IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 7.39 ⁜ Vol. 11, Issue 5, May 2022

DOI: 10.17148/IJARCCE.2022.11527

Data Collection for Machine Learning

Megha Kharat[1], Sheetal Wadhai[2]

Department of Computer Engineering, Universal College of Engineering, Pune.

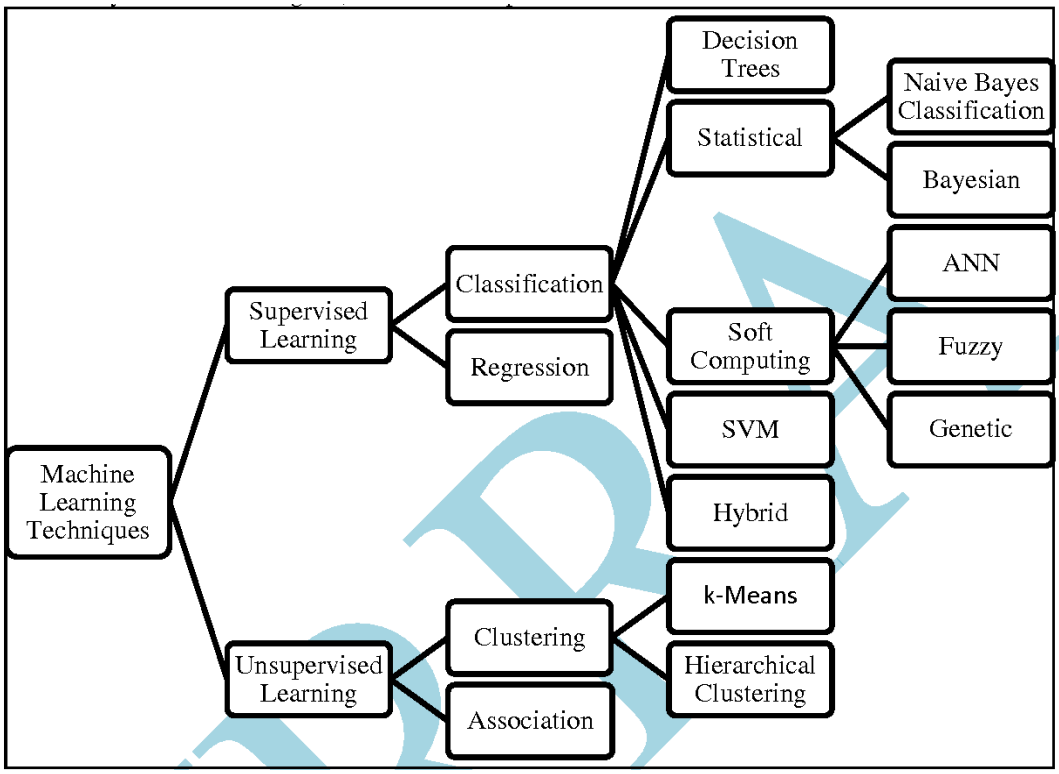
Abstract: We have presented an overview of strategies used in the applied behavioural sciences to assess variables. The majority of the methodologies are employed to varying degrees by quantitative/positivist and qualitative/constructivist researchers. Qualitative researchers prefer less regimented, more open-ended data collection procedures than quantitative researchers.

INTRODUCTION

At any given time, any software developer meets a circumstance in which the work at hand comprises several conditions and branches, and the addition of one additional input parameter can result in a whole rebuild of the entire solution. Or you can find yourself in a situation where you've exhausted all of your options, weighed all of the pros and disadvantages, and realised there's no way to fix the problem without utilising magic. You wish you could wave your magic wand and say "I wish..." to summon a solution capable of making sound decisions and even responding to new information. Furthermore, it would be ideal if the system could teach itself. It sounds like something out of a fairy tale until recently, it was a fairy tale.

Image recognition is one of these difficult tasks: objects, animals, images of interior human parts, faces, or even space objects. Each of those categories has an infinite number of variations.

WHY YOU NEED DATA FOR MACHINE LEARNING: HOW IT WORKS:



Waverley Software has been providing Machine Learning and Artificial Intelligence services to businesses ranging from start-ups to enterprises. Our data engineers have worked on a range of machine learning projects and have identified the major issues that clients confront in this difficult area.

