Sculpting Data for ML: The first act of Machine Learning

Rishabh Misra and Jigyasa Grover

{r1misra, jigrover}@eng.ucsd.edu

Abstract

In the contemporary world of Artificial Intelligence and Machine Learning, data is the new oil. Rightly so, giant leaps in this domain can be attributed to access to large-scale data. Hence, for Machine Learning algorithms to work their magic, it is imperative to lay a firm foundation by acquiring knowledge of curating good quality datasets. With the modern bloom of social networks, online retailers, streaming platforms, and knowledge and experience sharing platforms, there is no shortage of any form of data, be it textual, audio, or visual. All we need are the skills to identify valuable information and extract meaningful datasets to fashion more precise models. Sculpting Data for ML functions as the first act of the play of Machine Learning. It aims at enlightening Machine Learning and Artificial Intelligence enthusiasts, practitioners, and data scientists about one of the fundamental aspects of this realm, Dataset Curation. This work's distinctive feature is that it puts forward a step-by-step guide on constructing a good quality dataset from scratch. The hands-on tutorial ingrained in the work uses Python with tools like BeautifulSoup and Selenium to coach how to ethically gather data from various web sources. The whole flow is pinned on the fact that predictive models necessitate access to relevant, structured, and distinctive data to maneuver effectively. Overall, the work covers different techniques for dataset building, preprocessing, and engineering impactful features, thus highlighting the significance of data representation for Machine Learning models. Apart from molding data in its worthy format, this work also discusses ways to deal with noisy and unreliable data. Towards the end, it lays out various Machine Learning paradigms, and their data needs to showcase how to identify suitable learning algorithms to solve challenging problems effectively.

Introduction

The last couple of years have seen immense growth in the application and adoption of Machine Learning across diverse domains. There are various reasons behind this explosive growth: information overload, the need to automate mundane tasks, advancing the current state of technology, and sometimes just curiosity about the extent of possibilities. Following are some of the applications of Machine Learning in each of the abovementioned dimensions:

- *Information Overload*: Machine Learning tackles information overload by providing recommendations based on people's liking, like suggesting what products to buy, whom to connect with, what song to listen to, and what type of content to view, all based on their past engagements and inferred interests.
- *Automation of mundane tasks*: Machine Learning automates many everyday tasks like shortlisting resumes for a job posting, identifying grammatical mistakes in a text, transcribing audio in a video or a podcast, suggesting appropriate text responses, and so on. Not having to do such mundane tasks saves much human time and effort, which could be spent on more critical tasks.
- *Advancement of technology*: Academic institutions and business corporations are also plying Machine Learning to further the present state of scientific know-how. Instances of these include developing self-driving cars, improving healthcare quality by advancing the technology to diagnose ailments in their early stages, and enhancing agriculture produce using computer vision technology to monitor crops.
- *Testing the limits*: Creative use cases of Machine Learning include creating music or meaningful lyrics without human interference, synthesizing pictures and videos of non-existent people, automatically generating food recipes, and detecting sarcasm or spoilers in a text snippet.

Unquestionably, Machine Learning is the most used and abused sub-domain of Artificial Intelligence presently. Regardless of our fascination or loathe for it, it heavily influences our decision making power and dominates our lives presently. To describe, enabling computers to learn on their own is what encompasses Machine Learning. The power of spotting patterns without programming is the most prominent edge these decision-making systems have over the others. Researchers keep traversing unexplored territories of Machine Learning, whereas, according to experts, the businesses have just seen the tip of the algorithmic iceberg. According to finances online, US$28.5B was allocated to Machine Learning worldwide in just the first quarter of 2019, which is staggering since the figure was only US$1.3B in 2016. Almost 50% of companies have either started exploring or are planning to incorporate Machine Learning soon.

