Unsupervised Learning with R

The R Project for Statistical Computing provides an excellent platform for tackling data processing, data manipulation, modeling, and presentation.

This book introduces you to unsupervised learning and provides a high-level introduction to the topic. We quickly move on to discuss the application of key concepts and techniques for exploratory data analysis. The book then teaches you to identify groups with the help of clustering methods and building association rules. Finally, it provides alternatives for the treatment of high-dimensional datasets, as well as using dimensionality reduction and feature selection techniques.

By the end of this book, you will be able to implement unsupervised learning and various approaches associated with it in real-world projects.

Who this book is written for

This book is intended for professionals who are interested in data analysis using unsupervised learning techniques, as well as data analysts, statisticians, and data scientists seeking to learn to use R to apply data mining techniques. Knowledge of R, machine learning, and mathematics would help, but are not a strict requirement.

What you will learn from this book

- Load, manipulate, and explore your data in R using techniques for exploratory data analysis, such as summarization, manipulation, correlation, and data visualization
- Transform your data by using approaches such as scaling, recentering, scale [0-1], median/MAD, natural log, and imputation data
- Build and interpret clustering models using K-Means algorithms in R
- Build and interpret clustering models by hierarchical clustering algorithms in R
- Understand and apply dimensionality reduction techniques
- Create and use learning association rule models, such as recommendation algorithms
- Learn about and use the techniques of feature selection
- Install and use end-user tools as an alternative to programming in the R console directly

Unsupervised Learning with R

Work with over 40 packages to draw inferences from complex datasets and find hidden patterns in raw unstructured data
In this package, you will find:

- The author biography
- A preview chapter from the book, Chapter 1 'Welcome to the Age of Information Technology'
- A synopsis of the book’s content
- More information on Unsupervised Learning with R
About the Author

Erik Rodriguez Pacheco works as a manager in the business intelligence unit at Banco Improsa in San José, Costa Rica, where he holds 11 years of experience in the financial industry. He is currently a professor of the business intelligence specialization program at the Instituto Tecnológico de Costa Rica’s continuing education programs. Erik is an enthusiast of new technologies, particularly those related to business intelligence, data mining, and data science. He holds a bachelor’s degree in business administration from Universidad de Costa Rica, a specialization in business intelligence from the Instituto Tecnológico de Costa Rica, a specialization in data mining from Promidat (Programa Iberoamericano de Formación en Minería de Datos), and a specialization in business intelligence and data mining from Universidad del Bosque, Colombia. He is currently enrolled in an online specialization program in data science from Johns Hopkins University.

He has served as the technical reviewer of *R Data Visualization Cookbook* and *Data Manipulation with R - Second Edition*, both from Packt Publishing.

He can be reached at https://www.linkedin.com/in/erikrodriguezp.
Currently, the amount of information we are able to produce is increasing exponentially. In the past, data storage was very expensive. However, today, new technologies make it cheaper to store this information. So we are able to generate massive amounts of data, which it is also feasible to store. This means that we are immersed in a universe of data, of which we are not able to exploit the vast majority.

Among these large deposits of data storage there is valuable knowledge, but it is hidden and difficult to identify using traditional methods.

Fortunately, new technologies such as artificial intelligence, machine learning, and the management of databases converge with other disciplines that are more traditional such as statistics or mathematics to create the means to locate, extract, or even construct this valuable information from raw data.

This convergence of knowledge areas gives rise to, for example, very important subfields such as supervised learning and unsupervised learning, both derived from machine learning.

Both subfields contain a large quantity of tools to enhance the use of stored data so that it is possible to generate knowledge about the data and extract it in a human-interpretable way.

In this book, you will learn how to implement some of the most important concepts of unsupervised learning directly in the R console, one of the best tools for a data scientist, through practical examples using more than 40 R packages and a lot of useful functions.

Considering the wide range of techniques and knowledge related to unsupervised learning, this book is not intended to be in any way exhaustive. However, it contains some valuable knowledge and main techniques to introduce the reader to the study and implementation of this important sub field of machine learning.
What this book covers

Chapter 1, Welcome to the Age of Information Technology, aims at introducing the reader to the unsupervised learning context and explains the relation between unsupervised and supervised learning in the context of data mining. It also provides the reader with an introduction to the key concepts of information theory.

Chapter 2, Working with Data – Exploratory Data Analysis, is about some techniques for exploratory data analysis such as summarization, manipulation, correlation, and data visualization. An adequate knowledge of data, by exploration, is essential in order to apply unsupervised learning algorithms correctly. This assertion is true for any effort in data mining, not just for unsupervised learning.

