SathyaNarayanan Srinivasan

Sxs142031@utdallas.edu

Abstract

Implementation of the rpart package in R

Analysis using rpart package

Machine Learning Assignment - 3

# Requirements

* **Operating System** – Windows 7 or Higher
* **Product Version :** [**R-3.2.2**](https://cran.r-project.org/bin/windows/base/R-3.2.2-win.exe)
* **PDF Processor** – Adobe PDF Reader

# Prerequisites

* Please ensure that R ver. 3.2.2 is installed properly and working in the machine.
* Please ensure that rscript is accessible in the command prompt.i.e. when rscript is typed in the command prompt the system recognizes that it is either a internal or an external command. If not, please follow the steps here : [Debug Link](http://stackoverflow.com/questions/17339438/r-script-from-command-line)
* This program won’t compile in RStudio as it accepts command line parameters
* Also open Command Prompt in ELEVATED PRIVILEGES mode.

# Compiling and Running Instructions

* In a windows command prompt :
  + Type rscript <full path to the file Assignment3.R> with the .R extension
  + Press Enter
  + Wait for the packages to be installed – it will take a while depending on many factors
  + Press 1 to generate a RPlots.PDF similar to the file bundled with the source to see the output of Kyphosis data set
  + Press 2 to generate a RPlots1.PDF similar to the file bundled with the source to see the output of Solder data set
  + The output is ordered according the question of assignment 3.

# **Output – PLEASE NOTE (IMPORTANT)**

* **For the kyphosis data set, sometimes only ROOT node will be printed as a pruned tree. This is due to the randomly generated and cross validated Cp value. This is not an error and will be rectified by itself upon successive evaluations**.
* **“Assignment3withRattle.R” file has dependency issues with GTK+ that can be averted if run in a Mac or Winodws with GTK+ already installed** - This is just to create colorful charts.

|  |  |
| --- | --- |
| **Data Set** | **Accuracy** |
| 80 % training Set - Kyphosis | 82.3 % |
| 90 % training Set - Kyphosis | 77.8% |
| 80 % training Set – Solder | 62.5% |
| 90 % training Set - Solder | 66.6% |

OUTPUT AS IN COMMAND PROMPT

D:\DriveFiles\Fall15\ML\Assignment Solutions>dir

Volume in drive D is Data

Volume Serial Number is 3621-64B7

Directory of D:\DriveFiles\Fall15\ML\Assignment Solutions

09/25/2015 09:24 PM <DIR> .

09/25/2015 09:24 PM <DIR> ..

09/25/2015 11:20 PM <DIR> Assignment 3

0 File(s) 0 bytes

3 Dir(s) 533,442,695,168 bytes free

D:\DriveFiles\Fall15\ML\Assignment Solutions>cd "Assignment 3"

D:\DriveFiles\Fall15\ML\Assignment Solutions\Assignment 3>rscript Assignment3.R

trying URL 'http://cran.rstudio.com/bin/windows/contrib/3.2/rpart\_4.1-10.zip'

Content type 'application/zip' length 921967 bytes (900 KB)

==================================================

downloaded 900 KB

package 'rpart' successfully unpacked and MD5 sums checked

The downloaded binary packages are in

C:\Users\SathyaSrini\AppData\Local\Temp\RtmpIrBZ3o\downloaded\_packages

Type 1 for Kyphosis Tree and 2 for Solder Tree : 1

dev.new(): using pdf(file="Rplots2.pdf")

[1] "============== BEGIN KYPHOSIS RESULTS =====================================

==============="

Classification tree:

rpart(formula = Kyphosis ~ Age + Number + Start, data = kyphosis,

method = "class")

Variables actually used in tree construction:

[1] Age Start

Root node error: 17/81 = 0.20988

n= 81

CP nsplit rel error xerror xstd

1 0.176471 0 1.00000 1.0000 0.21559

2 0.019608 1 0.82353 1.1176 0.22433

3 0.010000 4 0.76471 1.1176 0.22433

[1] "Printing prediction for 80 percent of data : "

[1] absent absent absent absent absent absent absent absent absent absent

[11] absent absent absent absent absent absent absent

Levels: absent present

[1] "Printing accuracy for 80 percent of data : "

[1] 76.47059

[1] "Printing prediction for prediction for 90 percent of data: "

[1] absent absent absent absent absent absent absent absent absent

Levels: absent present

[1] "Printing accuracy for 90 percent of data :"

[1] 66.66667

[1] "============== END OF KYPHOSIS RESULTS ====================================

================"

