DSCI353-353m-453: 01a-f Intro Class

2001-353-353m-453-00a-f-IntroClass

Roger H. French, Peitian Wang

14 January, 2020

Contents

1.1.1.1 Reading, Homeworks, Projects, SemProjects

- Readings:
  - R4DS Chapters 1,2,3 In Explore section
  - If you are new to R, Read Peng-EDAwR
    - And his Youtube Playlists of Computing for Data Analysis
    - Peng-Computing For Data Analysis Playlist
- Laboratory Exercises:

  - LE0, a no credit excercise, is a useful intro to R
  - For those new to R
- SemProjects:
  - SemProjects: have 4 parts, we'll have reports on Sect. 1,2 then 3,4
    - SemProj Report Out #1 in Class w07a,07b, Tues/Thurs, Feb. 25,27
    - SemProj Report Out #2 in Class w10a,10b, Tues/Thurs, March 24,26
    - SemProj Report Out #3 in Class w14a,14b, Tues/Thurs April 21,23
    - SemProj Report #4 is the full, comprehensive project due at final exam.
  - These are Peer Graded
  - Assistance on SemProjects is done in DSCI352-352m-452 Class
    - DSCI352 meetings during Friday Community Hour, 12:45 to 1:45pm in Olin 303
    - Is taught by Prof. Laura Bruckman (lsh41@case.edu)
- Office Hours:
  - Mondays, Wednesdays 4pm to 5pm in White 540
- Final Exam
  - Thursday March 30th, 2020, 12 noon to 3pm

1.1.1.2 If you are new to R (Or want a quick refresher)

- You can do Lab Excercise LE0a and LE0b
  - These are from Chapter 1,2 of Open Intro Stats (OISv3)

1.1.1.3 Textbooks

1.1.1.3.1 Introduction to R and Data Science

For students new to R, Coding, Inferential Statistics

- Peng: R Programming for Data Science
- Peng: Exploratory Data Analysis with R
- OIS = Diez, Barr, Çetinkaya-Runde: Open Intro Stat v4

1.1.1.3.2 Textbooks for this class

- R4DS = Wickham, Grolemund: R for Data Science
- ISLR = James, Witten, Hastie, Tibshirani: Intro to Statistical Learning with R
- ESL = Trevor Hastie, Tibshirani, Friedman: Elements of Statistical Learning
- DLwR = Chollet, Allaire: Deep Learning with R

1.1.1.4 The DSCI courses and class sections

1.1.1.4.1 In these Applied Data Science (DSCI) classes

- We focus on teaching all necesary data science skills
  - Including coding in R
  - Use of Rmarkdown for data analysis reports and presentations
  - Git for code versioning and collaboration
  - Linear and non-linear regression and classification
  - Beyond linear modeling, including Support Vector Machines, Random Forest
  - Machine Learning, including Neural Networks, non-parametric regression
  - Deep Learning, including Keras/TensorFlow running on GPUs

1.1.1.4.2 The course sections

- DSCI35x (x = 1,3,2)
  - Is undergraduate class for “general” applied data science
- DSCI35xM (x=1,2,3) focuses on materials science systems
- DSCI45x (x=1,2,3)
  - Is a graduate level class
  - With the same class material and DSCI35x
  - Additionally the students do a 40 point Semester Data Analysis Project

1.1.1.4.3 The specific courses

- DSCI351, 351M, 451
  - Is an introduction to Exploratory Data Science
- DSCI353, 353M, 453
  - Focuses on Modeling, Prediction and Machine Learning
- DSCI 352, 352M, 452
  - Is a Semester long Data Science Project Class
  - Providing a data analysis for inclusion
  - In your Data Science Portfolio

1.1.1.4.4 Semester Data Science Projects

- Are done in DSCI352, 352M by students who have completed both DSCI351,3
- And by graduate students in DSCI 451, 453 and 452

For DSCI45x students, their Semester Project is developed in DSCI352 class

- With Prof. Laura Bruckman
- During team meetings during Friday Community Hour
  - 12:45 to 1:45 in Olin 303
- And during class office hours
  - Monday/Wednesday 4pm to 5pm in White 540
- There are weekly SemProj updates due each week on progress
- And 3 SemProj Presentations in DSCI35x class

