ML Best Practices

Machine learning is a complex and multifaceted activity that requires the combination of a number of success factors in order to work. In order to execute machine learning well, it is important to have a good understanding of the processes and variables that frequently come into play together. Here I document and lay out some of the best practices and ML processes that I have learned and continue to learn during my career.

A personal bugbear of mine is when people refer to machine learning as "an art rather than a science". It is true that machine learning involves random-seeming aspects that could be conflated with magic or wizardry. There are a lot of ways that poorly thought-through or executed machine learning projects can go wrong. However, there are well established protocols and principles that can guide a project towards success and we can execute projects in a risk-averse, incremental way allowing us to fail fast before a large amount of time and energy has been spent on a project.

Key Principles

Garbage In; Garbage Out

Machine learning is all about trying to infer statistical relationships between data. There are a lot of reasons why this might not work well which we will explore in depth. It's always important to remember that if you have bad data or a flawed hypothesis, your model is not going to perform well. Conversely, if your model is performing really well (perhaps "*too good to be true*" *levels of well*) this can be a red flag that something is wrong with your data or your process which leads into...

Professional Pessimism

I love this phrase that I picked up during my career stint as a QA Engineer. If something appears to be too good to be true, it probably is. Always question your results and double check your working. Typically if a machine learning model is achieving high-90% in appropriate performance metrics, it is worth your time to investigate whether something weird is happening. You may have training data in your test set (i.e. your model got hold of the exam answer sheet) or your model may have found a short cut. For example, if all the positive examples of fractured ribs come from x-rays of women and all the negative examples come from x-rays of men then the model may have just learned to separate male and female anatomy rather than what a fractured rib looks like.

Keep it Simple, Stupid (KISS)

As tempting as it might be to want to use the latest and greatest ML models, always start with the simplest approach possible and disqualify these approaches based on sound theoretical rationale or empirical experimentation:

- The more complex a modelling approach, the more uncertainty we invite into the process.
- Complex models are significantly more expensive to train and run and scaling them to run quickly on a large amount of data is a challenge.

Keep Receipts

Machine Learning is usually an iterative process. As you move through this process it is important to understand where you came from and why at all times and to be able to refer back to earlier results and hypotheses. Furthermore, in a business setting, you will likely be asked questions about how work that you did days or weeks ago compares to your current iteration of thinking.

Keeping receipts is about using simple tools (good note taking and ticket hygiene) in combination with more specific purpose-built AI and ML tooling to ensure that you can answer these questions at all times.

Deep Learning Best Practices

For deep learning I like to work with torch. My opinion on how torch code should be written aligns quite closely with this styleguide

---

Revision #8
Created 2022-11-05 06:42:16 UTC by James
Updated 2024-02-11 16:26:35 UTC by James