Chapter 3, Identifying and Understanding Groups – Clustering Algorithms, teaches the readers about one of the most used techniques in unsupervised learning, clustering. Identifying groups can help discover and explain some patterns hidden in data. It is frequently the answer for multiple problems in many industries or contexts. Finding clusters can help uncover relationships in data, which can in turn be used to support future decisions.

Chapter 4, Association Rules, covers another grouping technique, the association rules. The association process makes groups of observations and attempts to discover links or associations between different attributes of groups. This association becomes rules, which can in turn be used to support future decisions.

Chapter 5, Dimensionality Reduction, aims to explain some dimensionality reduction techniques. In machine learning, this concept is the process of reducing the number of random variables considered, and it can be subdivided into feature selection and extraction. The key is to reduce the number of dimensions, but preserve most parts of the information.

Chapter 6, Feature Selection Methods, explains some techniques for feature selection, also known as variable selection or attribute selection. The key point is to choose a subset of relevant features of variables for modeling and not to use features that seem to be redundant, considering correlation to simplify model construction.

Appendix, References, provides a list of links referenced in the book, which are sorted chapter-wise. Given the amount of package and functions used in this book, it is very difficult to cite references and authors within the text of each chapter, as it would appear intermittent for the reader.
Welcome to the Age of Information Technology

Machine learning is one of the disciplines that is most frequently used in data mining and can be subdivided into two main tasks: supervised learning and unsupervised learning. This book will concentrate mainly on unsupervised learning.

So, let's begin this journey right from the start. This particular chapter aims to introduce you to the unsupervised learning context. We will begin by explaining the concept of data mining and mentioning the main disciplines that we use in data mining.

Next, we will provide a high-level introduction of some key concepts about information theory. Information theory studies the transmission, processing, utilization, and even extraction of information and has been successfully applied in the data mining context.

Additionally, we will introduce CRISP DM because it is important to use a specialized methodology for management knowledge discovery projects.

Finally, we will introduce the software tools that we will use in this book, mentioning some of the reasons why they are highly recommended.

In brief, we will cover the following topics:

- The data mining concepts
- Machine learning
  - Supervised learning
  - Unsupervised learning
Welcome to the Age of Information Technology

- Information Theory
  - Entropy
  - Information gain
- CRISP-DM
- Benefits of using R

The information age
At present, the amount of data we are able to produce, transmit, and store is growing at an unprecedented rate. Within these large volumes of information, we can find deposits of valuable knowledge to be extracted. However, the main problem is to find such information and it is reasonable to say that it will become increasingly difficult.

Eric Emerson Schmidt, who was the chief executive of Google warned:

"Between the origin of the Earth and 2003 five Exabytes of information were created; today that amount is created every two weeks."

In this context, it is easy to understand that it is virtually impossible to identify these deposits of knowledge by manual methods, and this makes it necessary to resort to specialized disciplines such as data mining.

Data mining
The term began to be used in the ’80s by database developers. Data mining can be defined as the process of discovery of new and meaningful relationships by exploring large volumes of information.

Data mining, as an oriented knowledge discovery process, uses various disciplines, such as mathematics, statistics, artificial intelligence, databases, pattern recognition, or machine learning. Indeed, sometimes some of these terms are considered synonymous, which is in fact incorrect. Rather, there is an overlap of these disciplines.
The following diagram illustrates some disciplines involved in the process of data mining:

![Diagram of disciplines involved in data mining](image)

**Machine learning**

Machine learning is a subfield of computer science, and is defined as the study and creation of algorithms that are able to learn, in the context of this book, from the relationships between the information contained in a dataset.

In general terms, machine learning can be divided into several categories; the two most common ones are supervised learning and unsupervised learning.

**Supervised learning**

This is a task of machine learning, which is executed by a set of methods aimed to infer a function from the training data.

Normally, the training data is composed of a set of observations. Each observation possesses a diverse number of variables named *predictors*, and one variable that we want to predict, also known as *labels* or *classes*. These labels or classes represent the *teachers* because the models learn from them.

The ultimate aim of the function created by the model is the extrapolation of its behavior towards new observations, that is, prediction. This prediction corresponds to the output value of a supervised learning model that could be numeric, as in the case of a regression problem; or a class, as in the case of classification problems.
Welcome to the Age of Information Technology

To explain the process of supervised learning, we can resort to the following diagram:

For reference, some examples of supervised learning models are:

- Regression models
- Neural networks
- Support vector machines
- Random forests
- Boosting algorithms

We can divide the process into two stages **Modeling** and **Predicting**:

In the modeling stage, we start with raw data that will be used to train the model. The following is the definition of the variables used to build the model, as it is possible to reduce the number or transform them.