But, before we get into topics like ML and Data Science and try to explain how they function, we need to address a few questions:

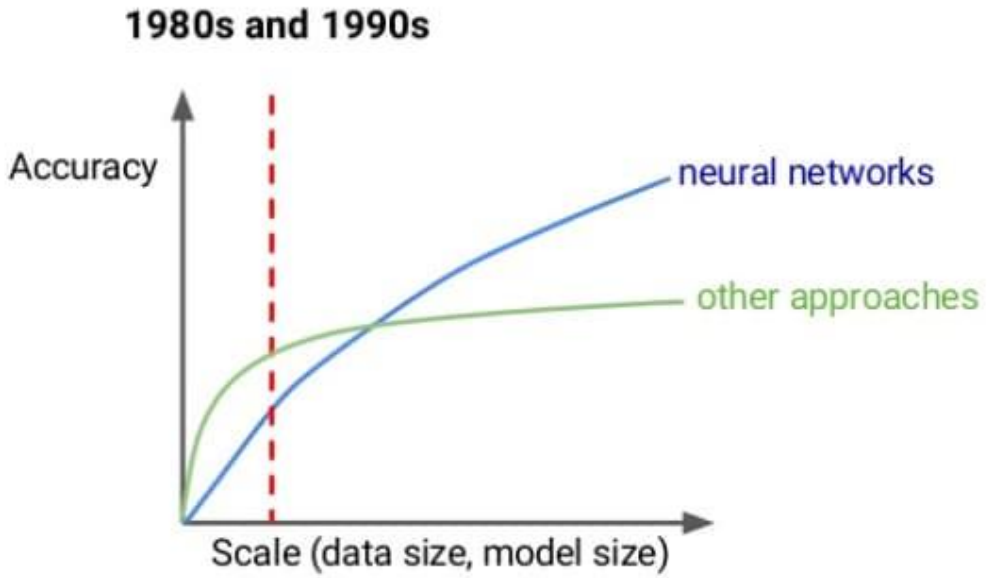
IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 7.39 ⁜ Vol. 11, Issue 5, May 2022

DOI: 10.17148/IJARCCE.2022.11527

• What can we do in business or on a project with the help of machine learning? What objectives do I hope to achieve with ML?
• Do I just want to be trendy, or will using ML actually improve user experience, increase profitability, or protect my product and its users?
• Do I need the system to be able to anticipate anything or only detect anomalies?
The answers to those questions substantially influence the source, format, and even quality of the incoming data. You might even realise you don't require ML at all.
Machine Learning Data Data is as essential to the proper operation of the ML system as oxygen is to living organisms. If you go back 30 years, you'll notice that the issue of data was particularly difficult. If data had not been completely digitalized, there would be hundreds of thousands of photos of individuals on the internet, fitness trackers would not be transmitting data to the cloud, and hospitals would not be able to store people's data in alphabetized files with attractive inscriptions. And the loss of a folder like that on a computer was a calamity because there were no backups available online.
The availability of vast amounts of structured and unstructured data has allowed practical applications of ML to explode in recent years. We have fantastic spam filters, auto-corrections of text input, convenient solutions for voice & text recognition, image search, or music fragment search, and soon – ubiquitous self-driving automobiles, thanks to ML. From an academic standpoint, the Stanford ML course of Andrew Ng from 2011 (at the time of writing this article) has nearly 4 (3.98) million students. This course is highly recommended for anyone interested in learning about machine learning
But let's come back to the importance of data in the process of learning. Maybe you've seen this image before:



Any Data Science specialist will tell you that having too much data is always preferable to having too little. And this is especially true for Deep Learning: the more samples you have, the more correctly the connections between neurons match to the chain of transitions on which the system will make a judgement.
There are also approaches for calculating the smallest dataset required based on the task at hand. For example, historically, the rule of thumb for Deep Learning classification would be 1k samples per class. From my own experience, I can tell you that if you utilise pre-trained models that are appropriate for your classification, this number can be reduced. In my example, using a pre-trained model for facial recognition or facial identification allowed me to successfully identify a person with only 10 photographs.
Of course, you should not overlook data quality. An unbalanced dataset, for example, will have a detrimental impact on the outcomes of a binary classification since one class would dominate in terms of the amount of samples inside a dataset. Using imbalance correction approaches, the problem can be solved by evaluating precision and recall rather than accuracy. However, according to this study, expanding the dataset will be a far better answer to this problem.

