Book Review

Reviewed by Zhongxian Wang
and Sylvain Jaume

Email: wangz@mail.montclair.edu
Email: sjaume@saintpeters.edu

Machine Learning and Data Science: An Introduction to Statistical Learning Methods with R (1st edition)
by Daniel D. Gutierrez
Published 14 September, 2015
by Technics Publications, LLC, 2 Lindsley Rd.,
Basking Ridge, NJ 07920, USA, 320pp
ISBN: 978-1-63462-096-3

Machine learning (ML) is a core component of data mining. ML explores the study and construction of algorithms that can learn from and make predictions on data.

Data Science is an essential skill for analysing and deriving useful insights from data, big and small (Lantz, 2013; Chiu, 2015).

This book is a practitioner's extremely valuable toolbox in machine learning and data science. It provides data scientists with the tools and techniques required to stand out in the data chain management: from data access, data munging, data analysis to both supervised and unsupervised machine learning, as well as model evaluation.

In Chapter 1, the author briefly overviews both data science (DS) and machine learning (ML). The reasons why DS is becoming ever demanding in the business world and how ML plays an integral role in DS have been demonstrated. The two types of ML, supervised learning and unsupervised learning, have been summarised. The author analyses the basic steps involved in the process of DS and discusses the power of R packages in doing ML. The first step in building a ML solution is to explore ways to access datasets in a variety of common formats. A general outline of the ML process has been drawn.

Data access is the theme of Chapter 2. The author regards data sources as the lifeblood of the ML algorithms. Data for ML comes in all shapes and sizes including CSV (comma separated values), Excel, JSON (*JavaScript Object Notation*), HTML pages, SQL databases, Twitter, Google Analytics, etc. Unstructured data has become popular data sources, which is free form and does not fall into the usual tabular format that typical corporate transactional data does. The author provides a useful data access toolset that a data scientist can reuse for subsequent ML projects.

In Chapter 3, the author emphasises the importance of 'Data Munging'. It involves the transformation of a dataset into a form more suitable for ML algorithms. Data munging is both a tedious and rewarding step in the ML process that often takes up to 80% of the time and cost involved in the overall project. Without this prerequisite, ML would meet obstacles, if it is not possible. Depending on the dataset and the

requirements underlying the ML problem, many data munging operations must be performed on the data, such as revise variable names, create new variables, discretise numeric values, handle data, binarise categorical variables, merge/order/reshape datasets, manipulate data using *dplyr*, handle missing data, and scale the features. Furthermore, the author introduces feature engineering, data sampling, and the data pipeline, as well as principal component analysis (PCA) for dimensionality reduction.

Performing *exploratory data analysis* (EDA) is the next step. In order to get familiar with the dataset, the author introduces many features of the R statistical environment that support this effort: numeric summaries, aggregations, distributions, densities review of factor variables, application of general statistical methods, exploratory plots, expository plots, and much more. The author outlines a cookbook for techniques in EDA in Chapter 4. A data scientist may need to revisit one of more data munging tasks in order to refine or transform the data even further. Another side benefit of EDA is to double-check and refine the selection of features that will be used later for ML. A data scientist may also need to revisit the feature engineering step to make an adjustment.

Chapter 5 covers regression, a method invented 200 year ago by the two independent mathematicians Carl Friedrich Gauss and Adrien-Marie Legendre. Regression, a best understood and widely used statistical tool for investigating the relationship between variables, is now at the centre of modern statistics and data science. Supervised learning is the most common type of ML, often associated with predictive analytics. The most common algorithms, such as simple linear regression, multiple linear regression and polynomial regression, are provided. Shrinkage methods like *ridge regression* and the *lasso*, which are designed to reduce variance by shrinking the regression coefficients towards zero, are new forms of regression.

The author introduces classification, another common form of supervised ML in Chapter 6. Classification is perhaps the most noticeable statistical learning technique because it is used for numerous problem domains such as churn prediction, spam and fraud detection, etc. The response variable is qualitative (or categorical) here, rather than quantitative as in the previous chapter. A data scientist can dramatically expand his or her ML toolbox by adding the following classification algorithms: logistic regression; classification trees; K-nearest neighbours; support vector machines (SVM); neural networks; Naïve Bayes; random forests; gradient boosting trees. A classifier, which is built on a training set, should perform well on the test set containing observations not used to train the classifier.

Chapter 7 evaluates model performance. The author explores the process of evaluating the results delivered by a ML algorithm and offers various techniques for obtaining better model performance. Instead of fit measure (R^2, adjusted R^2, RMSE, etc.), the predictability of a particular model based on new data is the real concern of decision makers. Evidence is born for supporting a theory when predictions are confirmed, and vice versa. You may need to iterate the previous steps: data munging, feature selection, EDA and statistical analysis, and even model selection by checking your first step from the beginning. Actually, data science is a never-ending process in the real world. A data scientist needs to frequently reevaluate the chosen models with newly available data, and accordingly retrain the algorithms to enhance predictive power.

The author brings together unsupervised learning with two clustering techniques in Chapter 8: hierarchical clustering and K-means clustering. This new methodology shows abundant capability for discovering previously unknown patterns from existing

unlabelled datasets, such as grouping similar data or detecting outliers. Unsupervised learning belongs to exploratory analysis, though being more subjective. Agglomerative approach, iterative partitioning approach, and principal component analysis are main techniques covered. There are no standard procedures and/or right answers for evaluating model performance or validating results. In the age of big data, the importance of unsupervised learning techniques is ever-growing in many application areas.

We especially enjoy the fact that the R statistical environment was chosen for this book. As a growing miracle, R (programming language) attracts many data scientists worldwide due to its open source policy. The advantage of an open source tool like R is that, as new data formats emerge, new packages able to process these data formats are also created.

This no-frills book is a guide to the Data Science Process. It is written for a fairly general audience from different disciplines, such as finance, sales, marketing, product development and others. According to McKinsey, an estimate of 500,000 strong workforce of data scientists will be needed in USA alone by 2018. The resulting talent gap is huge and must be filled urgently (Ledolter, 2013). The author's consulting experiences and wonderful writing style can nurture anyone who needs to jump-start his or her entry into data science using methodologies like machine learning.

References

Chiu, Y-W. (2015) *Machine Learning with R Cookbook*, Packt Publishing Ltd, 32 Lincoln Road, Olton Birmingham, B27 6PA, UK.

Lantz, B. (2013) *Machine Learning with R*, Packt Publishing Ltd, 32 Lincoln Road, Olton Birmingham, B27 6PA, UK.

Ledolter, J. (2013) *Data Mining and Business Analytics with R*, John Wiley & Sons.