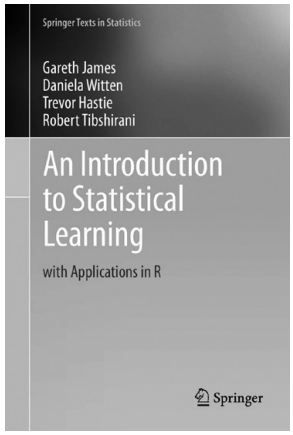
BOOK REVIEW



Gareth James, Daniel Witten, Trevor Hastie & Robert Tibshirani: An Introduction to Statistical Learning with Application in R: Springer, New York, 2013, Corrected at 8th Printing 2017, pp1-426+xiv.

- Reviewer: N.S. VISWANATH

The present book catches any researcher by its title wherein "Application in R" motivates for reading it. The four author written book has ten chapters covering critical aspects of data science for answering research questions raised in marketing, finance, biology and other social sciences. Statistics is a science, an art and is the only discipline whose rigour is rooted in mathematics and methods. The natural numbers in the well-defined number system acquires greater meaning and insights once they are used in the way in which crunching is done lead us to solutions. Data are collected with research design frame work to draw functionalities and testing hypotheses for definitive solutions for larger universal implications. The present book enriches any reader by providing insights in terms of accuracy, precision, interpretability, flexibility and so on functional models fitted by parametric or non-parametric methods. The reviewer got tempted to make a review to enable the target group get access to such a book for gaining competence and confidence in collection, collation and interpretation of data for the work envisaged to answer grey area of research.

The present book is divided into ten chapters. These chapters cover the concepts of statistical learning, Linear Regression, Classification, Resampling Methods, Linear Model Selection, Regularization, Moving beyond Linearity, Tree Based Methods, Support Vector Machines and Unsupervised learning Methods. Models, Estimation and Interpretation of results are critical to any research work. The first chapter discusses in detail regression-linear, linear discriminant analysis, and generalized linear models. Interestingly, reference is made to Generalized Additive Model (GAM). The chapters 1 & 2 discuss model functionality, quality of fit and tradeoff between bias and variance. Introduction to R begins in chapter-2. The limitations of the linear regression, model, coefficients, power of predictability of the model, limitations of linearity, what if the predictor or predictant are classified or quality variable and comparison of linear regression with K-class neighbours are done. The chapter makes an elaborate analysis of linear regression techniques.

Logistic regression for variables is discussed in the back drop of Bayes theorem. Theoretical foundations of Bayes and K-Nearest Neighbours have been discussed in chapter-3.Linear discriminant and Quadratic discriminant analyses have been analyzed using data sets from the real life. While classification is essential for attributes as dependent variable, resampling methods along with cross validation as techniques are

to be used for assessment of test results. Researchers face the problem of interpretability Vs flexibility. Chapter-6 discusses as to how one should mark for a tradeoff. Demonstration has been made of use of ridge regression, the lasso and of dimension reduction. Principal Component Analysis and the partial least squares are well addressed by examples. The shift towards non-linearity is done by polynomial regression. General Additive Model (GAM) for addressing such a problem for both variables and attributes are dwelt in style by the authors in chapter-7. Given a multi-faceted environment wherein interdependence of variables rule the world, decision trees, and random forests are talked by boosting and in chapter-8. Support Vector Machines (SVMs) for multiple classification and logistic regression is in chapter-9. The use of non-parametric methods form the core of fitting the data to a formatted elaboration. The use of ROC curves are discussed for a rational trade off. The use of Clustering and Principal Component Analysis are detailed in chapter-10 under unsupervised learning. In all, the book has several highlights of the tools and their uses in application disciplines. Any user of the book will be familiar with statistical tools after reading in the back drop of R language:

1. K-Nearest Neighbor (KNNN)
2. Logistic regression
3. Linear Discriminant Analysis
4. Srepwise Regression
5. Ridge Regression
6. Principal Component Regression
7. Partial Least squares
8. The Lasso
9. Single Input to Multiple Input Variables
10. Tree Based Methods - a. Bagging, b. Boosting & c. Random forests
11. Principal Component Analysis
12. K-means Clustering
13. Hierarchical Clustering
14. Interpretability Vs Flexibility
15. Model vs. Data Fitting
16. Parametric Vs Non-Parametric Methods.

The book is strongly recommended for researchers in others disciplines and for students of statistical science who would benefit on the application of tools for solving application oriented research problems.