D:\DriveFiles\Fall15\ML\Assignment Solutions\Assignment 3>rscript Assignment3.R

trying URL 'http://cran.rstudio.com/bin/windows/contrib/3.2/rpart\_4.1-10.zip'

Content type 'application/zip' length 921967 bytes (900 KB)

==================================================

downloaded 900 KB

package 'rpart' successfully unpacked and MD5 sums checked

The downloaded binary packages are in

C:\Users\SathyaSrini\AppData\Local\Temp\Rtmpwhtc7W\downloaded\_packages

Type 1 for Kyphosis Tree and 2 for Solder Tree : 2

[1] "============== BEGIN SOLDER RESULTS =======================================

============="

dev.new(): using pdf(file="Rplots3.pdf")

[1] "Printing prediction for 80 percent of data : "

2 3 5 11 18 20 23 25 31 34 40 41 42

Thick Thick Thick Thick Thick Thick Thick Thick Thick Thick Thick Thick Thick

43 48 51 54 55 57 62 63 64 67 68 76 81

Thick Thick Thick Thick Thick Thick Thick Thick Thick Thin Thin Thick Thick

92 94 101 104 108 120 121 134 137 152 154 155 156

Thick Thick Thick Thick Thick Thick Thick Thick Thick Thin Thin Thin Thin

158 159 165 166 167 173 176 181 190 193 207 223 237

Thin Thin Thin Thin Thin Thin Thick Thick Thick Thick Thick Thick Thick

240 249 257 265 277 278 285 287 290 299 304 306 309

Thick Thick Thin Thick Thin Thin Thick Thick Thick Thick Thick Thin Thin

316 318 319 329 330 333 337 341 343 344 347 355 358

Thin Thin Thick Thin Thin Thin Thin Thin Thin Thin Thin Thick Thin

366 369 370 371 378 381 390 393 409 415 421 422 428

Thick Thick Thick Thick Thick Thick Thick Thick Thick Thick Thick Thin Thin

444 450 458 463 466 469 493 495 499 505 518 522 524

Thin Thin Thin Thick Thick Thick Thick Thick Thick Thick Thin Thin Thin

530 532 534 536 540 545 546 549 551 553 556 564 585

Thin Thin Thin Thin Thin Thick Thick Thick Thick Thick Thick Thick Thick

589 592 594 599 605 610 611 615 616 623 631 637 638

Thin Thin Thick Thick Thin Thin Thin Thin Thin Thin Thin Thin Thin

646 655 664 674 680 689 698 705 709 712 713 714 717

Thin Thick Thin Thin Thin Thin Thin Thin Thin Thin Thin Thin Thin

718

Thin

Levels: Thick Thin

[1] "Printing accuracy for 80 percent of data : "

[1] 72.22222

[1] "Printing prediction for 90 percent of data: "

12 14 34 40 44 77 103 113 117 128 130 131 150

Thick Thick Thick Thick Thick Thick Thick Thick Thick Thick Thick Thick Thick

154 155 164 165 167 173 175 183 188 211 214 218 219

Thin Thin Thin Thin Thin Thin Thin Thick Thick Thick Thick Thin Thick

229 238 280 295 314 321 322 335 346 351 353 355 356

Thick Thick Thick Thick Thin Thick Thick Thin Thin Thin Thin Thick Thick

357 360 368 372 395 404 405 423 437 453 465 471 487

Thick Thin Thin Thin Thick Thick Thick Thin Thin Thick Thin Thin Thick

492 494 498 501 511 520 547 564 565 575 597 598 609

Thick Thick Thin Thick Thin Thin Thick Thick Thick Thin Thick Thick Thin

617 638 672 684 687 689 703

Thin Thin Thin Thin Thin Thin Thin

Levels: Thick Thin

[1] "Printing accuracy for 90 percent of data :"

[1] 62.5

[1] "============== END SOLDER RESULTS =========================================

==========="

D:\DriveFiles\Fall15\ML\Assignment Solutions\Assignment 3>rscript Assignment3wit

hRattle.R

trying URL 'http://cran.rstudio.com/bin/windows/contrib/3.2/rpart\_4.1-10.zip'

Content type 'application/zip' length 921967 bytes (900 KB)

==================================================

downloaded 900 KB

package 'rpart' successfully unpacked and MD5 sums checked

The downloaded binary packages are in

C:\Users\SathyaSrini\AppData\Local\Temp\RtmpYJ1yrw\downloaded\_packages

Warning: dependencies 'graph', 'RBGL', 'pkgDepTools', 'Rgraphviz' are not availa

ble

trying URL 'http://cran.rstudio.com/bin/windows/contrib/3.2/rattle\_3.5.0.zip'

