The Journal of Open Source Education

An R Companion for Introduction to Data Mining

Michael Hahsler [1]

1 Department of Computer Science, Southern Methodist University, USA

DOI: 10.21105/jose.00223

Software

- Review
- Repository
- Archive

Submitted: 31 May 2023
Published: 11 November 2024

Summary

An R Companion for Introduction to Data Mining is an open-source learning and teaching resource that covers how to implement data mining concepts using R. It is designed to accompany the popular data mining textbook *Introduction to Data Mining* (Tan et al., 2017) to study the implementation of the basic data mining concepts including data preparation, classification, clustering, and association analysis. The resource uses complete, annotated examples to demonstrate how data mining concepts can be translated into R code.

The materials have been made publicly available at: https://mhahsler.github.io/Introduction_to_Data_Mining_R_Examples/book/ and licensed under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 (CC BY-NC-SA 4.0) License.

Statement of Need

The textbook Introduction to Data Mining (Tan et al., 2017) has been one of the most popular choices to learn and teach data mining concepts. Some of the most important chapters have been made available for free by the authors on the books's website. One of the authors also provides Python Jupyter notebooks with examples, but complete R code examples were still needed. Given the R community's interest in data analysis, data science, and machine learning, and the broad support of R packages for data mining, there was a noticeable gap that was filled by this learning resource. This resource targets advanced undergraduate and graduate students and can be used as a component for a first introduction to data mining.

Learning Objectives and Content

The resource assumes basic knowledge of programming and statistics. The learning objectives are to learn how to

- prepare and understand data,
- perform classification,
- perform association analysis, and
- perform cluster analysis.

The resource presents self-contained and annotated R code examples that work with small datasets carefully chosen to show the learner many important aspects of data mining. The learner can copy and paste the examples into a new R markdown notebook to experiment with the code and the provided example data. Small exercises encourage the learner to modify the code by applying it to a different dataset. This learning-by-doing approach has worked well in preparing students to work with more complex real-world datasets by initially relieving them from dealing with too many low-level implementation details while exploring the concepts.

The resource mirrors the textbook's structure so it can be easily used along with the textbook. After a short introduction, Chapter 2 discusses data types in R, data quality concerns, and data preprocessing. In addition, data exploration and visualization examples are given. Chapters 3 and 4 cover classification methods, model selection, model evaluation, different types of classifiers, and essential practical issues like class imbalance. Chapter 5 introduces association analysis with a strong emphasis on visualization. Chapter 7 presents examples of cluster analysis, including popular algorithms, cluster evaluation, and the effect of outliers.

Instructional Design

This resource does not replace the Introduction to Data Mining textbook or instruction by a teacher, it instead provides supporting material for learning to implement data mining concepts in R. The learner is expected to have some programming experience and basic statistics knowledge.

The resource can be used for self-study by any interested person using it to accompany reading the Introduction to Data Mining textbook but its main purpose is to be used as a component for designing an introductory data mining course for advanced undergraduate or graduate students. To support instructors, in addition to the documented code examples, also complete presentation slide sets are provided on the book's GitHub page in PDF and PowerPoint format. The slides are organized in the same way as the resource. A direct connection between the slides and the code examples is provided by the R symbol on the slides where example code is available. The code examples can be assigned to be studied by the students outside of class or used by the instructor in class.

Designing assignments and assessments is left to the instructor since they depend on the level and field of study of the students (e.g., computer science, statistic, economics, or business). For example, for undergraduates, it is suggested to ask the students to apply the data mining techniques to a small, clean instructional data set (sample exercises are available in the resource at the end of each chapter), while graduate students may be asked to analyze larger real-world data sets, which may require a significant amount of cleaning and preprocessing.

Story of the Project

Since starting to teach data mining with R in the Spring 2013, I have been developing the Companion for Introduction to Data Mining resource mainly based on caret (Kuhn, 2008), and a set of packages developed with students to better support different data mining tasks (e.g., arules (Hahsler et al., 2005), seriation (Hahsler et al., 2008) arulesViz (Hahsler, 2017), and dbscan (Hahsler et al., 2019)). The resource grew from a collection of short, unconnected R scripts to a complete set of documented code examples that walk the learner step-by-step through how to implement data mining methods, and how to interpret the results. It went through an update to incorporate the popular tidyverse package collection (Wickham et al., 2019) and a transition from the 1st edition of the Introduction to Data Mining textbook to the second.

The companion resource has been used successfully in the department of Computer Science at Southern Methodist University for many years and by several instructors as a key component of an introductory data mining course delivered in person and in a distance education setting. It is also linked on the textbook website as an official resource. Faculty at the department actively maintains the resource, and we will update it with new R tools like tidymodels (Kuhn & Wickham, 2020) over time.

References

Hahsler, M. (2017). arulesViz: Interactive visualization of association rules with R. *R Journal*, *9*(2), 163–175. https://doi.org/10.32614/RJ-2017-047

Hahsler, M., Grün, B., & Hornik, K. (2005). arules – A computational environment for mining association rules and frequent item sets. *Journal of Statistical Software*, *14*(15), 1–25. https://doi.org/10.18637/jss.v014.i15

Hahsler, M., Hornik, K., & Buchta, C. (2008). Getting things in order: An introduction to the R package seriation. *Journal of Statistical Software*, *25*(3), 1–34. https://doi.org/10.18637/jss.v025.i03

Hahsler, M., Piekenbrock, M., & Doran, D. (2019). dbscan: Fast density-based clustering with R. *Journal of Statistical Software*, *91*(1), 1–30. https://doi.org/10.18637/jss.v091.i01

Kuhn, M. (2008). Building predictive models in R using the caret package. *Journal of Statistical Software*, *28*(5), 1–26. https://doi.org/10.18637/jss.v028.i05

Kuhn, M., & Wickham, H. (2020). *Tidymodels: A collection of packages for modeling and machine learning using tidyverse principles*. https://doi.org/10.32614/CRAN.package.tidymodels

Tan, P.-N., Steinbach, M. S., Karpatne, A., & Kumar, V. (2017). *Introduction to data mining* (2nd Edition). Pearson. ISBN: 978-0133128901

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., … Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, *4*(43), 1686. https://doi.org/10.21105/joss.01686