2001-353-353m-453-07b-p- MidTerm Exam Review.Rmd

Profs: R. H. French, L. S. Bruckman, P. Leu, K. Davis, S. Cirlos

TAs: W. Oltjen, K. Hernandez, M. Li, M. Li, D. Colvin

02 March, 2023

Contents

7.2.1.0.1 Reading, Lab Exercises, SemProjects

- Readings:
  - For today: ISLR9, (R4DS26-30)
  - For next class: ISLR10
- Laboratory Exercises:
  - LE4: Is due Thursday March 9th
  - LE5: Given out next Thursday
- Office Hours: (Class Canvas Calendar for Zoom Link)
  - Wednesdays @ 4:00 PM to 5:00 PM

  - Saturdays @ 3:00 PM to 4:00 PM
  - **Office Hours are on Zoom, and recorded**
- Semester Projects
  - Office Hours for SemProjs: Mondays at 4pm on Zoom
  - DSCI 453 Students Biweekly Updates Due
    * Update # is Due ** **
  - DSCI 453 Students
    * Next Report Out # is Due ** **
  - All DSCI 353/353M/453, E1453/2453 Students:
    * **Peer Grading of Report Out #1 is Due Thursday March 2nd**
  - Exams

* MidTerm: **Thursday March 9th**, in class or remote, 11:30 - 12:45 PM
  · **CWRU Spring Break is March 13th to March 17, so NO CLASS**
* Final: **Thursday May 4th**, 2023, 12:00PM - 3:00PM, Nord 356 or remote

7.2.1.0.2 Textbooks

- Text Books for DSCI353/353M/453

  - R4DS: Wickham: R for Data Science
  - ISLR: Intro to Statistical Learning with R, 2nd Ed.
  - DLwR: Deep Learning with R, Chollet, Allaire,
  - DLGB: Deep Learning, Goodfellow, Bengio, Courville

- Magazine Articles about Deep Learning

  - DL1 to DL13 are "Deep Learning" articles in 3-readings/2-articles/

- Books from DSCI351/351M/451

  - Peng: R Programming for Data Science
  - Peng: Exploratory Data Analysis with R
  - Open Intro Stats, v4
  - R4DS: Wickham: R for Data Science

7.2.1.0.3 Tidyverse Cheatsheets, Functions and Reading Your Code

- Look at the Tidyverse Cheatsheet

  - **Tidyverse For Beginners Cheatsheet**
    * In the Git/20s-dsci353-353m-453-prof/3-readings/3-CheatSheets/ folder
  - **Data Wrangling with dplyr and tidyr Cheatsheet**

  Tidyverse Functions & Conventions

  - The pipe operator `%>%`
  - Use `dplyr::filter()` to subset data row-wise.
  - Use `dplyr::arrange()` to sort the observations in a data frame
  - Use `dplyr::mutate()` to update or create new columns of a data frame
  - Use `dplyr::summarize()` to turn many observations into a single data point
  - Use `dplyr::arrange()` to change the ordering of the rows of a data frame
  - Use `dplyr::select()` to choose variables from a tibble,
    * keeps only variables you mention
  - Use `dplyr::rename()` keeps all the variables and renames variables
    * rename(iris, petal_length = Petal.Length)
  - These can be combined using `dplyr::group_by()`
    * which lets you perform operations "by group".
  - The `%in%` matches conditions provided by a vector using the c() function
  - The **forcats** package has tidyverse functions
    * for factors (categorical variables)
  - The **readr** package has tidyverse functions
    * to read_..., melt_... col_..., parse_... data and objects

Reading Your Code: Whenever you see

- The assignment operator `<-`, think **"gets"**
- The pipe operator, `%>%`, think **"then"**