Content type 'application/zip' length 3689738 bytes (3.5 MB)

==================================================

downloaded 3.5 MB

package 'rattle' successfully unpacked and MD5 sums checked

The downloaded binary packages are in

C:\Users\SathyaSrini\AppData\Local\Temp\RtmpYJ1yrw\downloaded\_packages

trying URL 'http://cran.rstudio.com/bin/windows/contrib/3.2/partykit\_1.0-3.zip'

Content type 'application/zip' length 1212533 bytes (1.2 MB)

==================================================

downloaded 1.2 MB

package 'partykit' successfully unpacked and MD5 sums checked

The downloaded binary packages are in

C:\Users\SathyaSrini\AppData\Local\Temp\RtmpYJ1yrw\downloaded\_packages

trying URL 'http://cran.rstudio.com/bin/windows/contrib/3.2/klaR\_0.6-12.zip'

Content type 'application/zip' length 318050 bytes (310 KB)

==================================================

downloaded 310 KB

package 'klaR' successfully unpacked and MD5 sums checked

The downloaded binary packages are in

C:\Users\SathyaSrini\AppData\Local\Temp\RtmpYJ1yrw\downloaded\_packages

Loading required package: RGtk2

Rattle: A free graphical interface for data mining with R.

Version 3.5.0 Copyright (c) 2006-2015 Togaware Pty Ltd.

Type 'rattle()' to shake, rattle, and roll your data.

Loading required package: grid

Loading required package: MASS

Type 1 for Kyphosis Tree and 2 for Solder Tree :

1

dev.new(): using pdf(file="Rplots4.pdf")

Classification tree:

rpart(formula = Kyphosis ~ Age + Number + Start, data = kyphosis,

method = "class")

Variables actually used in tree construction:

[1] Age Start

Root node error: 17/81 = 0.20988

n= 81

CP nsplit rel error xerror xstd

1 0.176471 0 1.00000 1.0000 0.21559

2 0.019608 1 0.82353 1.1765 0.22829

3 0.010000 4 0.76471 1.1765 0.22829

Error in apply(model$frame$yval2[, yval2per], 1, function(x) x[1 + x[1]]) :

dim(X) must have a positive length

Calls: kyfn -> fancyRpartPlot -> apply

Execution halted

D:\DriveFiles\Fall15\ML\Assignment Solutions\Assignment 3>rscript Assignment3wit

hRattle.R

trying URL 'http://cran.rstudio.com/bin/windows/contrib/3.2/rpart\_4.1-10.zip'

Content type 'application/zip' length 921967 bytes (900 KB)

==================================================

downloaded 900 KB

package 'rpart' successfully unpacked and MD5 sums checked

The downloaded binary packages are in

C:\Users\SathyaSrini\AppData\Local\Temp\Rtmp0OojSA\downloaded\_packages

Warning: dependencies 'graph', 'RBGL', 'pkgDepTools', 'Rgraphviz' are not availa

ble

trying URL 'http://cran.rstudio.com/bin/windows/contrib/3.2/rattle\_3.5.0.zip'

Content type 'application/zip' length 3689738 bytes (3.5 MB)

==================================================

downloaded 3.5 MB

package 'rattle' successfully unpacked and MD5 sums checked

The downloaded binary packages are in

C:\Users\SathyaSrini\AppData\Local\Temp\Rtmp0OojSA\downloaded\_packages

trying URL 'http://cran.rstudio.com/bin/windows/contrib/3.2/partykit\_1.0-3.zip'

Content type 'application/zip' length 1212533 bytes (1.2 MB)

==================================================

downloaded 1.2 MB

package 'partykit' successfully unpacked and MD5 sums checked

The downloaded binary packages are in

C:\Users\SathyaSrini\AppData\Local\Temp\Rtmp0OojSA\downloaded\_packages

trying URL 'http://cran.rstudio.com/bin/windows/contrib/3.2/klaR\_0.6-12.zip'

Content type 'application/zip' length 318050 bytes (310 KB)

==================================================

downloaded 310 KB

package 'klaR' successfully unpacked and MD5 sums checked

The downloaded binary packages are in

C:\Users\SathyaSrini\AppData\Local\Temp\Rtmp0OojSA\downloaded\_packages

Loading required package: RGtk2

Rattle: A free graphical interface for data mining with R.

Version 3.5.0 Copyright (c) 2006-2015 Togaware Pty Ltd.

