Preliminary Final Degree Project
OutliersLearn: R package to understand outlier detection

Author: Andrés Missiego Manjón
Tutor: Juan José Cuadrado Gallego
Degree: Degree in Informatics Engineering

December 2023

1 Introduction

In the dynamic landscape of data analytics, identifying outliers is critical. As The New York Times describes, outliers, "a statistical observation that is markedly different in value from the others of the sample" [1]. Detecting this anomalies can be a huge challenge at first sight but, thanks to past years investigations, data scientist can benefit from various algorithms that can solve this issue. To address this challenge, we are delighted to introduce the "OutliersLearn" package, a toolkit dedicated to learning outlier detection in R.

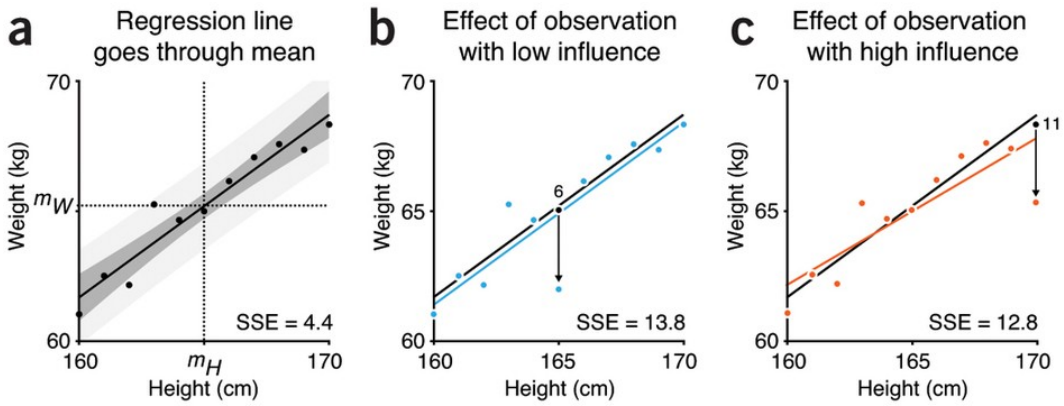
2 About outliers

As it was mentioned earlier, we can describe outliers as observations that has an abnormal distance from other values in a random sample from a population [2]. But, are all outliers the same? Do they have to be treated the same way? Do they have any type of effect on the algorithms used in data analysis, deep-learning, etc.? To understand outliers better and to solve this question, they can classified in two main types:

- Wrong data, coming from measurement errors, which must be eliminated because they will lead to data analysis with erroneous conclusions. This type of outliers are treated in the pre-processing stage (data cleaning) using various techniques that will lead to different results on the final dataset.
- Correct data, with a lot of significance, that deviates from the normal and must be analyzed very carefully because it can lead to important findings.

It's extremely obvious that this two types of outliers must not be treated the same. The first one (wrong data) includes "human errors" (e.g. misspelling a value), distorted measurements, etc. The second one are genuine values that really have an impact on the data (e.g. The Wall Street Crash of 1929), this type of data must be counted in the dataset because it's not an error, it's a significant value. Going back to the questions proposed earlier, outliers do have effect on the data. This can be for worse of for better, depending on the type of outlier (e.g. if the dataset used for a technique/algorithm includes wrong values, this can lead to worse results in the designated study).

This is an example of the effect of an outlier in a linear regression algorithm result [3]:



As we can see, including an outlier on a linear regression algorithm has effect on the result. If this outlier is wrong data, it will lead to worse predictions due to the fact that the regression line is modified by that/those value/s. If the outlier is a true value, then the predictions made by the regression line will take into account the anomaly and it will lead to more precise predictions.

Studies to identify anomalous events or outliers seek to find and categorize as outliers those events that are very different from the rest of those that make up the studied sample. To give a measure of how anomalous an event is we use the outlier score (its definition depends on the technique used). The degree of outlier is set arbitrarily by the data analyst taking into account the study being carried out. The classification of outlier identification techniques depends on the technique used in the analysis:

- Based on proximity: they look for events that are very separated from the rest of the events, they are based on the definition of distances (e.g. KNN)
- Based on density: they look for events that are in a spatial area in which there is a lower density than the observed average (e.g. LOF).
- Cluster-based: They apply to previously clustered events. (e.g BDSCAN clustering algorithm)
- Statistic methods can be classified in:
  - Organization (box and whiskers)
  - Dispersion
  - Regression

3 Objectives and field of application

This project has as field of study the study and application of outlier detection algorithms as a learning R package as well as the implementation of those algorithms. The most used algorithms in outlier detection will be studied, as well as the most modern ones (developing them in a package dedicated to learning outlier detection). The main objective that is chased by developing this R package is to provide an R package that can be used in the academic and professional fields to learn data anomaly detection algorithms in a simple and intuitive way. The package will allow users to gain hands-on experience in outlier detection. By providing a comprehensive learning environment, this project seeks to bridge the gap between theoretical knowledge and practical implementation. By using this R package, people will improve their knowledge of outlier detection algorithms, which will contribute to the advancement of data analysis and decision-making processes.

4 Work description

As it has been mentioned earlier, the main task is to develop an R package that helps on the learning process of outlier detection algorithms. To archive this, the first part of the project will explain this algorithms more deeply (using a more theoretical explanation). The second part will be centered on the development of the package in R (programming the functions/algorithms that have been explained on

the previous part). This package will include tools that can be used to detect algorithms in a learning environment (to understand how the algorithms work from a more practical point of view). Finally there will be included some examples, interesting cases and how outliers affect different algorithms depending on the situation.

5 Methodology and work plan

1. Outlier detection theory
   (a) Introduction to outlier detection
   (b) Most common algorithms
   (c) More outlier detection algorithms
   (d) Importance of detecting anomalies
2. Package development
   (a) Analysis
   (b) Design
   (c) Development
   (d) Tests
   (e) Documentation
3. Examples of use
4. How do outliers affect other algorithms

6 Media

For this Final Degree Project it is required:

- Computer with internet connection
- R & RStudio installed on the computer
- Sweave & noweb for documentation

7 Bibliography

1 Gladwell, Malcolm. "'Outliers.'" The New York Times, The New York Times, 28 Nov. 2008, www.nytimes.com/2008/11/30/books/chapters/chapter-outliers.html.

2 NIST. "What Are Outliers in the Data?" 7.1.6. What Are Outliers in the Data?, www.itl.nist.gov/.

2 Altman, Naomi, and Martin Krzywinski. "Analyzing Outliers: Influential or Nuisance?" Nature News, Nature Publishing Group, 30 Mar. 2016, www.nature.com/articles/nmeth.3812.