The number of startups focusing on just Machine Learning services are having a 14x rate of increase currently. 97% of mobile users use Machine Learning trained voice assistants on their devices, with a 40% search now powered just by voice. Advancement in the software aspect has also led to the projected growth of US$120B in global sales of AI-powered hardware by the end of 2025. Enterprise domains like security, analytics, and marketing are reaching new heights with Machine Learning with a 25%, 33%, and 16% increase in adoption rates, respectively. Dominant leaps leave the scientists, investors, policymakers, business leaders, and the audience enthralled, hinting that human-like intelligence in machines might be just around the corner. Nonetheless, progress in Machine Learning has been impressive, but there is a lot of pending explanation and examination, which keeps the research going on.

Significance of Data

For the state-of-the-art Machine Learning algorithms to work their magic, it is essential to focus on the three key dimensions:

- Well-Calibrated Data
- Sophisticated Algorithms
- Efficient Computation

The algorithms mature from the iterative process of experimenting and validating hypotheses. Furthermore, efficient computation requires optimizing the algorithm using distributed processing to run on a large scale. These are absolutely the essential aspects to consider; however, laying the foundation of the process with the perfect quality and quantity of data is the secret sauce that makes Machine Learning effective.

Since we have progressed from the primeval rule-based approach to a more data-driven approach, it goes without saying that we train Machine Learning algorithms to capture implicit patterns in the data provided. The type of data fed into the algorithm thus has a profound effect on the algorithm's success. Worthy data collection forms the foundation of the pyramid of the *AI Hierarchy of Needs* drawn parallel to *Maslow's Hierarchy of Human Needs* by Monica Rogati, a renowned Data Scientist and AI Advisor. Rogati puts forward that data literacy, data collection, and data flow form the basic needs that must be satisfied to achieve *self-actualization*, which would be the attainment of AI.

With the desire to climb the Data Science ladder and contribute to the fabrication of successful Machine Learning algorithms, one should also focus on dealing with data. Before we can employ a Machine Learning solution, it is crucial to understand the problem and the data requirements, followed by researching and accumulating data from the right source.

In most cases, we cannot use the collected data *as-is*. It might need some massaging with appropriate tools and techniques. Once that is done, we can bring into play suitable learning algorithms to address the problem at hand. At this point, it is good to note and understand that no quantum of algorithmic sophistication and efficient computation can make up for the low quality and quantity of data.

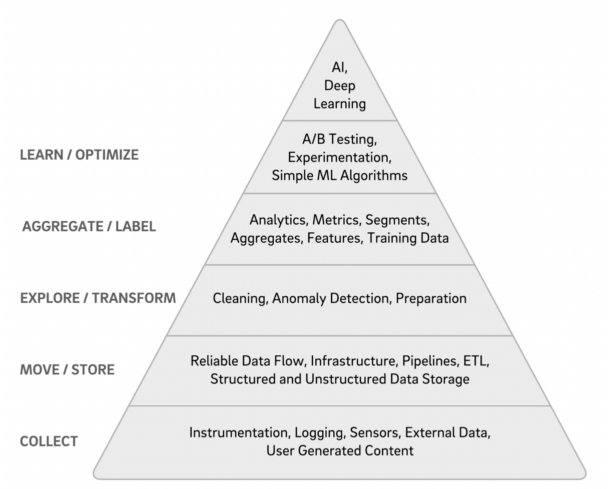
AI,
Deep
Learning
LEARN / OPTIMIZE
A/B Testing,
Experimentation,
Simple ML Algorithms
AGGREGATE / LABEL
Analytics, Metrics, Segments,
Aggregates, Features, Training Data
EXPLORE / TRANSFORM
Cleaning, Anomaly Detection, Preparation
MOVE / STORE
Reliable Data Flow, Infrastructure, Pipelines, ETL,
Structured and Unstructured Data Storage
COLLECT
Instrumentation, Logging, Sensors, External Data,
User Generated Content

Figure 1: The AI hierarchy of Needs by Monica Rogati.