Type 'rattle()' to shake, rattle, and roll your data.

Loading required package: grid

Loading required package: MASS

Type 1 for Kyphosis Tree and 2 for Solder Tree : 2

dev.new(): using pdf(file="Rplots5.pdf")

[1] "Printing prediction for 80 percent of data : "

5 7 18 21 25 28 37 41 51 55 57 64 66

Thick Thick Thick Thick Thick Thick Thick Thick Thick Thick Thick Thick Thick

68 73 74 76 80 89 90 94 107 110 115 128 135

Thin Thick Thick Thick Thick Thick Thick Thick Thick Thick Thick Thick Thick

142 146 149 152 154 160 162 167 169 172 177 182 186

Thick Thick Thick Thin Thin Thin Thin Thin Thin Thin Thin Thick Thick

190 194 196 200 209 216 218 221 227 228 232 240 249

Thick Thick Thick Thick Thick Thick Thin Thick Thick Thick Thick Thick Thick

263 274 275 277 279 298 306 318 321 323 324 327 353

Thin Thick Thick Thin Thick Thick Thin Thin Thick Thick Thin Thick Thin

355 358 360 361 364 365 369 376 377 379 387 394 396

Thick Thin Thin Thick Thick Thick Thick Thick Thick Thick Thick Thick Thick

397 407 413 430 442 446 447 448 451 455 457 465 467

Thick Thick Thick Thick Thin Thick Thick Thin Thick Thin Thin Thin Thick

471 473 482 484 490 493 499 501 503 506 516 522 523

Thin Thin Thick Thick Thick Thick Thick Thick Thin Thick Thin Thin Thin

524 526 529 534 543 546 555 556 570 571 572 584 587

Thin Thin Thin Thin Thin Thick Thick Thick Thick Thick Thick Thick Thin

599 600 605 607 609 610 612 617 624 629 636 637 639

Thick Thick Thin Thin Thin Thin Thin Thin Thin Thin Thin Thin Thin

645 649 654 662 669 681 689 698 701 710 711 715 718

Thin Thin Thin Thin Thin Thin Thin Thin Thin Thin Thin Thin Thin

720

Thin

Levels: Thick Thin

[1] "Printing accuracy for 80 percent of data : "

[1] 71.52778

[1] "Printing prediction for 90 percent of data: "

17 19 20 33 41 49 67 69 73 74 78 92 119

Thick Thick Thick Thick Thick Thick Thin Thin Thick Thick Thick Thick Thick

120 121 123 125 127 130 142 150 154 155 159 163 169

Thick Thick Thick Thick Thick Thick Thick Thick Thin Thin Thin Thin Thin

172 198 201 204 225 229 230 247 253 293 314 327 349

Thin Thick Thick Thick Thick Thick Thick Thick Thick Thick Thin Thick Thin

352 362 392 414 429 436 438 445 447 450 455 463 467

Thin Thick Thick Thick Thin Thin Thin Thick Thick Thin Thin Thick Thick

472 476 483 522 535 537 540 544 561 578 581 596 597

Thick Thick Thin Thin Thin Thin Thin Thick Thick Thin Thick Thick Thick

626 630 657 682 685 712 720

Thin Thin Thick Thin Thin Thin Thin

Levels: Thick Thin

[1] "Printing accuracy for 90 percent of data :"

[1] 63.88889

n= 720

node), split, n, loss, yval, (yprob)

\* denotes terminal node

1) root 720 360 Thick (0.50000000 0.50000000)

2) skips< 2.5 432 155 Thick (0.64120370 0.35879630)

4) Mask=B6 51 4 Thick (0.92156863 0.07843137) \*

5) Mask=A1.5,A3,B3 381 151 Thick (0.60367454 0.39632546)

10) skips< 0.5 210 68 Thick (0.67619048 0.32380952) \*

11) skips>=0.5 171 83 Thick (0.51461988 0.48538012)

22) Mask=A3,B3 112 48 Thick (0.57142857 0.42857143)

44) PadType=D7,L4,L6,L7,L8,L9,W4,W9 95 37 Thick (0.61052632 0.3894736

8) \*

45) PadType=D4,D6 17 6 Thin (0.35294118 0.64705882) \*

23) Mask=A1.5 59 24 Thin (0.40677966 0.59322034) \*

3) skips>=2.5 288 83 Thin (0.28819444 0.71180556) \*

D:\DriveFiles\Fall15\ML\Assignment Solutions\Assignment 3>Assignment3withRattle.

R