = function(){
  .self$flag <- TRUE
  # We randomly shuffle the indices of the rows and using these randomized indices
  ↪  shuffle the rows of the dataset and the target variable (in the same order)
  if(is.vector(.self$data)){
    ind <- sample(length(.self$data))
    D_dash <- .self$data[ind]
    y_dash <- .self$target[ind]
  }
  else if(is.array(.self$data)){
    ind <- sample(nrow(.self$data))
    D_dash <- .self$data[ind,]
    y_dash <- .self$target[ind]
  }
  else{
    ind <- sample(nrow(.self$data))
    D_dash <- .self$data[paste(ind),]
    y_dash <- .self$target[ind]
  }

  # We now create a list which indexes the rows in our dataset, we will use this to
  ↪  select groups of certain rows from the dataset.
  if(is.vector(.self$data)){
    subsets <- cut(seq(1,length(.self$data)), breaks = .self$k, labels = FALSE)
  }
  else{
    subsets <- cut(seq(1,nrow(.self$data)), breaks = .self$k, labels = FALSE)
  }

  # We assign empty vectors to fields k_test_errors and estimators so that we can
  ↪  assign them values in the upcoming for loop
  .self$k_test_errors <- vector(mode="numeric", length=.self$k)
  .self$estimators <- vector("list", length = .self$k)

  # Loop for carrying out k-fold cross-val
  for(i in 1:.self$k){
    # We segment our data into testing, D.test, and training, D.train, datasets
    testIndexes <- which(subsets==i,arr.ind=TRUE)
    if(is.vector(.self$data)){
      D.test <- D_dash[testIndexes]
      D.train <- D_dash[-testIndexes]
    }
    else{
      D.test <- D_dash[testIndexes, ,drop=FALSE]
      D.train <- D_dash[-testIndexes, ,drop=FALSE]
    }

    # We get the model matrix and predictor variables for the training data and then we
    ↪  find the estimator using our regression method
    if(is.null(.self$Feat_trans)){
      X.train <- model_matrix(D.train)
```

```
    }
    else{
      X.train <-model_matrix(.self$Feat_trans(D.train))
    }
    y.train <- y_dash[-testIndexes]
    w <- .self$Regr_method(X.train, y.train)
    .self$estimators[[i]] <- w

    # We get the model matrix and predictor variables for the testing data and then we
    ↪  calculate the error using our error function
    if(is.null(.self$Feat_trans)){
      X.test <- model_matrix(D.test)
    }
    else{
      X.test <-model_matrix(.self$Feat_trans(D.test))
    }
    y.test <- y_dash[testIndexes]
    test.error <- .self$E_fun(y.test, X.test %*% w)

    # We add the testing error to our error vector
    .self$k_test_errors[i] <- test.error

  }

  # We find and return the cross-val error
  .self$cv_error <- sum(.self$k_test_errors)/.self$k
  .self$cv_error
},

show = function() {
  cat("head(data)    =", head(.self$data), "\n", sep=" ")
  cat("head(target)  =", head(.self$target), "\n", sep=" ")
  cat("k             =", .self$k, "\n", sep=" ")
  cat("Regr_method   =", .self$Regr_method.name, "\n", sep=" ")
  cat("E_fun         =", .self$E_fun.name, "\n", sep=" ")
  if(.self$flag){
    cat("k_test_errors =", head(.self$k_test_errors), "\n", sep=" ")
    for (v in 1:length(head(.self$estimators))) {
      cat(paste("estimators[",as.character(v), "] =", sep=""),
      ↪  head(.self$estimators)[[v]], "\n", sep=" ")
    }
    cat("cv_error      =", .self$cv_error, "\n", sep=" ")
  }
},

# When called with problem type (at the moment only support regression:"r") computes
↪  the cv error and returns it
getCv_error = function(t="r") {
  if (t == "r"){
    .self$regr_cv()
    return(.self$cv_error)
  }
  else{
```

```
    stop("Not given valid problem type")
  }
},

getEstimators = function() {
  return(.self$estimators)
},

getK_test_errors = function() {
  return(.self$k_test_errors)
},

# Allows user to change the data that cv will be performed on (user must give both data
↪  and targets)
setData = function(data, target) {
  .self$data <- data
  .self$target <- target
},

# Checks that the function passed to it has closure
check_closure = function(f){
  if (typeof(f) == "closure"){
    return(TRUE)
  }
  else{
    return(FALSE)
  }
},