1.1.1.5 Syllabus

1.1.1.6 Operating Systems: Windows, OSX and Linux

Command Line Environments

- Linux: Bash on Linux, or Git Bash on Windows
- Mac OSX: Bash in Terminal
- Windows: Command.com Terminal
- In R: R Console, or Console in RStudio

Item	Linux OS	X Mac Wi	ndows
folder demarcation	/	/	“\” don’t use
directory listing	ls	ls	dir
present work. dir	pwd	pwd	
change directory	cd	cd	cd
drives	root	root	drive letters

Item	Linux OS	X Mac Wi	ndows
NO SPACES in	filenames	spaces	don't work

1.1.1.6.1 Basic/Universal Rules

- No Spaces in Filenames
- Only 1 period in a filename, before file extension
- No other periods
- Only Letters, Underscore (_), and Dashes (-) in Filenames
- In code scripts, use forward slash in all file paths and directorys
- You can use CamelBack or snake_case in variable or file names
  - To make code easier to read.
- Code Style is Rstudio or Google R style
- No use of = for Assignments
- Only use <- as the Assignment Operator in R
  - Rstudio Cheat Sheet says <- is "Alt -" in R code

1.1.1.7 Quick Introduction to R/Rstudio/Git

R is the statistical programming language

Rstudio is the Integrated Development Environment (IDE)

Git is the distributed content versioning system

1.1.1.8 Things you need to do

1.1.1.8.1 Online accounts

- Sign up for our Class Slack with your personal or case.edu email
- Sign up for a bitbucket.org account
  - with your case.edu address

- Sign up for a twitter account,
  - then follow @frenchrh, @hadleywickham, @dataandme, @JennyBryan
  - @minebocek, @juliasilge, @rdpeng, @jtleek, @robjhyndman
  - and others as you want, such as
  - @fchollet, @TensorFlow, @ylecun, @GoogleAI, @egorzakharovdl
- Sign up for a stack overflow account on stack exchange

1.1.1.8.2 Also get ODS VDI access

- You should have access to the following resources
  - Citrix Workspace. After installing the http address is https://myapps.case.edu
- Or CWRU AWS Portal to Citrix Xen Desktop for VDIs
  - A Open Data Science (ODS) VDI

1.1.1.8.3 High performance computing (HPC) resources we will use

- We will use Kaggle.com for Deep Learning with TensorFlow on GPUs
- We will also use CWRU's HPC Data Science Cluster
  - This will get you familiar with working in HPC

