Università degli Studi di Torino
Dipartimento di Informatica

Corso di Laurea Magistrale in Informatica

Tesi di Laurea Magistrale in Informatica

Hands On Data Analysis

Relatore:
Prof. Enrico Bini

Correlatore:
Prof. Yves Van Ingelgem

Candidato
Damiano Gianotti

Sessione di aprile 2022
a.a 2020/2021

Dichiarazione di originalità

Dichiaro di essere responsabile del contenuto dell'elaborato che presento al fine del conseguimento del titolo, di non avere plagiato in tutto o in parte il lavoro prodotto da altri e di aver citato le fonti originali in modo congruente alle normative vigenti in materia di plagio e di diritto d'autore. Sono inoltre consapevole che nel caso la mia dichiarazione risultasse mendace, potrei incorrere nelle sanzioni previste dalla legge e la mia ammissione alla prova finale potrebbe essere negata.

Statement of Originality

I declare that I am responsible for the content of the paper I am submitting for the purpose of obtaining the degree, that I have not plagiarized all or part of the work produced by others and that I have cited the original sources in a manner consistent with current regulations on plagiarism and copyright. I am also aware that if my statement is false, I may incur the penalties provided by law and my admission to the final examination may be denied.

Sommario

In questo documento viene discusso il problema

Abstract

In this document the problem of

Contents

Chapter 1

Introduction

1.1 Data Analysis

Nowadays, it's hard to go anywhere now without hearing about AI and machine learning and data, particularly. It's everywhere. Research has suggested that every two years, we generate more data than ever existed before. So the amount of data is doubling every two years now, that is absolutely an astronomical amount, but the thing is that, of course, this data doesn't necessarily mean anything. The fact is: you can create tables of data, but unless you understand what's in them and what they mean, you haven't got any knowledge. Here we can see an important distinction between having **data** and having **knowledge**. As a species, we're producing a huge amount of data, even if a lot of it doesn't get used, and it sits there on a hard disk waiting for someone to look at it.

If we want to extract knowledge from data we are going to need some tools and processes to do this in a formal way and that's where data science comes into play. So, perhaps, if you do this for your job, then data analysis could be useful for you. Maybe your company's generating data, and you want to analyze this data? On other hand perhaps you (reader) are just a consumer and companies are using data on you. They're generating data on you, and they're profiting from data on you. These are sometimes life-changing decisions that are being made on your data and so it's empowering to know how this process works. Let's do a simple example: suppose you go online to book some flights for a holiday, and then you decide that actually, two flights via an intermediate airport is cheaper than a single flight; for taking that decision you're doing data analysis, taking lots of different data sources and working out the optimal route. This could also happen automatically as well depending on the flight website that you're using. Let's try to formalize this process, what is the meaning of the different topics listed so far? One problem is that multiple definitions existed with a slight difference and, on the other hand, a lot of these terms are used completely interchangeably; AI

is a classic example. You can't buy a product without it having been having AI added to it, when most of the time, the manufacturer are referring to machine learning.

The idea of AI is that we're training a machine to perform a task without explicitly programming it to do so. A good example of AI that isn't machine learning would be a mouse in a maze where all you're doing is telling it to turn left or right at random, not learning anything about the environment itself. It doesn't understand what the maze is, but it will eventually get to the end right: that's a kind of rudimentary artificial intelligence that doesn't involve learning anything Another possible approach, Machine learning, is about **not** giving the mouse different operational conditions, but more about "feeding" it with examples and hoping it will learn to perform most tasks itself. Hence, here is why machine learning is a subset of AI, and it shouldn't be used interchangeably to avoid user confusion. If we use the latter what we'll end doing is training our model based on existing samples of data to either tease out information or make predictions on this data. One of the main operational problems, that I could try in this experience, is that not all data is made nice equal; some of it's noisy and messy, maybe we don't know what it is and don't know whether we can apply a certain technique/idea to it. And so from this come the necessity to clean this data up. This will involve taking this data, understanding what it is and extracting some knowledge so that we can then apply this AI or machine learning techniques to it.