# Sets the regression method passed to it to Regr_method if function is of type
↪  "closure"
setRegr_method = function(Regr_method){
  if (.self$check_closure(Regr_method)){
    .self$Regr_method <- Regr_method
    .self$Regr_method.name <- as.character(substitute(Regr_method))
  }
  else{
    stop("Regr_method doesnt have type: closure")
  }
},

# Sets the error function passed to it to E_fun if function is of type "closure"
setE_fun = function(E_fun){
  if (.self$check_closure(E_fun)){
    .self$E_fun <- E_fun
    .self$E_fun.name <-as.character(substitute(E_fun))
  }
  else{
    stop("E_fun doesnt have type: closure")
  }
},

# Sets the value of k and makes sure its an integer
```

```
  setk = function(i){
    if (is.integer(i)){
      .self$k <- i
    }
    else{
      stop("setk was not given an integer")
    }
  },

  setFeat_trans = function(Feat_trans){
    if (.self$check_closure(Feat_trans)){
      .self$Feat_trans <- Feat_trans
      .self$Feat_trans.name <-as.character(substitute(Feat_trans))
    }
    else{
      stop("Feat_trans doesnt have type: closure")
    }
  }
)
```

The first function above is the *initialize function*, it initializes all the fields in the correct manner. We then have *regr_cv* which carries out cross validation for a regression problem (note we have set up the class such that we could add the ability to do classification at a later stage) using information from the objects fields. We then have multiple setter function such as *setRegr_method* which checks what we are trying to set the new regression method to is of type closure and *setData* which allows the user to set the predictor and target variables to be worked on. Note that through encapsulation the user is only allowed to set **both** the predictor and target variables, this is on purpose so that a user doesn't accidentally update one and then forget to update the other (usually new predictor variables will correspond to new target variables or vice-versa). We also have multiple getter methods such as *getCv_error*, this computes and returns the cross-validation error according to what the objects fields are set to. Finally, we have a *show* method which when called prints information about all the fields of the class in a nice easy to read format so the user may see clearly the current state of the object.

Now we will test out this class, first we create some data to test it on and initialize the class,

```
x <- runif(100,0,10)
y <- 2*x -1 + rnorm(100)
cv <- CrossValidation$new(data=x,  target=y) ;cv
```

```
## head(data)    = 9.477833 2.934199 5.659784 2.354751 9.476448 5.512583
## head(target)  = 19.70247 4.930106 8.99729 4.175376 16.1883 8.577505
## k             = 10
## Regr_method   = LLS
## E_fun         = E_l2
```

Now we run cross validation on our toy data:

```
cv$regr_cv();
```

```
## [1] 11.2263
```

```
cv
```

```
## head(data)    = 9.477833 2.934199 5.659784 2.354751 9.476448 5.512583
## head(target)  = 19.70247 4.930106 8.99729 4.175376 16.1883 8.577505
## k             = 10
```

```
## Regr_method   = LLS
## E_fun         = E_l2
## k_test_errors = 12.22132 4.104099 6.854663 9.985063 14.0192 9.685678
## estimators[1] = -0.974611 1.983977
## estimators[2] = -1.086697 2.004172
## estimators[3] = -1.064591 2.000383
## estimators[4] = -0.8719819 1.979565
## estimators[5] = -0.9420648 1.994819
## estimators[6] = -1.044356 1.994205
## cv_error      = 11.2263
```

How about if we want to change the regression method and k that we are using? let's do that now:

```
cv$setRegr_method(LLS_R(2))
cv$setk(as.integer(5))
cv$getCv_error("r")
```

```
## [1] 22.27975
```

Here we set the regression method with the LLS_R function from stattools with the lambda parameter set to 2, also note we are actually doing some functional programming here, we only pass the hyper-parameter for the regression method so it returns a function (which will use the hyper-parameter we set) and give this as the regression method to be used by the object.

What about if we want to change the data? let's do that:

```
x <- women$height
y <- women$weight
cv$setData(x,y)
cv$getCv_error("r")
```

```
## [1] 139.5472
```

We have successfully changed the data and then carried out cross-validation again, however, our cv error has gone up alot! perhaps doing a polynomial feature transform will help us better model this new dataset:

```
ft <- poly_feat_trans(2)
cv$setFeat_trans(ft)
cv$getCv_error("r")
```

```
## [1] 5.782755
```

We see that now we are getting a much lower cv error. We now will implement this class in the stattools package, this concludes the section on reference classes and this chapter of the portfolio.