Day:Date	Foundation	Practicum	Readings (optional)	Due (optional)
w01a:Tu:1/14/20	Open Data Science	R, Rstudio IDE, Git		(LE0)
w01b:Th:1/16/20	Intro R Markdown	Forking Class Repo	(R4DS-1,2,3)	
w02a:Tu:1/21/20	Statistical Learning	Pred. Analytics	ISLR2	(LE0)
w02b:Th:1/23/20	Data Analytic Style	Tidy Data Manip.	(R4DS-4,5,6)	LE1
w03a:Tu:1/28/20	Lin. Regr.	Pairs Plots	(OIS7)	
w03b:Th:1/30/20	Mult. Lin. Regr.	Test Stats	ISLR3 (R4DS-7,8)	
w04a:Tu:2/4/20	Logistic Regr.	Interaction Terms	ISLR4	**LE1:Due**, LE2
w04b:Th:2/6/20	Classification		(OIS8)	
w05a:Tu:2/11/20*	Resample Cross-Valid.	Cluster Analysis	ISLR5	
w05b:Th:2/13/20	Bootstrap	Steps of Data Analysis	DL1,2 (R4DS9-16)	**LE2:Due**
w06a:Tu:2/18/20	LMS: Subset		ISLR6	LE3
w06b:Th:2/20/20	LMS: Feature Selec.	Coeff. Uncertainties	DL3,4 (R4DS17-21)	
w07a:Tu:2/25/20*	BeyondL: Spline, GAM	**SemProj1-453**	ISLR7	
w07b:Th:2/27/20	MidTerm Review	**SemProj1-3/452**	DL5,6 (R4DS22-25)	**LE3:Due**
w08a:Tu:3/3/20	**MIDTERM EXAM**			
w08b:Th:3/5/20	Dim. Reduc. & PCA		ISLR8 (R4DS26-30)	
Tu:Th:3/9-13/20	**SPRING BREAK**			
w09a:Tu:3/17/20	Regr. Trees	Dec. Trees	ISLR9	LE4
w09b:Th:3/19/20	Bagging, Boosting	How DT Work	ISLR10	
w10a:Tu:3/24/20*	Support Vector Mach.	**SemProj2-453**	ESL11	**LE4:Due**
w10b:Th:3/26/20*	ML Overview, Caret	**SemProj2-3/452**	DLR1	LE5
w11a:Tu:3/31/20*	Neural Networks	MNIST digits	DLR2	
w11b:Th:4/2/20*	NN Topo., Types, Train	ImageNet	DLR3	**LE5:Due**
w12a:Tu:4/7/20*	R-Keras/TensorFlow	CNN w TF	DLR4	LE6
w12b:Th:4/9/20	CNN w/TF	EL Image Sup. ML		
w13a:Tu:4/14/20	CNNs w/small data	DLwR 2.1		**LE6:Due** LE7
w13b:Th:4/16/20	Tboard, TFestimators	pretrained CNNS		
w14a:Tu:4/21/20	R Packaging	**SemProj3-453**		
w14b:Th:4/23/20	Final Exam Review	**SemProj3-3/452**		**LE7:Due**
	FINAL EXAM	**Th. 3/30/2020, 12-3pm**	Nord 356	

Figure 1: Modeling, Prediction and Machine Learning Syllabus

  - Using Linux on GPU compute nodes
  - For Keras/TensorFlow with R
  - For Deep Learning

1.1.1.8.4 Lab Exercises are submitted and graded on Canvas

- Assignment turn in pages will be posted when LE are given out.

1.1.1.8.5 Your Class Git Repo

- My “Professor” Repo is 20s-dsci353-353m-453-prof
  - On bitbucket, you will fork this repo to your own account
  - Each day prior to class, update your fork from my prof. repo

1.1.1.9 Intro to some R: Data Types

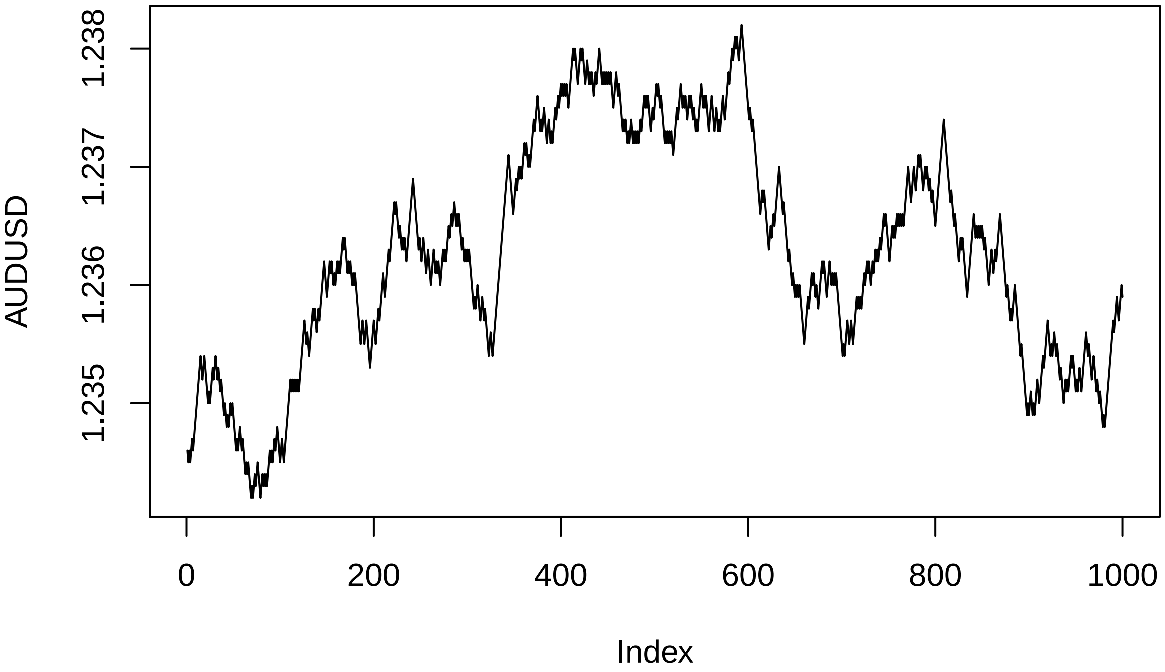
- Primitives (numeric, integer, character, logical, factor)
- Data Frames
- Lists
- Tables
- Arrays
- Environments
- Others (functions, closures, promises..)

