frrnq8--bmg,mpe-/.,26..7-0 ggq 0./6 43 5/

Issues in Information Systems
Volume 19, Issue 2, pp. 65-71, 2018

TEACH MBA DATA SCIENCE USING R

Jianfeng Wang, Indiana University of Pennsylvania, jwang@iup.edu
Linwu Gu, Indiana University of Pennsylvania, lgu@iup.edu

ABSTRACT

This is a pedagogical study of teaching data science classes in our business school. Given challenges there, we find some effective ways of mixing R programing with excel-based tools in teaching this course. The use of textbook, data sets and instructional materials is discussed. So is the arrangement of the course contents. We believe this paper will help colleagues from other universities in teaching such a class or proposing a similar course.

Keywords: Data Science, R, Excel, Algorithms.

INTRODUCTION

Data science has been very technological and a tough topic to teach at business schools. There are discussions of challenges in teaching data science courses (Wang and Gu, 2016; Yap and Drye, 2018). We have offered this class in our business school for a few semesters. Our experience may be helpful to other professors in teaching data science or related topics for MBA students, who usually have no strong background in statistics and programming. It is a real challenge to teach MBA students who want to learn about data science but don't have strong background in statistics and programming.

In this paper, we will discuss how we teach this class. We first introduce the outline of our course design, then we discuss the tools we use in the class: Microsoft PowerBI, R project, and R console and studio. At the end, we discuss the textbooks and course arrangement.

COURSE DESIGN

Data science is an interdisciplinary area in which useful insights are retrieved from data mining on different kinds of data sets such as numeric or textual, structured or unstructured (Provost and Fawcett, 2013; O'Neil and Schutt, 2014; Wang and Gu, 2016). Structured data sets are often ready for analysis. Unstructured data sets must be munged before they can be analyzed. Unstructured data sets must be processed and reformatted such that they can be read by machines. In machine learning, data analysis is a bit like blackbox, a bit hard for business students to understand. Successful completion of such a class requires that a student has backgrounds in linear algebra, statistics, programming, SQL, and other computer related areas (Provost and Fawcett, 2013). But when students come to a data science class, they often don't have strong backgrounds in so many areas. Though they all have backgrounds in basic statistics, basic statistics is not enough for a little bit high level discussion of statistical programming using R or Python. What we do first is to teach our students excel-based data processing and analysis, and a little bit review of statistics. Most steps of data formatting and processing in Excel are viewable in Windows and easier to understand. After a few weeks of discussion and practice using Excel-based tools and review of basic statistics, we introduce a bit more technical discussion of data mining algorithms and tools using R. The pace is very important. Too fast or too slow may make some students bored or decide to quit. It is an art to balance the materials and control the pace while introducing these challenging topics.

MICROSOFT POWERBI

Microsoft has a rich set of tools for data analysis with Excel as a platform. With Excel Pivot Table, there are PowerPivot, PowerMap, PowerView, and PowerQuery (Clark, 2014). PowerQuery can function as interface between Excel and Database Servers and Other tools under Microsoft Azure Cloud. PowerBI is a quite useful toolset for

students to learn. Through the practice of Excel and PowerBI, students are trained to be familiar with the rich functions available in Excel and PowerBI. Many students only know basic functions of Excel through their computer literacy courses. Most students don't know that Excel can be used to train data, to handle missing data, to find outliers in the data sets, and can also be used to handle text data. PivotTable is great to do some data aggregation and trend and multidimensional analysis. In Excel worksheet or PowerPivot, data is usually organized column by column, which is similar to large scale data processing by R, Python or Spark (Matloff, 2011; White, 2013). The experience of working with Excel can help students develop some concepts, experience and interests in data analysis, which can be helpful when R or Python is introduced for much larger data processing. The limit of using Excel for data processing lies in the size of a data set that can be analyzed at each time. The data set to be processed by Excel should not be more than the memory available in the computer (Clark, 2014). In other words, Excel or Excel-based tools are very much using one computer, not using parallel computing on multiple computers. Excel Add-ins from Microsoft can be downloaded for free or come as a part of Office package. Excel Add-ins from third party vendors such as solver.com will charge a license fee.
Other than regular regression and statistics functions, what is available for data analysis in Excel itself is quite limited (CarlBerg, 2016). Even the data mining add-in designed for SQL Server is quite limited in data mining functions (Ferrari, 2012). Data mining is not convenient for classroom teaching because you must set each student account with some DBA privileges.

