More on Classification & Clustering.. (Explore and work through the given R-scripts)

Thilanka Munasinghe

Data Analytics

ITWS-4600/ITWS-6600/MATP-4450/CSCI-4960

Group 3 Lab 5, October 11, 2022

- During the today's Lab, you are asked to go over and explore on the R scripts that are given in the course repository (group2 on RPI Box: https://rpi.box.com/s/2xx9uI1fmc6bf5ff8h4jreae69emikmf
- **What is expected**: You are asked to Explore, Inspect the code/scripts in the Rstudio environment and get familiar with those scripts.

- As you are working on those given scripts, and your goal is to understand the R functions that are available in those scripts by using the help() function in RStudio and searching them on the web.

Plot tools/ tips

Make sure to read these articles:

Combining Plots:

http://statmethods.net/advgraphs/layout.html

How to Read and Use Histograms in R:

http://flowingdata.com/2014/02/27/how-to-read-histograms-and-use-them-in-r

More script fragments in R available on the web site:

https://rpi.box.com/s/2xx9ul1fmc6bf5ff8h4jreae69emikmf

Today on web under group2/

- lab2_abalone.R lab2_kknn1.R
  lab2_nbayes1.R lab2_nbayes2.R
  lab2_nbayes3.R lab2_nbayes4.R
  lab2_swiss.R
  lab3 ctree1.R
  lab3 ctree2.R
  lab3 ctree3.R

Scripts – work through these

See in folder group2/

lab1_kmeans1.R

lab1_nyt.R

lab1_bronx1.R

lab1_bronx2.R

lab1_pairs1.R

lab1_splom.R

lab1_gpairs1.R

lab1_mosaic.R

lab1_spm.R

lab1_wknn.R

lab1_kknn1.R

lab1_kknn2.R

lab1_kknn3.R

Do over…

- **Make sure that you go over the lab1_nyt.R (Good Practice for Assignment3)!**
- Make sure that you get to the “bronx” dataset and group2/lab1_bronx1.R and lab1_bronx2.R **explore the script fragments**
- You need it for the up coming A4!!
- And group2/lab1*.R scripts – work through them all

Reading Assignments:

Reading Assignment: Read prior to next week's class.

Some reading material are related to upcoming labs/lectures:

You need to understand the confusion matrix and contingency tables in order to interpret the results properly in upcoming labs and lectures **(we will cover these topics during next week(s) classes/labs, but first, ream these articles!).**

- **Chapter 8: Class Textbook: Introduction to Statistical Learning with R (ISLR)**
- https://en.wikipedia.org/wiki/Confusion_matrix
- https://www.dataschool.io/simple-guide-to-confusion-matrix-terminology/
- https://towardsdatascience.com/understanding-confusion-matrix-a9ad42dcfd62
- https://en.wikipedia.org/wiki/Contingency_table
- http://mathworld.wolfram.com/ContingencyTable.html

We will cover these topics next week…

Reading Assignment:

- **Creating a confusion matrix using *cvms* package in R.**
- **Install these Libraries: cvms , tibble**

```
library(cvms)
library(tibble)
```

- During today's Lab: Work on the examples listed in this reading assignment.
- https://cran.r-project.org/web/packages/cvms/vignettes/Creating_a_confusion_matrix.html

We will cover these topics next week…

Reading Assignment:
Understand the Confusion Matrix

- In the field of machine learning and specifically the problem of statistical classification, a confusion matrix, also known as an error matrix

Read: https://en.wikipedia.org/wiki/Confusion_matrix

		Actual Values	
		Positive (1)	Negative (0)
Predicted Values	Positive (1)	TP	FP
	Negative (0)	FN	TN

Image/Photo Credit: https://towardsdatascience.com/understanding-confusion-matrix-a9ad42dcfd62

We will cover these topics next week…

Model Performance

- Metrics to evaluate the model performance

Accuracy Rate:

$$\text{Accuracy Rate} = \frac{Number\ of\ Correct\ Predictions}{Number\ of\ Total\ Predictions}$$

$$\text{Accuracy Rate} = \frac{TP+TN}{TP+TN+FP+FN}$$

True Positive (TP)
True Negative (TN)
False Positive (FP)
False Negative (FN)

Confusion matrix

Read: https://en.wikipedia.org/wiki/Confusion_matrix

We will cover these topics next week…

Model Performance

Error Rate :

$$\text{Error Rate} = \frac{Number\ of\ Incorrect\ Predictions}{Number\ of\ Total\ Prediction}$$

$$\text{Error Rate} = \frac{FN+FP}{TP+TN+FP+FN}$$

$$\text{Error Rate} = 1 - Accuracy\ Rate$$

True Positive (TP)
True Negative (TN)
False Positive (FP)
False Negative (FN)

We will cover these topics next week…

Reading Assignment:

- Go over this article:
- https://www.rdocumentation.org/packages/caret/versions/3.45/topics/confusionMatrix

- Metrics to evaluate the model performance

Accuracy Rate:

$$\text{Accuracy Rate} = \frac{Number\ of\ Correct\ Predictions}{Number\ of\ Total\ Predictions}$$

$$\text{Accuracy Rate} = \frac{TP+TN}{TP+TN+FP+FN}$$

True Positive (TP)
True Negative (TN)
False Positive (FP)
False Negative (FN)

Confusion matrix

Read: https://en.wikipedia.org/wiki/Confusion_matrix

Project Datasets check !

1) Make sure to check-in with the instructor or TA about your **project proposal dataset**.

(By now you should have selected a project or a plan to choose a dataset, if not speak to me before end of the Friday…)

2) Show your "dataset" that you are planning work with, we need to document it so that we know if there are multiple students working on the same dataset. (**This is important! Make sure to check your project dataset before end of the Friday's class**)

Update your GitHub

- Push your Lab codes to your GitHub before end of today's lab (make sure to push your codes that you completed during the Labs before you leave the classroom)
- You might not complete all of them during today's lab but, push to GitHub what you completed during the class).
- **Push your Lab 3, Lab 4 and Lab 5 codes to GitHub**, TA and I will be checking your GitHub repo during this week.

Assignments to come

Reminders:

- Assignment 3: Due: October 14th , 2022 (by 11:59pm EST) Submission method: written document posted on LMS Assignment. Please use the following file naming for electronic submission: DataAnalytics_A3_YOURFIRSTNAME_YOURLASTNAME.xxx
- Assignment 5: presentations will be next week
- Assignment 5 Due: Monday, October 18th, 2022 (Presentation slides due at the beginning of the class) Presentation method: Presentations during the class time
- **Slides MUST be submitted to LMS at the beginning of the class on October 18th, 2022.**