Let's give to important informal definitions; Data science: take data and prepare it in a way that then it can be later used understood. Data analysis: the idea of using statistical measures to try and work out what's going on Perhaps, sometimes, just using statistics to analyze the data isn't enough; you can't learn everything about it, you can mathematically grasp how it works, but you might not understand what it all means. So this is where visualizing the data can be really helpful, that's going to be charting it, plotting it, trying to work out trends and links between different variables. Analysis and visualization, jumping back and forth in a cycle, you could do both of these things numerous times, trying to work a way out

Another important aspect that we should talk about is Data pre-processing. Often you'll be finding your recording much more data than you need. This is certainly true of an online shop, suppose I'm a random customer. As such I'm going to be looking at a lot of products, that I don't end up buying, and I was never really going to buy. In this case, the shop-owners have got to sort of 'weed out' this information to work out what it is that they might better convince me to buy. So this is going to require preprocessing data and removing nonsense and drilling right down to the useful stuff. This is pre-processing and is going to be in the loop together with analysis and visualization, as we can repeat these operations, drill-down and whittle down our data into the most usable sort of the core of knowledge that we can get the most out of it.

Now it may be the case that just analyzing the data is enough, but sometimes you want to take things a little further, we could use machine learning or modelling, which perform two fundamental jobs. One is to classify data like - is there another car on the road? Or, Does this patient have cancer? The other is to make predictions about future outcomes like - will the stock go up? or, Which blog do you want to read next? for predicting what's going to happen next,

Finally, there is data mining and big data. I'm not sure what data mining is because I don't think anyone knows what is it. It's a bit of a buzzword Really what data mining is a combination of pre-processing your data and may be using clustering to extract some knowledge from it So that's our sort of it's a word that's come to be used in place of those things If someone says they're doing data mining, that's what they're doing. It's a night it's a cool sounding word, but you're not mining anything, right? You're just doing what everyone else does on data. Let's assume we've collected a lot of examples of a specific topic, a huge number, and each of our examples is quite complicated, it has a lot of variables and so the amount of data we've got is sort of unwieldy, right? I would argue, perhaps, that big data is not data that you can run on your laptop like you might be using cloud computing infrastructure or certainly parallel processing in some way to pre-process and analyze this data So this is exactly where the line is, how Big Data is.

1.2 What is Data?

We talked a lot about data in the last section and while it is important that we can analyze and understand data, but what is data? Understanding what data is it's a prerequisite for being able to use it properly, perhaps the most important thing as far as we're concerned

1.3 Data Cleaning

1.4 Data Transformation

1.5 Data Reduction

1.6 More advanced technique

Queste sono opzionali

1.6.1 PCA

1.6.2 Classification

1.6.3 Regression

1.6.4 Clustering

Chapter 2

Background

2.1 Why Zensor exist?

Today, most often, technical data sheets coupled with the knowledge of a number of unique experienced individuals are used to determine when maintenance is required for the asset. Product quality only becomes an issue when customers start complaining and repairs are done when it's already far too late. All of these puts tremendous strain on the people responsible for asset, while it could be avoided. Unexpected shutdowns are costly and very demanding for the workforce involved; a possible solution would be making assets smart in order to increase the availability. The only way to validate the actual health is by having a continuous look at a broad data set and adding a specific multi-aspect monitoring setup consisting of different sensor types that follow the behavior of the assets general state-of-health.

2.1.1 About the company

Zensor [1] provides full, integrated and intelligent monitoring solutions for the industrial production, renewable energy and infrastructure sectors. This allows to Zensor's customers to convert the potential contained in the world of IoT and Industry 4.0 into their reality. The company enable digitalization, but with an interface oriented towards the real human: an expert solution without the need for internal experts. It provides not only monitoring devices or data analysis, but offers a full, standardized and standalone end-to-end product that leverages the value contained in a subset of the following aspects:

- operational efficiency
- predictive maintenance
- energy efficiency
- ageing and degradation

- safety

A standard offering consists in several aspects (if required) ordered logically below; as such Zensor takes end-to-end responsibility in monitoring the health and efficiency of structures and processes.