Though excel skills are very helpful and important and highly sought after in the job markets, tools based on R or Python are pivotal for larger data processing. In our class, we introduce R after we discuss Excel and PowerBI tools.

R-PROJECT

R is a statistical programming language. It is first used for academic purpose and teaching (Matloff, 2011). For the past few years, the commercial use of R has gained momentum. R has been ranked as a top choice for data processing and analytics and visualization for a few years by the knowledge discovery news site kdnuggets.com. Microsoft has acquired revolutionary R and tried to integrate R into Microsoft software offers.
R is an integrated suite of software facilities for data manipulation, calculation and visualization. Functions in Excel are all available in R but in different ways. R is pretty good in data training and munging. R is designed to process data types such as vectors, matrices, lists and data frames. The data frame in R is compatible with pair-wise column-based data structures, which can be handled by big data processing systems such as Hadoop and Spark. R script can run on top of Spark to process very large data sets (Zaharia et al., 2016).

R is available as free software under the terms of the Free Software Foundation's GNU General Public License in source code form. It compiles and runs on a wide variety of UNIX platforms and similar systems (including FreeBSD and Linux), Windows and MacOS. There are about 12465 R packages contributed and maintained by developers from all over the world (check for updates at https://cran.r-project.org/web/packages/). Users can simply download such packages when they need to use any such packages. The CRAN, the comprehensive R archive network, is available at many web servers all over the world, from which users can download R packages needed for their programs (Matloff, 2011). There are R packages, which provide methods for data mining and machine learning such as topicmodel, wordcloud, and TM for text mining; c50 and rweka for decision tree analysis; e1071 for naïve Bayes algorithm; fpc and cluster for clustering analysis; forecast and dtw for time series; Rlof for outlier detecting; igraph, d3network, and RNeo4j for social network and graph mining; RgoogleMaps for spatial data; ff, ffbase and filehash for very large data storage; ggplots for graphics; arules for association analysis, and many others. For the details of these packages, please refer to CRAN project documents at cran.r-project.org. Users can also visit for r-bloggers.com for more example-based discussions.

There are some very helpful websites dedicated on discussing and sharing R scripts such as r-bloggers.com, rdatamining.com, etc.

R CONSOLE AND R STUDIO

R console, is an R interpreter and environment that can process data sets. R console is free to download from www.r-project.org. R studio non-commercial free version is downloadable from www.r-studio.com. R-studio is a tool that integrate R code editor, R console, R environment logger, and file browser. R-studio is based on R console. To use R-studio to run r script, there must be R console installed first. Both R console and R-studio are very easy to install and configure, great tools for teaching and learning R coding. R console comes with some base installation including class and other packages that can do basic calculation, statistics and regression analysis (Matloff 2011). Most data mining algorithms come from packages not available in the base installation (Lantz, 2013). Users simply need to load a package from CRAN mirror network and make the package library available in R session. Figure 1 shows some packages for data mining processing such as c50, cluster, e1071 etc. They were installed in our computers. To use functions from e1071, just click the checkbox next to the name. All the functions from the package library will be ready to run and process users' data sets (Teeter, 2011; Matloff, 2011; Lantz, 2013). If the packages are not downloaded and installed, users cannot see the package names in the list. Through R studio, just click "Install" to install a known package. To update an installed package, just click "Update".

If the package is not in the list, click "INSTALL" to download the package from CRAN mirror sites. When one installs a package, the other pre-required packages may also need to be installed.

