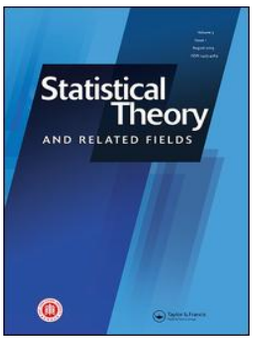

Statistical Theory and Related Fields

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/tstf20

An introduction to statistical learning with applications in R

by Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani, New York, Springer Science and Business Media, 2013, $41.98, eISBN: 978-1-4614-7137-7

Fariha Sohil, Muhammad Umair Sohali & Javid Shabbir

To cite this article: Fariha Sohil, Muhammad Umair Sohali & Javid Shabbir (2021): An introduction to statistical learning with applications in R, Statistical Theory and Related Fields, DOI: 10.1080/24754269.2021.1980261

To link to this article: https://doi.org/10.1080/24754269.2021.1980261

Published online: 26 Sep 2021.

Taylor & Francis
Taylor & Francis Group

BOOK REVIEW

OPEN ACCESS

An introduction to statistical learning with applications in R by Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani, New York, Springer Science and Business Media, 2013, $41.98, eISBN: 978-1-4614-7137-7

In the twenty-first century, Machine learning is a hot trend procedure for handling the real-life problem. Several books and research articles are now available in literature on this topic. An introduction to statistical learning with applications in R by Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani is one of them which provide the fundamental and modern machine learning material with its application using numerous real-life data sets. This book consists of 10 chapters having 440 pages with index.

A master data scientist/statistician must be aware of basic statistical terminologies to evaluate the capacity of the fitted model. For example,

- Bias-variance trade-off
- Difference between classification trees and regression problems
- The difference between supervised learning and unsupervised learning
- Trade-off between accuracy and interpretability and many others.

In 'An introduction to statistical learning with application in R', data scientists will find out all with a comprehensive introduction of machine learning and its applications in *R*. For novice researchers, this book has many attractive features which are helpful in understanding data science with a board aspect.

Chapters 1 and 2 explain the basic terminology and concepts behind statistical learning. Throughout these chapters, readers can answer 'what is statistical learning'. Simply, statistical learning (major algorithms in machine learning are statistical learning) means a procedure or algorithms for modelling based on data. Usually, we find a model that can help us in Prediction (estimate output given input) or Inference (understanding relationships in data) or both Prediction and Inference. These estimation procedures are supervised or unsupervised. Data scientists must find out the suitable fit or understand given data best so that they use metrics such as MSE, Accuracy to evaluate the tool. But in data science, we have 'no free lunch' because of the trade-off between bias-variance and the trade-off between accuracy and interpreter. Hence, data scientist should know about this trade-off theory to find the best model that produces the fitted values more accurately. All detailed explanation is given in Chapter 2 of this book with Lab codes.

The detailed discussion on 'supervised learning' with numerous applications is given in Chapters 3–9. In Chapter 3, the simple linear regression model is defined with application for quantitative and qualitative predictors. Classification settings are defined in Chapter 4 by explaining Logistic Regression, Linear Discernment Analysis (LDA), and Quadratic Discriminant Analysis (QDA). Chapter 5 discusses on resampling by considering the topics of Cross Validation (k-Fold) and Bootstrapping. So, we are able to choose the best model and the best hyper-parameter that bestfit on data. Data dimension reduction is presented in Chapters 6 and 7. In Chapter 8, the authors discuss Tree Based Methods i.e., Regression trees and Classification trees with their Lab codes. The topic Support Vector Machines is given in Chapter 9.

For 'unsupervised learning', the authors present k-means, hierarchical clusters in Chapter 10. All these algorithms are explained very clearly with its practical application.

Overall the book is a nice blend of theory and applications with *R* and we enjoyed reading it. It is clearly a useful resource for researchers who want to work in machine learning. The book could have been made even more appealing by including compressive discussion on Model Based Estimation procedures.

The presentation has been done in an efficient way and the explanation is clear, which enables the researcher to digest the machine learning idea.

Fariha Sohil
Department of Education, TheWomen University, Multan, Pakistan
Muhammad Umair Sohali
Department of Mathematics and Statistics, Institute of Business Management, Karachi, Pakistan
Javid Shabbir
Department of Statistics, Quaid-I-Azam University, Islamabad, Pakistan
umairsohailch@gmail.com

https://doi.org/10.1080/24754269.2021.1980261