An introduction to applied data science

A 5-day workshop towards practical big-data applications and analytics

Eduardo Barbaro and Coen Jonker

Data Scientistists at Mobiquity Inc

About the workshop and this manual:

Welcome to “An introduction to applied data science” a 5-day original workshop thought and written by Eduardo Barbaro and Coen Jonker. At the end of this week you will be able to (i) understand the basics of data science, (ii) mathematically describe data, and (iii) munge[1] it into an easy and communicable form. As the course develops, we will teach you how to access and look at data in innovative ways, as well as to extract value from big data. We aim a significant part of this training to help you create and interpret a diverse set of numerical models, as well as to calculate descriptive analytics while fully understanding their meaning. We also focus on presentation, and how to show your results in beautiful smart charts and tables.

Although all the computations will be performed in a cloud environment, we will guide you through all the necessary steps to install any software necessary for this course in your local machines. We take advantage of a cloud computing environment to get to run our code in a fast, secure, and controlled domain.

This manual is divided in 11 independent short chapters. In Chapter 1 we set the stage with a short introduction to then answer together a bold question: “what is data science?”. In Chapter 2 we will spend some time to recap some basic mathematical concepts needed throughout the entire workshop. Chapters 3 and 4 are devoted to (important) practicalities, such as installing software/libraries and getting familiar with the cloud environment. We also teach you some basic programming skills, so we are all on the same page. We move forward to our first data harvest, in Chapter 5. We teach you how to manipulate data, as well as to properly describe and clean it. Chapter 6 covers the basics of designing a smart numerical experiment.

Chapters 7 to 10 cover more advanced topics such as machine learning tools, databases (SQL and no-SQL) as well as a mapreduce implementation (Hadoop). In particular, in Chapter 10 we show in detail how to load, transform, and extract value from real big data. Finally, in Chapter 11 we come back to basics to show you rudiments of plotting and how to communicate your results.

We wish you a pleasant learning!

[1] describes the constructive operation of tying together systems and interfaces that were not specifically designed to interoperate

Contents

1

A historical view on data science

In this Chapter we briefly try to shed some light on the question “what is data science?”. First of all, it is important to realize that data science is not a new concept. There are many complex definitions out there, but in our view, data science is simply the coupling of very well-established disciplines, such as mathematics and statistics, with a relatively young discipline, computer sciences.

Have a look at Fig. 1.1. It shows the evolution of 5 fundamental disciplines related to data science (Mathematics, Statistics, visualization, technology, computer science). Note that this cartoon depicts events as old as the “invention” of modern calculus by Newton/Leibniz (in the 17^{th} century) or foundation of probability theory by Cardano (in the 16^{th} century). You can see that back in those times all the disciplines were really self-contained areas without much interconnection. As we enter the 20^{th} century things start to get more interconnected. Statistics and Mathematics cannot be separated so clearly any more (e.g. stochastic models, survival models), and computer sciences and technology started to gain more terrain (e.g. relational databases, tree-based models).

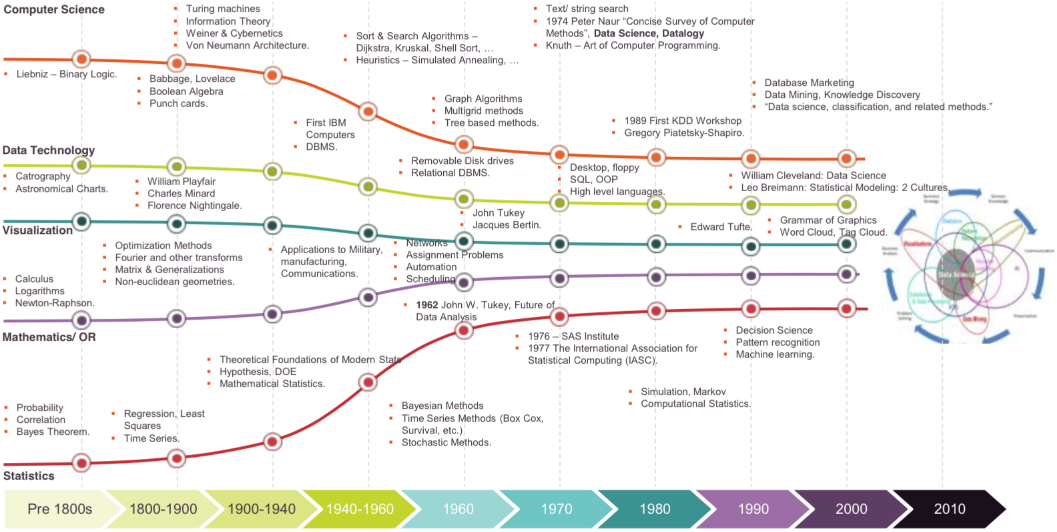

Figure 1.1: Most important events in the history of data science. *Credit: Mamatha Upadhyaya*

As early as 1962, John Wilder Tukey, an American Mathematician, wrote in his "The future of data":