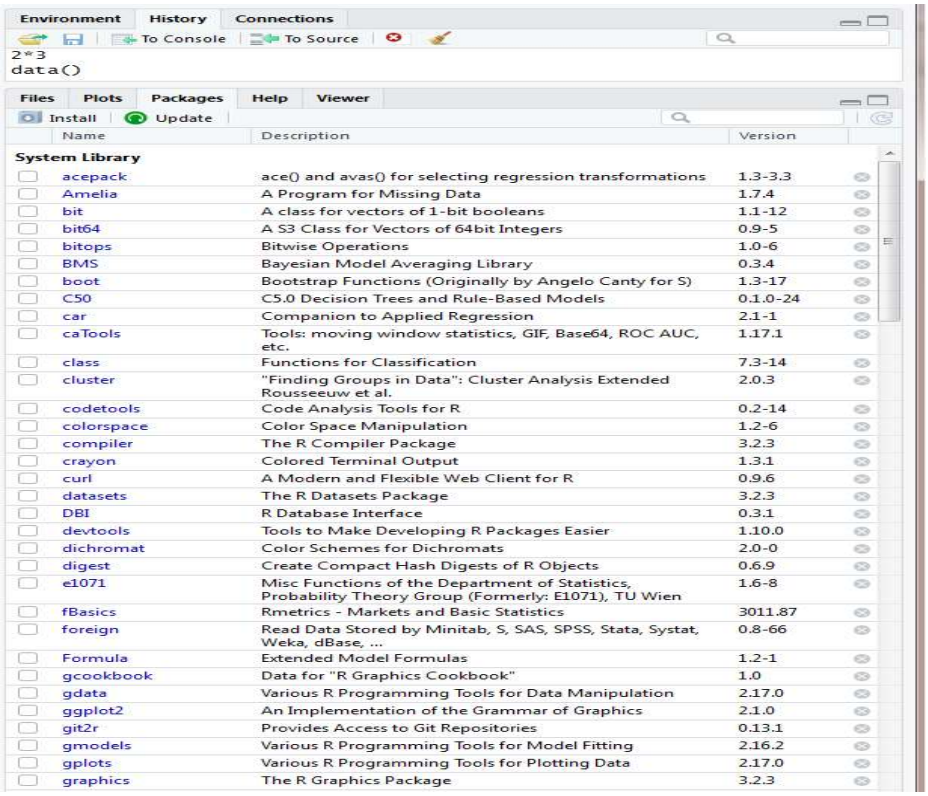

Figure 1. R Package Examples

R console is session-based. When a user opens R console or R studio, he starts an R session. Any data sets and packages loaded to the current session will just be available for the session. If he closes the session and exit, the data sets and packages will be removed from the prime memory. When next time the user needs to run the same data sets and use the same packages, he must load the data sets and packages again. But these setups can be done easily and quickly.

WHICH ALGORITHMS TO COVER AND WHAT DATA SETS TO USE

There are many data mining algorithms, which can be covered in a semester. Some algorithms are a little hard for MBA students to understand, such as naïve Bayes, logistic regression, neural network. Others are easier to understand, such as decision tree classification and prediction, k-means, and k-nearest neighboring, etc. There are variations with these algorithms, which are programmed to fit more complicated analysis. Such enhanced versions are usually more challenging to understand. Overall data mining or machine learning algorithms present a complex framework of analysis only students with enough statistics and programming backgrounds and willing to work hard can walk through. Even if a professor just tries to discuss these concepts and models in a superficial way, low-GPA students will have trouble understanding.
Another important issue is what size of a data set is good for classroom use. University of California Irvine provide a portal at https://archive.ics.uci.edu/ml/datasets.html, with a long list of data sets commonly used for machine learning and data science practice. These data sets are all small-sized and pretty fit for classroom use. The R base installation also carries many small data sets, which can be used easily (Figure 2). We use different data sets to introduce different algorithms, such as credit risk data for k nearest neighboring, iris data for k-means, mushroom data set for decision tree classification, email text files for text mining. For brief introductions to a short list of popular algorithms in data mining, please refer to Provost and Fawcett (2013); for a short list of R packages carrying these algorithms, please refer to Wang and Gu (2016) and Lantz (2013). Analyze a data set from a specific domain is a case study. In real data analysis, domain knowledge is very important (Provost and Fawcett, 2013).