We proceed to train the model, and finally carry out the evaluation. It is important to note that training, construction, and validation of the model form an iterative process aimed at achieving the best possible model, which means we may need to return to a previous step to make adjustments.
The second stage is the prediction. We already have the model and a number of new observations. Using the model that we built and tested, a prediction for new data is executed and the results are generated.

**Unsupervised learning**

The unsupervised learning objective of this book is a machine learning task that aims to describe the associations and patterns in relation to a set of input variables. The fundamental difference from supervised learning is that input data has no class labels, so it has no variables to predict and rather tries to find data structures by their relationship.

We could say that unsupervised learning aims to simulate the human learning process, which implies learning without explicit supervision, that is, without a teacher as is the case with supervised learning.

In unsupervised learning, we can also speak of two stages: **Modeling** and **profiting**:

In the modeling phase, we take the input data and proceed to apply techniques of feature selection or feature extraction. Once we define the most convenient variables, we proceed to choose the best method of unsupervised learning to solve the problem at hand. For example, it could be a problem of clustering or association rules.
After choosing the method, we proceed to build the model and execute an iterative tuning process until we are satisfied with the results.

In contrast to supervised learning, in which the model value is derived mostly from prediction, in unsupervised learning, the findings obtained during the modeling phase could be enough to fulfill the purpose, in which case, the process would stop. For example, if the objective is to make a customer group, once done, the modeling phase will have an idea of the existing groups, and that could be the goal of the analysis.

Assuming that the model was subsequently used, there is a second stage, which is when we have the model and want to exploit it again. We will receive new data and use the model that we built to run on them and get results.

Throughout this book, we will explain in greater depth, many aspects of unsupervised learning.

**Information theory**
Information theory studies the transmission, processing, utilization, and even extraction of information and has been successfully applied in the data mining context.

Information theory, also known as the *mathematical theory of communication* or the *mathematical information theory* is a theory proposed by Claude E. Shannon and Warren Weaver in the late 1940s.

Information theory is a concept that has been extrapolated to other contexts and is widely used in relation to machine learning and unsupervised learning. Considering that in several examples of this book, some concepts will be mentioned. In this regard, we will explain them in this section.

Information theory defines the degree of dependence between two variables based on the concept of mutual information; that is, the information that is common between two variables and therefore, it can be considered a measure of the reduction of uncertainty about the value of a variable once we know the other.

In relation to the above, there are two important concepts that we want to clarify: the entropy and information gain.
Entropy

Entropy, also known as the information media, gives the mean value of the information by a variable. It can be considered a measure of uncertainty, because it is a measure of how pure or impure a variable is. The entropy ranges from 0 when all instances of a variable have the same value, to 1 when there exists an equal number of instances of each value.

Formally, the entropy can be defined with the help of the following formula:

$$ I = - \sum_{i=1}^{k} P(value_i) \cdot \log_2(P(value_i)) $$

Explaining the mathematical concepts of information theory is beyond the scope of this book. However, considering its importance, we will explain the concept of entropy and information gain using an example:

Suppose we have the following dataset consisting of four variables (Color, Size, Shape, and Result) and 16 observations:
Considering it contains 16 instances: 9 TRUE and 7 FALSE, we proceed to apply the formula of entropy as follows:

\[
I(Ex) = - \left[ \left( \frac{9}{16} \right) \log_2 \left( \frac{9}{16} \right) + \left( \frac{7}{16} \right) \log_2 \left( \frac{7}{16} \right) \right]
\]

\[
I(Ex) = 0.9887
\]

The entropy for the example is 0.9887, which makes much sense because 7/16 and 9/16 is almost a coin flip; hence, the entropy is close to 1.

**Information gain**

When we are trying to decide the relevance of an attribute, we can examine the information gain associated with the variable. Information gains are usually a good measure to decide the relevance of an attribute. It is a measure related to entropy and can be defined as the expected reduction in entropy caused by a partitioning of features. In general terms, the expected information gain is the change in information entropy.

We can calculate the expected entropy of each possible attribute. In other words, the degree to which the entropy would change.

Continuing with the example, we consider the variable size and proceed to the calculation of information gain.