HOW TO WORK WITH EXISTING DATA: DATA CLEANING, LABELLING

What is Machine Learning?

Now that we have data, we must determine what Machine Learning is. In layman's terms, ML is the process of extracting knowledge from data.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

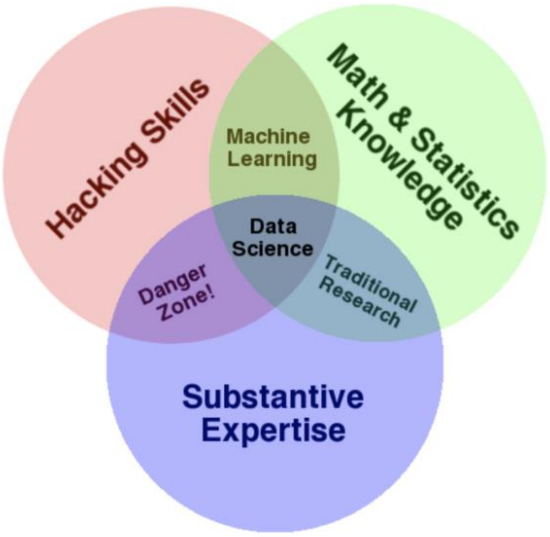
International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 7.39 ⁜ Vol. 11, Issue 5, May 2022

DOI: 10.17148/IJARCCE.2022.11527

"Machine Learning" is defined as "a branch of research that provides computers with the ability to learn without being explicitly programmed" - Arthur Samuel.

Looking at Drew Conway's data science Venn diagram, we can identify three distinct domains that interact with ML: computer science, math, and statistics. On the diagram, you'll also notice that ML is a subset of Data Science, but we'll get to that later.



.

There are numerous subfields and applications of machine learning (ML), such as neural networks (NN), genetic algorithms, data mining, computer vision, natural language processing (NLP), and others. We can define three sorts of ML based on what we want to achieve from the output and what data we have on the input:

- **Supervised learning:** The goal here would be to train a model that can make predictions based on previously unseen future data. Data must be labelled in order for this to happen.
- **Unsupervised learning:** This type of learning works with unlabelled data, and its purpose is to discover hidden patterns and, most likely, useful information.
- **Reinforcement learning:** The goal here is to create a system that learns and improves over time by interacting with its surroundings.

The decision between the three is determined by the problem we're attempting to solve, which is determined by the questions we should have asked ourselves (and preferably answered) at the outset. Supervised learning is our first choice for problems involving classification (distinguishing cats from dogs in a photograph) or regression (predicting the weather for the next month). Unsupervised learning is used when we have unlabeled data and need to conduct clustering (segmenting consumers of an online business), dimensionality reduction (removing superfluous features from a model), or anomaly/outlier detection (finding users with abnormal or suspicious website visiting behaviours). As you can see, these two forms of ML tackle a wide range of problems, and the only difference between them, besides from the tasks, is data: Unsupervised learning does not always require labelled data, but supervised learning must.

WHERE DOES LABELLED DATA DWELL?

Data labelling is the process of categorising or annotating data for use in machine learning.

Labels differ and are unique for each dataset, depending on the task at hand. Labels in the same dataset can have distinct meanings and be used for different activities. For example, the classification of cats and dogs can be expanded to include animals with and without spots on their fur.