There are hundreds of data sets available from www.kaggle.com. Many companies post their data sets at kaggle.com and invite developers to write competing code to analyze their data sets with awards. Such data sets can often be downloaded and assigned to students for projects.

Are bigger data sets good for classroom discussion? That depends on lab computer capacities available at different colleges. At our business school, the lab our department uses is equipped with computers using the latest generation of Intel microprocessors and 32G prime memory. Such configuration can handle hundreds of megabytes of data easily. A few years ago, the computers we used only had 12G memory each and older chips, which made processing 200MG data a freezing task. For classroom practice, bigger the data set, more problems there may be. If a computer can be connected with a data center to use cloud sources, there may be some other unexpected problems in networking.

But R can be programmed to process very large data sets. On top of Spark, a core big data platform, R code is runnable to handle very large data sets, among Scala and Python. In an introductory class, it makes sense to use R as R can be used to process very large data sets too (Zaharia et al., 2016).

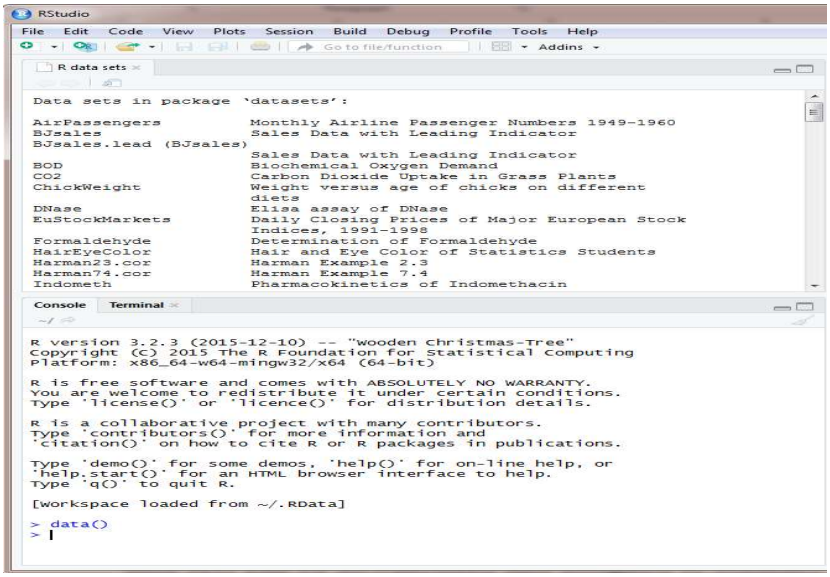

Figure 2. Data sets with R base installation

R is such a well-designed statistical programming language that to simply use R for data science and business intelligence does not require a user to be an advanced R programmer. With a minimal level of knowledge of R syntax and types of variables, a user can call functions to read data sets to R sessions and analyze data (Matloff, 2011; Lantz, 2013; Teetor, 2011). A user certainly has to have statistics background in order to interpret the results from R. If a beginner plan to use python, a general programming language, he will need to learn data types, loops, data structures and, then Numpy and Pandora, and then other advanced packages for data mining algorithms. If he plan to learn R coding, he can start with learning vector, matrix, list, and data frame syntax and indexing. He can then quickly move to learn some data mining packages and their main functions. The R learning curve is a bit steep but short (Matloff 2011). To use main functions in data mining packages, users don't have to develop a detailed understanding of how the function is written or programmed.

Because of this convenience, R has been quite popular in academic circles and often selected for teaching in statistical programing and data science classes (Hicks and Irizary, 2016).