> *For a long time I thought I was a statistician, interested in inferences from the particular to the general. But as I have watched mathematical statistics evolve, I have had cause to wonder and doubt. I have come to feel that my central interest is in data analysis. Data analysis, and the parts of statistics which adhere to it, must take on the characteristics of science rather than those of mathematics data analysis is intrinsically an empirical science.*

That served as inspiration to Tukey who, in 1977, published his most well-known book "Exploratory Data Analysis". In that book, he argued that exploratory and confirmatory data analyses must proceed alongside. Also in the 1970's, Peter Naur, Danish computer science pioneer, published the important "Concise Survey of Computer Methods". In this book, the term data science was used for the first time. According to him, data science can be defined as:

> *The science of dealing with data, once they have been established, while the relation of the data to what they represent is delegated to other fields and sciences.*

In the beginning of the 21^{st} century all the disciplines related to data science merged. In 2008, Jeff Hammerbacher (Facebook) and DJ Patil (LinkedIn) used the terminology "Data scientist" to define their work and teams. Since then, the use of this terminology has fully infiltrated the vernacular, and did not stop growing.

Today, data science and computer sciences (through machine learning) have been put together as almost synonyms. There are new terminologies appearing every year, such as Deep Learning, Big Data, Data Mining. Today, we understand that data science and big data do not mean (just) "lots of data". Instead, it means the creation of a new paradigm in how we do analysis and combine the use of our traditional tools (Mathematics and Statistics) with the technology available nowadays.

2
Some basics and beyond

Before we start with some hands-on data science, we think it is fundamental to ensure we are all on the same page on the Mathematics and Statistics we need throughout this course. But don't worry, it is nothing fancy. We, by no means, have the intent to *formally* teach you any Mathematics. Instead, we just want to refresh some of the concepts we already know, but may have forgotten throughout the years.

2.1 Mean, median, and standard deviation

We would like to start with a good look at Fig. 2.1. This type of graphical representation is called a box-plot[1]. It shows the variation in miles per gallon (MPG) for cars with different numbers of cylinders[2]. For now, let's focus only on cars with 4 cylinders. The black line just above 25 MPG is called median, i.e. the middle of the dataset. It indicates the point such that there is an equal probability of a MPG value falling above or below it. The top/bottom edges of the box indicate the upper and lower quartile, respectively. These quartiles tell us that there is only 25% chance a

[1] The box-plot was created by John Wilder Tukey, the American Mathematician we talked about in Chapter 1.

[2] This data was extracted from the 1974 Motor Trend US magazine

car with 4 cylinders will use more than 30 or less than 23 MPG. *Did you realize that already 50% of our data lies within the box?* The horizontal lines above/below the box represent the maximum/minimum MPG values excluding outliers. Here, outliers are defined as MPG which appear less often than 10% of the cars with 4 cylinders, i.e. above 34 MPG or below 21 MPG.

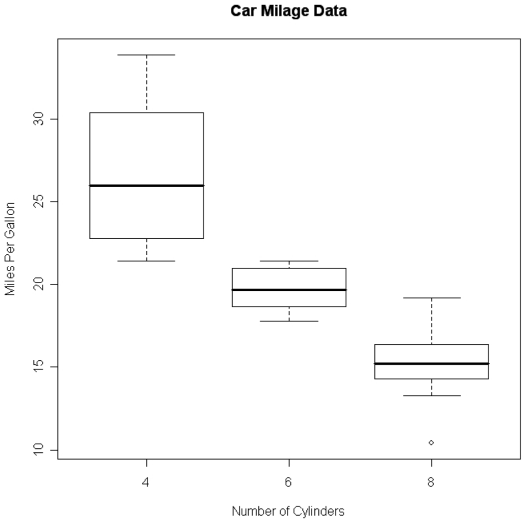

Figure 2.1: Box-plot of miles per gallon by car cylinders. This data was extracted from the 1974 Motor Trend US magazine.

Note that the thick horizontal black lines inside the boxes represent the respective *medians* for 4, 6 or 8 cylinders. Note that this value is different than the mean value. Here is why: the "mean" represents the "average" we all know. Simply sum up all the numbers and then divide them by the total number of occurrences. The "median" represents the "middle" of a ordered series of numbers. This is important because in practice we always find distributions which are highly skewed. In those cases, the median is more representative than the mean.

The mean or the median alone do not tell us the whole story about our data. We still have to understand the data *distribution* and how the individual points are spread around the mean. To help us with that we will introduce the concept of standard deviation. In a nutshell, it represents by how much the individual samples differ from the mean. We depict it in Fig. 2.2.

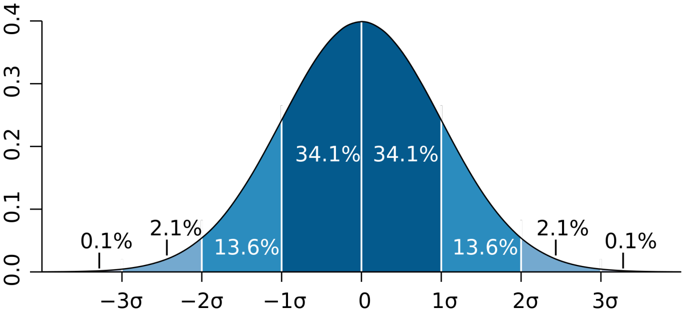

Figure 2.2: Normal distribution where each band accounts for ± 1 standard deviation

The standard deviation is a measure of the dispersion of your data. That means a small standard deviation indicates little dispersion around the average. In contrast, a high standard deviation indicates significant dispersion around the average. As shown in Fig. 2.2, the width of one standard deviation "covers" roughly 68% of the data dispersion. That number increases to over 99% for three standard deviations.

