Classical ML - Data Enhancement Opps via LLMs

Many classical ML algorithms can be improved with high quality data, with LLM assistance.

colum.mccoole@analect.com
colum-mccoole-746b946a
analect/technical-docs-hierarchy

Docs: Level 1 High-level Concepts

Topic Navigation

How does this topic relate to other relevant content in this space.

Datacentric AI: Enhancing Pre-Existing Data

Traditionally, Artificial Intelligence (AI) development has been heavily model-centric. This means that the primary focus was on building and optimizing complex algorithms and model architectures to achieve better performance. Data was often seen as a fixed resource used to train these models.

Datacentric AI, on the other hand, flips this paradigm. It emphasizes the critical role of data quality and consistency in achieving superior AI performance. Instead of solely tweaking models, datacentric AI focuses on systematically improving the data used to train them.

Key Principles of Datacentric AI:

- **Data quality matters more than model complexity:** Small, high-quality datasets can often outperform large, noisy datasets.
- **Iterative data improvement:** Data is treated as a dynamic asset that can be continuously improved through systematic processes.
- **Focus on data consistency and labeling:** Ensuring data is accurate, consistent, and labeled correctly is paramount.
- **Tools and techniques for data engineering:** Specialized tools and techniques are employed for data cleaning, augmentation, and transformation.

LLMs Remain a Small Part of a Larger Ecosystem

Generative AI is merely an emerging solution in a larger established ML landscape. The past decade was one dominated by supervised learning (as evidenced by the green bubble in Figure 3). Within this realm, really large neural-networks demonstrated improved performance with size and so big-tech companies focused their efforts on assembling ever-larger training datasets to train these models.

The emerging trend from the past few years is about adding Generative AI to the mix (orange bubble in Figure 3). The core of generative AI is using supervised learning (input/output mappings) to predict the next word, but these models are becoming ever-more powerful at a host of general-purpose applications, including aiding with labelling data-sets, to enable pre-existing proven algorithmic solutions in the supervised learning domain.

Both classical ML methods and LLMs will each play a part in evolving AI applications. While LLMs have captured the lime-light, they can be expensive to train and infer results, when pre-existing ML algorithms can often be a much faster and cost-effective solution. Finding that mix will be key.

AI as a Collection of Tools

Within AI, let's focus on machine learning (ML), which are AI systems that are able to perform tasks as a result of a ***learning*** process that relies on data. ML is best thought of as being on a continuum with traditional statistical methods, rather than its own field.

These solutions are general-purpose in nature and so have the ability to be used across a broad set of business problems.

Huge value remains to be created using supervised learning. Layered on top of that, Generative AI brings even more opportunities.

1. **Supervised Learning** - the task of the ML algorithm is to infer the value of a predefined target (or output) variable based on known values of feature (or input) variables. The existence of labelled data (ie data with known values for the target in question) is a prerequisite for supervised learning. This learning process is also referred to as model-training. Inferences can then be performed on unlabelled data with the trained model. It looks to solve mainly *regression* (numerical prediction) and *classification* (categorical prediction) problems.

2. **Unsupervised Learning** - involves the identification of patterns and relationships in data without there being a pre-defined relationship of interest (ie. no labels). This approach can solve problems such as:
   - **Cluster analysis**, where the aim is to group units of observation based on similarities / dissimilarities between them. (eg. customer segmentations). Differs from classification, as categories aren't predefined.
   - **Association analysis**, where the goal is to identify salient relationships between variables within a dataset. (eg. customers interested in X also tend to be interested in Y)

3. **Reinforcement Learning** - relies on an 'agent' exploring an environment. The learning process relies on a reward function that provides feedback on the actions taken. The agent aims to maximise its reward and thus improve its performance through an iterative process of trial and error. Examples include problems in:
   - robotics or game playing.
   - Applications in finance might include trading or dynamic pricing.

4. **Generative AI** - is a class of models that creates content from user input. Generative AI can take a variety of inputs and create a variety of outputs, like text, images, audio, and video. It can also take and create combinations of these.

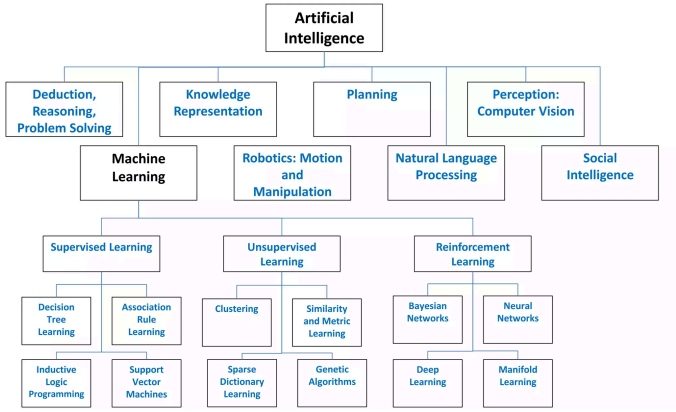

Figure 1. Nazre and Garg - https://www.slideshare.net/ajitnazre/deepdive-in-aiml-venture-landscape-by-ajit-nazre-rahul-garg

Input (A)	Output (B)	Application
Email	Spam? (0/1)	Spam filtering
Ad. user info.	Click? (0/1)	Online advertising
Image, radar info.	Position of other cars	Self-driving car
Ship route	Fuel consumed	Fuel optimisation
Image of phone	Defect? (0/1)	Visual inspection
Restaurant reviews	Sentiment (pos/neg)	Reputation monitoring

Figure 2. Supervised Learning Use-cases (Andrew Ng) - Derived from https://www.youtube.com/watch?v=5p248yoa3oE

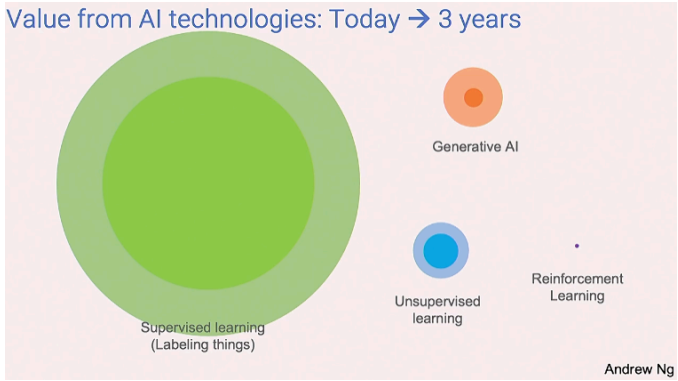

Figure 3. Andrew Ng's assessment of growth (shaded part) in fields of ML. See Fig. 2 for link to source presentation