WHICH TEXTBOOKS TO USE

Most of our MBA students don't have strong backgrounds in statistics or programming. That may be the case in similar universities. So, data science textbooks used by or written by professors in computer science or statistics professors are not fit for our MBA students or senior business undergraduates. Shmueli et al. (2018) is a good textbook for students with strong backgrounds in statistics and programming. Many instructors use Data Science for Business by Provost and Fawcett (2013). Provost and Fawcett (2013) is a nice book with minimal mathematical discussion and no coding at all. Most discussions in the book stay at conceptual level. Their book is quite fit for MBA students with some background in statistics but without any background in programming. Provost and Fawcett (2013) provide a framework of general analytical approaches in data modeling, introduce algorithms for clustering, classification, association, regression, and prediction with minimal math, and offer insights on how to support business strategies using data-driven techniques. We use this book as a general reference of data science for business analysis. But we don't think this book is enough for some ambitious students, who would like to develop their skills in R coding and practice data analytics using R, or would like to choose data analytics as their career.
As we mentioned in the early section of the course design, we first introduce Excel-based PowerBI, then discuss minimal R coding so that students can understand syntax and indexing of R data types and structures, before we discuss R packages for a selective list of data mining algorithms. We prepared many worksheets and materials ourselves. We often discuss more technical materials in the second half of a semester for such a data science class.

ARRANGEMENT OF THE COURSE CONTENTS

Most of the MBA students in our MBA program don't have strong backgrounds in statistics or programming. But many of them truly wanted to learn data science and data analytics skills. We found it very helpful to introduce Excel-based tools first in a class. At Excel, every step for data analysis and reformatting is viewable through Windows. A hover over an object in a worksheet and right-click on the mouse may provide some hints of what to do next. Mini toolbar and hints from Excel are very helpful even in data analysis. Many Excel functions and features for data formatting and processing can find their counterparts in R, but much easier to learn. For example, we can use such functions as if(), isnull(), isnumber(), istext() in excel to handle missing data and blank cells or reformate cell values. Very similar functions available in R. Excel PivotTable functions are very similar to tapply() functions in R, which is designed to handle multi-dimensional aggregation in R. Functions such as tapply() can take a list or data frame as an argument and use specific functions and a list of factors to provide multidimensional aggregation analysis (Matloff, 2011). But it is easier to understand PivotTable than tapply() function. The practice of PivotTable can certainly students better understand tapply() function in R.

PowerView and PowerMap are useful tools for visualization. PowerView provide a dashboard or scorecard interface by which users can understand a data analysis report using map and charts. PowerMap can provide a geographic distribution of aggregate data trends. PowerPivot allows loading of larger data sets from database servers and other sources to be analyzed using PivotTable functions available in Excel (Ferrari, 2013). PowerQuery, now part of Data Menu in Excel, is an interface to load data from database servers and other sources. PowerQuery allows users to edit

queried data sets and later load to Excel worksheet for further analysis and integration into a data model when needed (Clark, 2014). Both PivotTable and PowerPivot use Microsoft DAX and MDX libraries of functions to provide data analysis for business intelligence.

Introduction of Excel-based tools such as these mentioned above may take 3 to 4 weeks. It can help students to learn basic data analysis in a real hands-on way and set up a foundation for them to learn further about data mining using R, which is mostly a blackbox process. We truly think learning Excel-based tools help students a lot. Lots of public data sets available through www.data.gov can be downloaded for practice. We generated lots of materials and worksheets using public data sets and other data sets we used before.

Next in the sequence is an introduction of syntax and indexing of R data structures and basic functions. R is unique where R's basic type of variables is vector. There is no scalar variable in R. There are vector, matrix, list and dataframe, all with similar indexing syntax but it will take some time for students to get used to and remember such syntax. The syntax is very flexible, which makes R ideal for data analysis and statistical programming. The visualization tools through R ggplot package and others are very easy to use, making R very rich in tools for data visualization (Teeter, 2011). Introduction of R vector, matrix, list and data frame may take about 2 to 3 weeks.