Depending on the size and complexity of the dataset, the size of the in-house Data Science team, as well as the time and money, we can organise the Data Labelling process in numerous ways:

- **Crowdsourcing**: A third-party service provider provides a platform for individuals and organisations to outsource their procedures and jobs.
- **Outsourcing**: Hiring freelancers or contractors;
- **Specialized teams**: Employing Data Labelling teams that have been educated and managed by a third-party firm;
- **In-house teams**: assigning Data Labelling responsibilities to an internal team of workers or data scientists

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 7.39 ⁜ Vol. 11, Issue 5, May 2022

DOI: 10.17148/IJARCCE.2022.11527

Each of them has advantages and disadvantages (for example, the quality of the products, the cost of the task, or the speed with which the labelling is accomplished), and one method that works for one endeavour may not work for another. Furthermore, you can mix and match them as you go.

DATA COLLECTION PROCESS

The first step in developing any data science algorithm is determining the desired objectives. In our instance, we wanted to keep track of these types of actions. In the Horse Analytics project, we intended to distinguish four basic training activities of a horse: standing, walking, trotting, and galloping.
To train an algorithm to recognise any action, you must offer the correct data. As the fundamental constituent, data is sent through a neural network until it begins to detect patterns and develop conclusions based on similarities.
Keep in mind that only high-quality data can be used to build an accurate information model. But here's the thing: when you're working on a one-of-a-kind software, you're unlikely to stumble across an organised database or, in some cases, a searchable database, any records whatsoever.

1. Make a data-gathering strategy
Make a strategy outlining the categories of data you'll need, the amount of data you'll need, and the subjects of your data collection before you start acquiring it. You should also be aware of the data requirements' maximum and minimum. Your data scientist is in charge of all of these requirements.

2. Organize a team to gather data
As the investigation proceeds, you'll gain a clearer understanding of what you should include and exclude from the plan. You'll see that certain data simply adds noise to the analytical process, while others improve precision. As a result, you should analyse and adapt your plan on a regular basis based on your own circumstances.
When it comes to data, having a team of specialists you can rely on is essential; someone who understands the importance of acquiring the right information. They should be aware that violations of the data collection flow result in data corruption, thus it is their responsibility to monitor the data collection flow and detect any issues that arise.
These data scientists should be able to work alone with little supervision, allowing you to delegate jobs later. It's vital to have a team that helps you grow and engage new people without your direct involvement.

3. Organize data-gathering tools
To collect data, you'll need specific equipment and software tools, which can vary depending on the job. It's critical to realise that each piece of hardware collects data in its own unique method. When comparing data from two different device kinds, for example, you may notice some variances because the sensors are different. To avoid this and improve data accuracy, we used a few telephones during the data collection method.
It is vital to be consistent when gathering data. We tried to be as precise as possible by placing the device in the same pocket and gathering data in the same way during each collection session.
While data collecting is a mechanical process, its utility is controlled by human characteristics. Ensure that your entire data science team The entire crew is on the same page.

4. Expect low efficiency throughout the initial iterations
At initially, everything proceeds slowly, which is expected considering that everyone is new to the process. You should spend some time at the start double-checking each stage of the data collection procedure. But don't worry; once you've beyond the "trial-and-error" stage, the data collection procedure will be considerably faster and smoother.

5. Always go through the information you've obtained
Every data scientist's nightmare is putting in a lot of time and effort to acquire data only to discover that it's damaged afterwards. Anything can go wrong: certain sensors may malfunction or stop working entirely, while others may produce irregularities. As a result, you should always analyse the data you receive and try to identify any flaws as soon as possible so that they may be corrected.

6. Prepare a data preparation toolbox
Pre-processing is required if you want to obtain rapid feedback on difficulties with the data collection toolbox.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 7.39 ⁜ Vol. 11, Issue 5, May 2022

DOI: 10.17148/IJARCCE.2022.11527

HOW TO PREPARE YOUR DATASETS FOR MACHINE LEARNING?

1. Examine the accuracy of your data

Do you have faith in your data? That is the first question you should ask. Even the most advanced machine learning algorithms will fail in the presence of poor data. We go into more detail about data quality in another article, but there are a few key points to remember.

What is the magnitude of human error? If your data is acquired or tagged by humans, test a subset of it to discover how often errors occur.

Were there any technical difficulties during the data transfer? For example, the same documents could be duplicated as a result of a server failure, or you could have experienced a storage disaster or a cyberattack. Examine how these events affected your data.