1.1.1.9.1 Simple Types

```
x <- 1
class(x)
## [1] "numeric"

y <- "Hello World"
class(y)
## [1] "character"

z <- TRUE
class(z)
## [1] "logical"

as.integer(z)
## [1] 1
```

1.1.1.9.2 Example: Generating Random Data

```
randomWalk <- function(N)(cumsum(ifelse(rbinom(prob = 0.5, size = 1, N) == 0,-1,1)))
AUDUSD <- 1.2345 + randomWalk(1000)*.0001
plot(AUDUSD, type = 'l')
```



1.1.1.10 Recommended R Libraries

We're running R 3.6.2, named "Dark and Stormy Night"

All our "Standard R Packages" are loaded in the ODS VDI

1.1.1.10.1 Basic useful packages (and many more than this)

- Rcpp - Convenient C++ interface
- zoo/xts - Time series libraries
- Matrix - Enhanced matrix library

1.1.1.10.2 Hadley Wickham Tidyverse packages

- This is the content of R for Data Science (R4DS) book.
  - Using Pipes "%>%" to replace loops
  - Makes syntax more compact and readable
  - Makes code faster
- Tidyverse Style Guide
  - Using tidy dataframes
- ggplot2 - Mini-DSL (domain specific language) for data visualization
- plyr/reshape - Data reshaping/manipulation
- dplyr
- data.table - Faster data.frame manipulation
- knitr - for markdown processing
- among others like purrr etc.

1.1.1.10.3 Statistical and Machine Learning

- e1071 Functions for latent class analysis, short time Fourier transform, fuzzy clustering, support vector machines, shortest path computation, bagged clustering, naive Bayes classifier etc (142479 downloads)
- MASS tools for variable selection etc.

- rpart Recursive Partitioning and Regression Trees. (135390)
- igraph A collection of network analysis tools. (122930)
- nnet Feed-forward Neural Networks and Multinomial Log-Linear Models. (108298)
- randomForest Breiman and Cutler's random forests for classification and regression. (105375)
- caret package (short for Classification And REgression Training) is a set of functions that attempt to streamline the process for creating predictive models. (87151)
- kernlab Kernel-based Machine Learning Lab. (62064)
- glmnet Lasso and elastic-net regularized generalized linear models. (56948)
- ROCR Visualizing the performance of scoring classifiers. (51323)
- gbm Generalized Boosted Regression Models. (44760)
- party A Laboratory for Recursive Partitioning. (43290)
- arules Mining Association Rules and Frequent Itemsets. (39654)
- tree Classification and regression trees. (27882)
- klaR Classification and visualization. (27828)
- RWeka R/Weka interface. (26973)
- ipred Improved Predictors. (22358)
- lars Least Angle Regression, Lasso and Forward Stagewise. (19691)
- earth Multivariate Adaptive Regression Spline Models. (15901)
- CORElearn Classification, regression, feature evaluation and ordinal evaluation. (13856)
- mboost Model-Based Boosting. (13078)

1.1.1.11 Links

http://www.r-project.org

Rory Winston, for the Learning R intro http://www.theresearchkitchen.com/archives/1017

R for Data Science http://r4ds.had.co.nz/

- Or pull the R4DS repo from Bitbucket https://bitbucket.org/cwrudsci/r4ds
- Peng-Computing For Data Analysis Playlist

Kaggle, Runs Open Data Science Competitions https://www.kaggle.com/