2.2 Interpreting a plot, tracing a linear fit, and understanding R^2

Another important topic we want to talk is how to interpret plots. As we show in Chapter 11 plots are a fundamental way to communicate your data-science results within your company. For now, we will focus on extracting some interesting information from Fig. 2.3. Later, we go in much greater detail on how to design plots and interpret such model results.

We observe a negative correlation between MPG and weight of the car. As the weight increases, MPG decreases. Note that this follows intuitively, since we expect heavier cars to use more fuel and therefore have lower MPG. The blue line represents a linear fit between the variables. The criterion to obtain the linear fit is to minimize the distance between the fit and the data points. This difference is called residuum (shown in red in Fig. 2.3). In Chapter 6 we will cover the design of such simple numerical model in much greater detail.

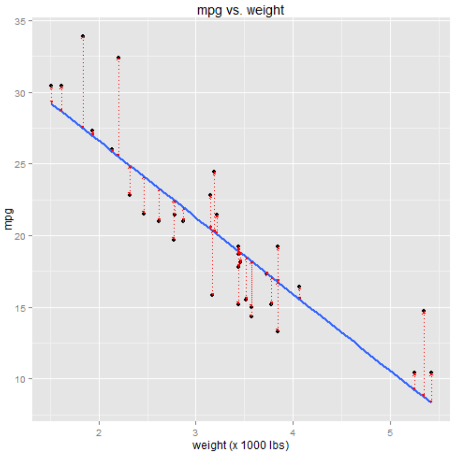

Figure 2.3: Miles per gallon versus the weight of the cars. The blue line draws a linear relation between MPG and the weight of the cars. The red dotted-lines represent the difference (residuum) between the linear fit and each data point. This data was extracted from the 1974 Motor Trend US magazine.

For now let's assume we know how to implement such model - but how do we measure its performance? A simple way to achieve that is by calculating the so-called *determination coefficient*, known as (R^2). R^2 is a metric that uses the residua information to determine if a linear fit is acceptable or not. It ranges from 0 to 1, where zero means no correlation between variables and 1 means a perfect fit. In the case shown in Fig. 2.3 we find that $R^2 = 0.75$, indicating a good linear fit between MPG and weight of the car. In other words, the weight of the car can explain 75% of the MPG's variance. Generally speaking we define $R^2 > 0.5$ as acceptable.

2.3 Histograms

The last topic we want to cover in this Chapter is histograms. A histogram is nothing more than a bar plot whose area represents the frequency of a given variable. In Fig. 2.4 we can visualize a histogram showing the frequency of the MPG variable:

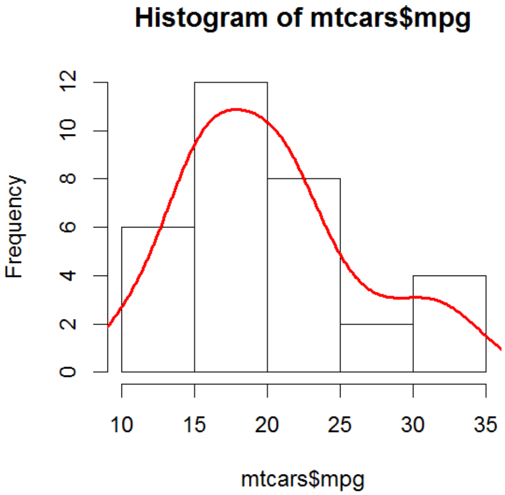

Figure 2.4: Histogram showing the MPG frequency distribution. The red line emphasize its distribution. This data was extracted from the 1974 Motor Trend US magazine.

Note that the width of the bars represents the class interval. In our case every bar shows the frequency for an interval equal to 5 MPG. We can conclude from this plot that most cars have a MPG usage between 15 and 20 followed by 20-25 MPG (see also the red line). This plot also triggers many more questions, such as why is the frequency of cars using 30-35 MPG higher if compared to 25-30 MPG? It is not possible to answer this question by only looking at this plot. The correct strategy is to combine a number of different analyses and plot more variables to understand this behaviour. We will exemplify that in Chapters 5 and 6.

3

Practicalities: Software Installations and AWS cloud

4

Practicalities: Programming language and libraries

5
Harvesting data

6
Numerical Modeling

6.1 Basics

6.2 Designing smart numerical experiments

7

Machine learning tools

7.1 Supervised methods

7.2 unsupervised methods

8 Databases

8.1 SQL

8.2 NoSQL

8.2.1 MongoDB

9 Map Reduce

9.1 Hadoop

10 Big Data

11

Communicating Results

11.1 Visuals