2. To make data consistent, format it

Another phrase for data formatting is the file format you're using. It's also not difficult to convert a dataset into the file format that works best for your machine learning system. We're referring to the consistency of the format of the recordings themselves. If you're merging data from several sources or your dataset has been manually updated by multiple people, it's worth double-checking that all variables inside a specific attribute are written consistently. Date formats, monetary quantities, addresses, and so on are examples. The input format should be the same throughout the dataset.

3. Data should be minimized

Because Well, huge data! It's tempting to incorporate as much data as possible. That is a blunder. Yes, you should gather as much information as possible. It's best to decrease data if you're compiling a dataset with certain goals in mind.

You already know what the target property is (the value you want to forecast). Without any forecasting input, you may guess which variables are crucial and which would add more dimensions and complexity to your dataset.

This method is known as attribute sampling.

Record sampling is another way. Simply removing records (objects) with missing, erroneous, or less representative data improves forecast accuracy. The methodology can also be used later on when you need a model prototype to determine whether a machine learning method you've chosen achieves the expected results and evaluate the ROI of your ML endeavour.

You can also reduce data by aggregating it into larger records by splitting the whole attribute data into several categories and drawing the number for each category. Instead of looking at the most popular items on any given day over a five-year period, combine them into weekly or monthly rankings. This will help to reduce data volume and processing time without causing any problems discernible prediction losses.

4. ETL and Data Warehouses

The first is warehouse data storage. These storages are often used to hold structured (or SQL) records that fit into traditional table shapes. This category includes all of your sales records, payrolls, and CRM data. Another common component of dealing with warehouses is data transformation before loading it into a warehouse. This post will go through data transformation options in further depth. However, it assumes that you know what data you need and how you want it to appear, so you analyse it all before saving it. This is known as the Extract, Transform, and Load approach (ETL). The problem with this approach is that you never know which data will be useful and which will not be. As a result, warehouses are commonly used to visualise the metrics that we know we need to track via business intelligence interfaces. There is also a third choice.

5. ELT and Data Lakes

Data lakes are storage systems that may hold both organised and unstructured data, such as images, videos, voice recordings, and PDF files. Even though data is arranged, it is not altered before being stored. You'd import the data in its current state and then decide how to use and handle it on-demand. This is known as the Extract, Load, and Transform approach.

6. Managing the human element

Another factor to consider is the human factor. Data collection can be a time-consuming job that overloads your staff with too many directions. If employees are expected to keep records on a regular and manual basis, they are likely to reject these tasks as yet another bureaucratic whim and quit their jobs.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 7.39 ⁜ Vol. 11, Issue 5, May 2022

DOI: 10.17148/IJARCCE.2022.11527

7. Identify the issue as soon as possible

Knowing what you want to forecast might help you decide which data is most useful to collect. When formulating the issue, conduct data exploration and try to think in the areas of classification, clustering, regression, and ranking that we mentioned in our whitepaper on the commercial application of machine learning. In layman's words, these responsibilities are divided as follows:

Clustering

You need an algorithm to determine the categorization criteria and the number of classes. The main distinction between this and classification jobs is that you have no idea what the group and division principles are. This is common, for example, when you need to segment your customers and tailor a different approach to each group based on their attributes.

Regression

You'd need an algorithm to generate a numerical value. For example, if you spend too much time determining the appropriate pricing for your product because it is dependent on so many variables, regression techniques can assist you in estimating it.

REFERENCES:

1. https://www.hindawi.com/journals/sp/2021/7880477/fig9/?msclkid=c7999a61ffc817625f8cf8e3950828f7&utm_source=bing&utm_medium=cpc&utm_campaign=HDW_MRKT_GBL_SUB_BNGA_PAI_DYNA_JOUR_X_PJ&utm_term=imaging&utm_content=JOUR_X_PJ_Stroke%20Research%20and%20Treatment
2. https://waverleysoftware.com/blog/data-collection-for-machine-learning-guide/
3. https://www.suntec.ai/blog/a-complete-guide-to-data-collection-for-machine-learning/
4. https://www.suntec.ai/blog/a-complete-guide-to-data-collection-for-machine-learning/