7.2.1.0.4 Syllabus

Day:Date	Foundation	Practicum	Readings(optional)	Due(optional)
w01a:Tu:1/17/23	Markov Cluster	R, Rstudio IDE, Git		(LE0)
w01b:Th:1/19/23	Stat. Learning, Approach	Bash, Git, Class Repo	ISLR1,2 (R4DS-1-3)	
w02a:Tu:1/24/23	Lin. Regr. Bias-Var.	SemProjs; Regr. Ovrvw	ISLR3,(R4DS-4-6)	(LE0:Due) LE1
w02b:Th:1/26/23	Train/Test, Bias vs. Vari.	Tidyverse Review	DL01 DL02 (R4DS-7,8)	
w02Pr:Fr:1/27/23	**ADD DROP**	**DEADLINE**		**453 Update 1**
w03a:Tu:1/31/23	Logistic Regr. Classif	Pred. Analytics, Regr.	DL03,ISLR4	
w03b:Th:2/2/23	LDA/QDA	ggPlot2, Code Expect.	DL04, DL05	**LE1:Due, LE2**
w03:Sa:2/4/23				**LE1:Due**
w04a:Tu:2/7/23	Resample Cross-Valid.	ggplot	ISLR5	
w04b:Th:2/9/23	DL, ML Overview	Multilevel Mod.	ISLR6 (R4DS9-16)	
w04Pr:Fr:2/10/23				**453 Update 2**
w05a:Tu:2/14/23	Resampling: Bootstrap	Bootstrap Mixed Effects	DL2R1, DL06,07	**LE2:Due, LE3**
w05b:Th:2/16/23	Subset Selec., Shrink.	Dim. Red. PCA	DLwR2	
w05Pr:Fr:2/17/23				**453 Rep. Out 1**
w06a:Tu:2/21/23	ML with NNs	ggplot, clustering	DLwR3	
w06b:Th:2/23/23	Beyond Linear Modls	Feature Select., Caret	ISLR7 (R4DS22-25)	**LE3:Due, LE4**
w06Pr:Fr:2/24/23				**453 Update 3**
w07a:Tu:2/28/23	Dec. Trees, Rand. Forest	Tidy Modeling	ISLR8, DL08,09	
w07b:Th:3/2/23	MidTerm Review, SVM	SVM, SVR, ROC	ISLR9 (R4DS26-30)	**Peer Review 1**
w08a:Tu:3/7/23	ML Overview	, Keras/TF2, Torch	ISLR10.1,10.2	
w08b:Th:3/9/23	**MIDTERM EXAM**		DL10,11	**LE4:Due LE5**
w08Pr:Fr:3/10/23				**453 Update 4**
Tu:3/14/23	**SPRING**	**BREAK**	ISLR10.3,10.4	
Th:3/16/23	**SPRING**	**BREAK**	ISLR10.5,10.6,	
w09a:Tu:3/21/23	Deep Learning	TF2 Keras Intro		ISLR10.7,10.8, DLwR3
w09b:Th:3/23/23	Computer Vision, CNN	CNN w/TF2, Overfit	DLwR4, DL12,13	
w09Pr:Fr:3/24/23				**453 Rep. Out 2**
w10a:Tu:3/28/23	Deep Learn Intro	NN Types	DLwR5 Hinton ImageNet	
w10b:Th:3/30/23	DL CNN,RNN ImageNet	NN Types, CNN wTF2		
w10Pr:Fr:3/31/23				**453 Upd.5 & PrRev 2**
Sa:4/1/23				**LE5:Due LE6**
w11a:Tu:4/4/23	Fitting NNs	AUC,Prec,Recall Fruit		
w11b:Th:4/6/23	NLP, Graphs & ML		LeCun DL Rev. 2015	
w12a:Tu:4/11/23	Graphs & ML	NLP with sequences	DLwR6	
w12b:Th:4/13/23	NLP w attention	Graph Repr Proc Wrk-flw		**LE6:Due LE7**
w13a:Tu:4/18/23	DL Frameworks	Explaining DL w Lime		
w13b:Th:4/20/23	Linux Distros XGBoost	Explain Preds	Deep Dream	
w13Pr:Fr:4/21/23				**453 Rep. Out 3 Due**
w14a:Tu:4/25/23	Tranformers			
w14b:Th:4/27/23	Final Exam Review	Torch NN & DeepLearn		**LE7:Due**
w14Pr:Fr:4/28/23				**Peer Rev 3 Due**
	FINAL EXAM	**Th. 5/4/23, 12-3pm**	Nord 356 & Zoom	
	453 Final PDF Report	**Fr. 4/29, 11:59pm**		

Table 1: DSCI353-353M-453 Weekly Syllabus. R4DS-x.y, OISx.y, ISLRx.y, DLwRx.y, DLGBx.y refers to chapters and sections assigned as reading in our textbooks. DLx are deep learning articles.