Once students understand how to manipulate R list and dataframe and able to call functions, they can learn packages for data science algorithms. We introduce linear regression, logistic regression, k-means, K-nearest neighboring, decision tree classification and prediction, naïve Bayes, text mining, and neural network. These are most popular algorithms. Data Science for Business (Provost and Fawcett, 2013) provides decent discussion about data modeling and pros and cons of different data mining algorithms. R scripts for linear regression, logistic regression, k-means, K-nearest neighboring, decision tree classification and prediction, naïve Bayes, text mining and neural network are available from different sources with different data sets analyzed. Both linear regression and logistics regression are easier to handle in the classroom as packages for these two usually won't change much over time. When an instructor introduce relatively new R packages, he should be careful with possible changes and updates in the packages carrying data mining algorithms. Younger R packages may experience more updates as developers try to perfect the coding and fix bugs. If there is any update in a package, then there may be minor changes in how to call those functions. As a result, old available scripts may not work completely. So instructors have to rewrite the R code themselves. It is not going to be difficult as they can usually find help or hints from online sources such as R-bloogers.com. But instructors indeed should check their example scripts if they want to run the scripts in a classroom setting. The introduction of these algorithms may take about 6-7 weeks.

CONCLUSIONS

This paper summarizes our experiences and thoughts in teaching data science classes for our MBA students in a university as ours, Indiana University of Pennsylvania (IUP). IUP is AACSB accredited and has been selected in Princeton Review for Best Business Schools since its first edition. Just as those in many similar universities in the USA, most of our MBA students lack strong backgrounds in statistics and programming. We believe that the way we teach this class and our arrangement of the course contents can help colleagues from similar schools in teaching data science or related topics using R. If MBA students in a class have pretty strong backgrounds in statistics and programming, then the arrangement of the course contents will be very different.

REFERENCES

Baumer, B, (2015). A Data Science Course for Undergraduates: Thinking with Data, *American Statistician*, *69*(4), 334-342

Basesens, B., Bapna, R., Marsden, J., Vanthienen, J., & Zhao, L. (2016). Transformation Issues of Big Data and Analytics in Networked Business. *MIS Quarterly*, *40*(4), 807-818.

CarlBerg, C. (2016). Regression Analysis Microsoft Excel. Que Publishing PTG.

Chang, W. (2012). R Graphics Cookbook. O'Reilly, Sebastopol, CA.

Clark, D. (2014). Beginning PowerBI with Excel 2013. Apress, New York.

Dremel, C., Wulf, J., Herterich, M., Waizmann, J., & Brenner W. (2017). How AUDI AG established Big Data Analytics in its digital Transformation. *MIS Quarterly Executive*, *16*(2), 81-100.

Ferrari, A. (2012) Microsoft SQL Server 2012 Analysis Services. Microsoft Press PTG.

Hicks, S. & Irizary R. (2016) A guide to teaching data science, working paper, retrieved from https://arxiv.org/ftp/arxiv/papers/1612/1612.07140.pdf on Apr 21, 2018.

Lantz, B. (2013). Machine Learning with R. Packt Publishing, Birmingham-Mumbai

Matloff, N. (2011). The Art of R Programming, No Starch Press, San Francisco, CA.

Provost, F., & Fawcett, T. (2013). Data Science for Business. O'Reilly.

Shmueli, G., Bruce, P., Yahav, I., Patel, N., & Lichtendahl, K. (2018). Data mining for Business Analytics: Concepts, Techniques, and Applications in R. John Wiley& Sin Inc.

Teetor, P. (2011). R Cookbook, O'Reillly, Sebastopol, CA.

Wang, J. & Gu, L. (2016). Challenges of Teaching Data Science in a Business School. *Issues in Information Systems*, *17*(3), 209-217.

White, Tom. (2013) Hadoop: The Definitive Guide, O'Reilly.

Yap A., & Drye S. (2018). The Challenges of Teaching Business Analytics: Finding Real Big Data for Business Students. *Information System Education Journal, 16*(1), 41-50.

Zaharia, M., et al. (2016). Apache Spark: A Unified Engine for Big Data Processing. *Communication of ACM*, *59*(11), 56-65.