1. Hardware {sensors and acquisition units}
2. Installation and Commissioning {engineering and CAD}
3. Data Management {data transfer, storage, coupling to existing data sources (SCADA, weather, operational...), data cleaning and treatment}
4. Analysis and Reporting {predictions, trend and event detection, real-time reporting through online dashboards.}

Table 2.1: Assets, Industries and Infrastructure for which Zensor has specific products

Asset	Industries	Infrastructure
Rolling cranes	Metal Production	Offshore wind
Grinders and crushers	Mining and Materials	Rail
Flattener rollers	Food Production	Civil Infrastructure
Rolling mills	Glass Production	Energy
Conveyor belts	Discrete Manufacturing	
Tunnels	Textiles	
Chain transporters		
Sieves		
Bridges		

2.1.2 Philosophy

Vision Technological advance can only bring real value to society when the user-facing component is driven by Simplicity and Clarity for the end-user. Zensor sees this as the fastest and most certain route to a world where man-made structures affect the sustainability of our planet in the least possible way:

1. they are intrinsically safe;
2. their useful life is optimized to the maximum;
3. their impact on environment and society is quantified and communicated.

Mission Translate technological innovations in monitoring and analysis into easy-to-understand, tangible and relevant information that we share in the way tailored for either production managers, management or maintenance professionals... As such we are the knowledgeable and easy-to-reach companion for owners and operators in making their assets increasingly safe, efficient and sustainable. If up to us, till eternity.

2.2 Zensor Approach

2.2.1 What are the stages

Hardware and/or other Existing Data sources Sometimes not enough data is available from the beginning. Deriving valuable insights from a monitoring system states from the data: identifying and locating the relevant data in existing databases/data warehouse or putting the right sensors, with appropriate settings, on the right positions and measurement conditions; afterwards reading them out in the optimal way. All of this are defined clearly in every asset-specific package.

Installation and Commissioning Push the button of Industry 4.0. Initially links to the existing data sources are established and data gets ingested. Where required a set of acquisition units and sensors is installed on the machine or structure. After a final verification on the spot (SAT) the monitoring system is launched: the assets enter the IoT.

Data Management Data is continuously streaming in from individual setups as well as historian sources. Structuring, verifying and cleaning the data sets is an essential prerequisite to allow for a profound analysis afterwards: on your way to an automated, continuous and smart follow-up.

Analysis and Clear Reporting Advanced insights are unlocked using algorithms based on physics as well as big data approaches. Clear dashboards, warnings and periodic reports inform the owner or line manager about the present state and upcoming issues. Surprises are avoided, standstills reduced.

2.2.2 Features

Availability Have a continuous idea of availability, automatically as the platform combines different input streams and contextual information.

Performance Based on the data collected and machine-learning based methods for determining the operational condition the performance is calculated.

Warnings Whenever values start to deviate, or data streams stop, warnings are sent. This avoids 'black holes' in the insights of the production line or assets.

Quality Coupling to existing databases or using human input fields the product quality is linked to operational process parameters.

MTTF The Mean Time Till Failure is tracked continuously, for each asset covered the overall 'disturbance free' operation is displayed.

MTBF As events and operating conditions are automatically detected the Mean Time Between Failures is determined continuously, giving a good insight on where optimization is possible.

OEE Using all parameters cited above the Overall Equipment Effectiveness is determined, a major parameter for optimizing asset management strategies and future investments, with a huge cost savings potential.

Factory information systems Such systems are crucial for obtaining operational excellence. When well managed they maximize efficiency and effectiveness. Automated data collection and advanced analysis makes this possible.

2.3 Stack/Pipeline in a nutshell

Chapter 3

What did I do in Zensor?

3.1 Monitor electricity consumption

3.1.1 Student Dorms

3.1.2 V U B Campus

3.1.3 MOBI Hospital

3.2 Improve existing Industrial production

3.2.1 Analyse blade grinder vibration

[2]

3.2.2 Increase efficiency of tomato company

[3]

Chapter 4

What did I learn

4.1 From success

4.2 From failures

Appendix A

development environment

Software	**Library**	**Versions**
Grafana		7.5.8
Python		3.10.0
	zipp	3.5.0
protoc		3.15.6

Table A.1: List of software and libraries used.

Bibliography

[1] Maarten and Yves. *Zensor: our approach*. [Online; accessed 16-January-2022]. 2022. URL: https://www.zensor.be/our-approach.

[2] @2022 Stumabo. *Stumabo*. [Online; accessed 23-January-2022]. 2022. URL: https://www.stumabo.com/en.

[3] @2022 Stoffels Tomaten. *Stoffels About*. [Online; accessed 23-January-2022]. 2022. URL: https://www.stoffels-tomaten.be/en/about-stoffels.