Figure 1: DSCI351-351M-451 Syllabus

7.2.1.1 Care and feeding of your files and your repository Before next week's Thursday in-class MidTerm Exam

- Check `git status` of you class repository on Markov/OnDemand or computer
- Make sure you can `git pull`
- And that you can `git push`
- And `git add --all :/` to stage your local changed files to the commit list
- Then `git commit -m 'updating my repo'`
- Then `git push`

Its essential to have an up-to-date copy of your repo in bitbucket

If you want to make backup copies of any files

- For safety for example
  - Copy them to a new folder in your `/home/caseID` directory
  - Copy them to your google drive
  - Copy them to your local computer

Also you can do the midterm on either

- Markov Data Science Cluster
  - Using an Rstudio Server (rxf131) {The 22.04 version}
  - via (http://ondemand.case.edu)
- Or on ODS Win10 Desktops
  - Using an Rstudio on the ODS Desktop
  - via (http://myapps.case.edu)

So go and login to both of these

- And setup your Git server
  - Check your Git Server Configuration
    * `git config --list`
  - Its essential to initially configure your git server
    * `git config --global user.name "[name]"`
    * `git config --global user.email "[email address]"`
- And clone your personal class repository
  - to both Markov and ODS Desktop
  - `git clone https://vuvlab@bitbucket.org/vuvlab/22s-dsci353-353m-453-e1453-e2453-CaseID.git`
  - But with your CaseID in the command

7.2.1.2 DSCI353-353m-453 MidTerm exam overview   You can use the resources in your repo,

And the resources in R and Rstudio

- such as the help system

Turn in your .Rmd file and the compiled pdf

- to the canvas midterm exam assignment page

7.2.1.3 You'll want to review

- the class notes,
- and more particularly
  - the .Rmd/pdf class notes on concepts and ideas.
- and the readings in textbooks
  - R3DS
  - ISLR2
  - DLwR2

- And the Deep Learning articles

You'll also want to be familiar with

- ISLR Chapters 1, 2, 3, 4, 5, 6 and 7
- R4DS Chapters 1-25

And ISLR chapters 8, 9, 10, 12 aren't covered on the midterm exam

- ISLR 8, Tree-based Methods
    - Decision Trees and Random Forest Machine Learning
- ISLR 9, Support Vector Machines (SVM)
- ISLR 10, Deep Learning
- ISLR 12, Unsupervised Learning

And if you want you can review

- Regression and Classification
- in Open Intro Stats v4
    - OIS Chapters 8 and 9
    - On linear regression and multiple regression

7.2.1.4 The Practicum problems on the midterm

- are using the `Boston` dataset
    - which is part of the `MASS` package.

It records the median value of houses

- for 506 neighborhoods around Boston.

This dataset is also used in ISLR Chapter 3,

- section 3.6 Lab Linear and Multiple Regression.

- Sections 3.6.1 to 3.6.7, Pages 109 - 119.

7.2.1.5 Use good code style

- **Comment your code**
- Use `ggplot2` for plots
- **Make plots with proper axis labels and titles**
- Not just variable names for axis labels
- **Use tidyverse commands, and pipes in your code**

7.2.1.6 You have 1 hour and 15 minutes

- The midterm is worth 10 points.

- Points for each question are as follows
    - 1. 3 pts
    - 2. 3 pts
    - 3. 4 pts

7.2.1.7 Good code style will count in the grading

- Good code style
    - Commenting your code to make it clear
    - `<-` assignment operator
    - spaces around operators

- – spaces after commas
  - – indentation of multiline commands
  - – indentation of loops and functions
- Using `ggplot2`
- Using `tidyverse` packages and `dplyr` pipes
  - – as appropriate

7.2.1.8 Problems on the MidTerm

7.2.1.8.1 Problem 1: Foundation Questions (3 pts)

7.2.1.8.2 Problem 2: Practicum- Linear Regression (3 pts)

7.2.1.8.3 Problem 3: Practicum- Multiple Regression (4 pts)

7.2.1.8.4 Cites

- G. James, D. Witten, T. Hastie, and R. Tibshirani, An Introduction to Statistical Learning: 2nd Ed., with Applications in R, 2nd ed. 2021 edition. New York: Springer, 2021.