We want to calculate the information gain (entropy reduction), that is, the reduction in uncertainty using the feature size.
The first thing to do is calculate the entropy of each subset of the variable:

<table>
<thead>
<tr>
<th>Feature: Size</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Small</strong></td>
</tr>
<tr>
<td>Color</td>
</tr>
<tr>
<td>-------</td>
</tr>
<tr>
<td>Y</td>
</tr>
<tr>
<td>G</td>
</tr>
<tr>
<td>Y</td>
</tr>
<tr>
<td>G</td>
</tr>
<tr>
<td>Y</td>
</tr>
<tr>
<td>Y</td>
</tr>
<tr>
<td>Y</td>
</tr>
<tr>
<td><strong>Large</strong></td>
</tr>
<tr>
<td>Color</td>
</tr>
<tr>
<td>-------</td>
</tr>
<tr>
<td>G</td>
</tr>
<tr>
<td>Y</td>
</tr>
<tr>
<td>Y</td>
</tr>
<tr>
<td>Y</td>
</tr>
<tr>
<td>Y</td>
</tr>
<tr>
<td>Y</td>
</tr>
<tr>
<td>Y</td>
</tr>
</tbody>
</table>

Then we must add the entropies calculated according to the proportion of observations. In the example, both Size = Small and Size = Large contain eight observations:

\[
\begin{align*}
I(S_{\text{Size}}) &= \left( \frac{8}{16} \right) \cdot 0.8113 + \left( \frac{8}{16} \right) \cdot 0.9544 \\
I(S_{\text{Size}}) &= 0.8828
\end{align*}
\]

As the information gained by definition is the change from entropy:

\[IG = Entropy(S) - Entropy(S_{\text{Size}})\]

\[IG = 0.9887 - 0.8828\]

\[IG = 0.1059\]

So, we gained 0.1059 bits of information about the dataset by choosing the size feature.
Welcome to the Age of Information Technology

Data mining methodology and software tools

To conclude this introductory chapter, we consider it important to note two additional points: a suggested methodology for data mining projects and some important aspects of the software that we use in this book.

CRISP-DM

CRISP-DM is an acronym for Cross Industry Standard Process for Data Mining. Although it is a process model for data mining projects in general, it is a good framework to use in unsupervised learning projects. It is not the only existing standard, but currently, is the most often used.

CRISP-DM, is divided into 4 levels of abstraction organized hierarchically in tasks ranging from the most general level to the most specific cases and organizes the development of a data mining project, in a series of six phases:

<table>
<thead>
<tr>
<th>CRISP-DM Stages</th>
<th>Purpose</th>
</tr>
</thead>
<tbody>
<tr>
<td>Business understanding</td>
<td>This aims to understand the project objectives and requirements from a business perspective and convert this knowledge into a data mining problem.</td>
</tr>
<tr>
<td>Data understanding</td>
<td>This pretends to get familiar with data, to identify quality problems, and to get first insights into the data.</td>
</tr>
<tr>
<td>Data preparation</td>
<td>This covers all activities to construct the final dataset to feed into the modeling tool. The data preparation phase might include tasks such as attribute selection, data transformation, data cleaning, and any other task considered necessary.</td>
</tr>
<tr>
<td>Modeling</td>
<td>The modeling techniques are selected, calibrated, and applied.</td>
</tr>
<tr>
<td>Evaluation</td>
<td>Before proceeding to the final deployment of the model, it is important to perform a more thorough evaluation, reviewing the steps executed for its construction, and to be sure it properly achieves the business objectives.</td>
</tr>
<tr>
<td>Deployment</td>
<td>This is the exploitation phase of the project.</td>
</tr>
</tbody>
</table>
These phases interact in an ordered process, as shown in the following diagram:

We will not delve into an explanation of each of these phases. Instead, we simply suggest the methodology as a framework. However, if you want to investigate further, there is much information available online.

**Benefits of using R**

This book is based entirely on the use of R—a tool that was originally developed by Robert Gentleman and Ross Ihaka from the Department of Statistics at the University of Auckland in 1993. Its current development is the responsibility of the R Development Core Team.

There are many tools for data mining, so let's take a look at a few of the benefits of using R:

- R is Free! And not just free, R is open source software.
- It is probably the most used tool for the scientific community to carry out research, and certainly the most used by professionals working in data mining.
Welcome to the Age of Information Technology

- Perhaps one of the best features it has is a giant collaborative repository called CRAN, which currently has more than 7,300 packages for many different purposes. Very few applications have this diversity.
- It has a very active community along with multiple forums where we can discuss our queries with others and solve our problems.
- R has great capacities for information visualization.
- And a huge so on...

Summary
In this chapter, we have contextualized the concept of unsupervised learning relative to machine learning, supervised learning, and the theory of information.

In addition, we presented a methodology for data mining project management. Finally, we presented the software we plan to use throughout the chapters of this book.

In the next chapter, we will explain some exploratory techniques and thus proceed to execute the next step in the CRISP-DM methodology.
Where to buy this book

You can buy Unsupervised Learning with R from the Packt Publishing website.

Alternatively, you can buy the book from Amazon, BN.com, Computer Manuals and most internet book retailers.

Click here for ordering and shipping details.