Current times are witnessing accelerated advancement in intelligent systems. Just about every Machine Learning and Data Science enthusiast, conventional developer, and corporate organization is attempting to hop on the bandwagon. In that act of haste, it is widespread to skip the modus operandi of sculpting the data before training and evaluating the Machine Learning model, and eventually running into a wall. Hence, to avoid the misspending of time, money, and efforts in the race of building the perfect artificially intelligent archetype, it is pivotal to follow the process systematically

Honing Dataset Curation Skills

As we discussed above, well-calibrated data is what fuels the Machine Learning engines. There are various sources from which one can pick an already preprocessed dataset that might not require further trimming, anonymizing, standardizing, normalizing, and so on. This data can directly be fed into a model without any changes, thus providing us an easy way to get up to speed with the experimentation of sundry Machine Learning techniques. Using a preprocessed dataset also comes in handy when one just wants to check the working of the end-to-end Machine Learning pipeline; however, it might not guarantee excellent performance. There could be cases when the datasets available do not precisely match our problem's wavelength or are not present in enough quantity required by our use case. At times like these, it is convenient to synthesize our dataset and pack it up with information potentially valuable to the use case in the desired quantity. The contemporary world has volumes of all kinds of crude data available on the web. Equipped with a Machine Learning mallet, we can hammer all the nails in this data-oriented world with an added skill of identifying and extracting meaningful datasets.

Importance in Academia

World over, the scientific community engaged in higher education and research has been hustling quite a lot with Machine Learning and Data Science these days. Core Machine

Learning and Data Science research groups in academia are oriented more towards novelty and advancing the field, many times weighing it higher than money-making logic or performance scaling. In this attempt to work on the new problems, finding collaborators from pertinent domains, and seeking funding for the research projects are not the only obstacles they face. They also have to look for relevant data sources in the absence of in-house data, contrary to the corporate giants who usually have access.

One way to lay hands on the relevant data is to collaborate with industry researchers; however, that is not always feasible. In that scenario, the only way to overcome this hurdle is to collect a reasonable dataset by themselves. Although challenging, many renowned academic researchers have been finding creative ways to achieve that lately. Conquering this barrier allows them to be self-sufficient and gives them a chance to lead research in new domains. Another reason curating datasets is becoming a highly regarded skill within the research community is that it fosters transparency and encourages reproducibility of the results.

Importance in Industry

Globally business-oriented corporations are increasingly investing in Machine Learning as they realize the value technology adds to their product and business. In contrast to academia, the industry prioritizes profit-generating logic and high scale performance higher than novelty and advancement of theoretical knowledge. Therefore, most use cases involve utilizing learning algorithms' performance on a large scale to improve the user experience or revenue generation. Established organizations seldom have obstacles in obtaining computation power, hiring folks with expertise in corresponding domains, or accessing relevant data. The challenge, however, comes in scaling up their solutions to their massive user base. Any organization would have a lot of unstructured data available on its hands as raw logs; however, developing an efficient data processing pipeline remains a task. For creating such resilient pipelines, apart from the relevant technological knowledge, we also require the skills to identify and curate meaningful and unbiased datasets from a sea of unstructured data.

Concluding Remarks

There is a modern bloom of social networks, online shopping portals, blogs, video streaming platforms, and so many other knowledge and experience sharing platforms enabled with all kinds of media, be it textual, visual, or audio. Consequently, a vast magnitude of raw data is available via numerous sources nowadays. This work aims to provide us an in-depth guide about one of the most rudimentary aspects of Machine Learning - *Dataset Curation*. Dataset Curation often does not get its due limelight but has high relevance in academia and industry. We shall walk through the process of constructing robust datasets, like (Misra 2022; Misra and Arora 2019; Misra, Wan, and McAuley 2018), from scratch using Python with tools like BeautifulSoup and Selenium. So turn the page[1], and start curating datasets suited to cater to all the needs!

[1] Available at amazon.com/dp/B08RN47C5T

References

Misra, R. 2022. News Category Dataset. *arXiv preprint arXiv:2209.11429* .

Misra, R.; and Arora, P. 2019. Sarcasm Detection using Hybrid Neural Network. *arXiv preprint arXiv:1908.07414* .

Misra, R.; Wan, M.; and McAuley, J. 2018. Decomposing fit semantics for product size recommendation in metric spaces. In *Proceedings of the 12th ACM Conference on Recommender Systems*, 422–426. ACM.