Applied statistics and Machine learning in Python with subsurface applications

Enrico Riccardi, University of Stavanger

Mar 2, 2024

Who to talk to?

We established that machine learning and artificial intelligence is a huge field. The consequence is that one need to know how to approach the field in order to get the proper guidance.

As a parallel, if you are sick, you need to contact the right person. Contacting a surgeon will most likely result in a waste of time, your and her/his. Yet surgeons are between the most competent people in medicine.

Who does what

Different disciplines and thus actors can be involved:

- Data Scientists: They specialise in analysing and interpreting complex datasets using a variety of techniques, including machine learning.

Data scientists can help you understand whether your problem can be addressed through data analysis and machine learning, and they can develop models to solve it.

- Computational Engineers: These professionals specialise in designing, building, and implementing machine learning models.

They have a strong background in software engineering along with expertise in machine learning algorithms.

- AI Researchers or Specialists: If your problem is highly specialised or requires cutting-edge machine learning techniques,

consulting with AI researchers or specialists in the specific field of

Who does what

Different disciplines and thus actors can be involved (continue):

- Industry-Specific Experts with ML Background: Depending on your problem, experts in a particular industry who have experience with machine learning can provide valuable insights.

For example, for a healthcare-related problem, a bioinformatician or medical informatics professional with machine learning experience could be ideal.

- Consultancy Firms specialising in AI and ML: There are firms that specialise in providing AI and ML solutions to a range of problems across different industries.

They can offer a *team* of experts to work on your problem.

- Academic Institutions: Universities and research institutions often have departments or labs dedicated to machine learning and AI research.

Why is it so complicated?

A problem to be addressed by advanced statistical methods needs to be sectioned into its components.

- Data acquisition, data curing/wrangling, data preprocessing and filtering,

Domain expertise, computational engineers.

- Model selection, model training, testing and validation.

Data scientists/Engineers.

- Output stability and reliability.

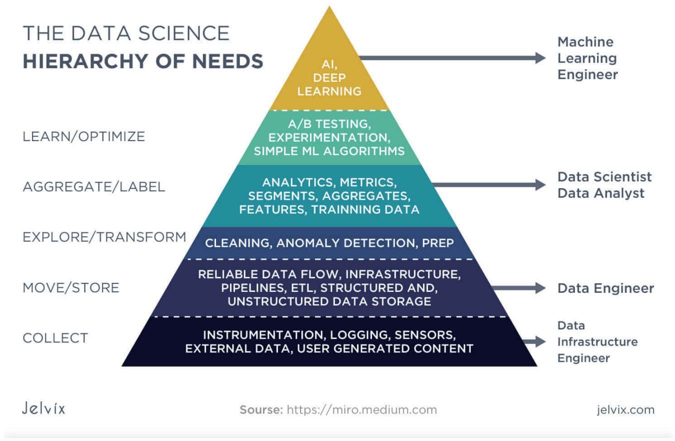
Data scientist.

Bias in approaches selection

As ML users often specialise in one or more methods, there is a consequent propensity to apply such a method into the largest set of applications possible.

While this is humanly understandable, it is not the best practice and can lead to bad solution.

Hierarchy of needs


THE DATA SCIENCE
HIERARCHY OF NEEDS
AI, DEEP LEARNING
A/B TESTING, EXPERIMENTATION, SIMPLE ML ALGORITHMS
ANALYTICS, METRICS, SEGMENTS, AGGREGATES, FEATURES, TRAINNING DATA
CLEANING, ANOMALY DETECTION, PREP
RELIABLE DATA FLOW, INFRASTRUCTURE, PIPELINES, ETL, STRUCTURED AND, UNSTRUCTURED DATA STORAGE
INSTRUMENTATION, LOGGING, SENSORS, EXTERNAL DATA, USER GENERATED CONTENT
LEARN/OPTIMIZE
AGGREGATE/LABEL
EXPLORE/TRANSFORM
MOVE/STORE
COLLECT
Machine Learning Engineer
Data Scientist
Data Analyst
Data Engineer
Data Infrastructure Engineer
Jelvix
Sourse: https://miro.medium.com
jelvix.com

Black Box approaches

If the result is correct, do you care that it is so for the right reason?

Do you even care? Maybe not and it is good like that.

Why should you?

Handling a Black Box

If your intent is not to become a ML developer, you will have to accept to use a black box approach.

We do not know how a smart-phone works, yet we (assume to) know how to use it!

Is it useful? Is it safe? Is it ethical?

Opening a Black Box

But if you need to develop an insight, consider interpretable AI methods: it is a whole new exciting field that is now emerging.

Interpretable AI can be pivotal for ML development, because it allow the consideration of different ethical concerns.

Garbage in, Garbage out

This is one of the fundamental baselines of ML. It doesn't matter what approach one uses. If data are garbage, model prediction will be garbage. Collecting data hoping that something will come out is often the perfect recipe for disasters.

No free lunch theorem

The theorem states that all optimization algorithms perform equally well when their performance is averaged across all possible problems.

- The performance of all optimization algorithms are identical, under some specific constraints.
- There is probably no single best optimization algorithm or machine learning algorithm.

If one algorithm performs better than another algorithm on one class of problems, then it will perform worse on another class of problems

Data properties

- All starts from data: what are data-properties?
- Are there such things as good data and bad data?

Main lesson (Exam question)

- Data DO NOT always have value.

- TRASH in TRASH out

MetaData properties

Data without metadata are just numbers (i.e. if they are integers, they are still good to play lottery)

Metadata can be pretty much anything. Depending on the application, we can distinguish between:

1. Descriptive : used for discovery and identification. It includes elements such as title, abstract, author, and keywords.
2. Structural : describe how compound objects are put together. It describes the types, versions, relationships, and other characteristics of digital materials.
3. Administrative : to help manage a resource, like resource type, permissions, and when and how it was created.
4. Reference : to indicate the information about the contents and quality of statistical data.
5. Statistical : (or process data), may describe processes that collect, process, or produce statistical data.
6. Legal : creator, copyright, licensing.

MetaData aim

These characteristics shall all be considered when constructing data repositories.

They are a must for:

- Code repositories
- Data repositories

Please click on the link and explore. Those are just some of the open (scientific) repositories. The aim/hope is to allow people to extract information. How to make value from that information... is another story.

MetaData for sharing and re-use

More considerations:

- Metadata is more and more important in a digital open world.
- Researchers and automatic algorithms would benefit from importing data directly.
- FAIR research is an important part of Open Science revolution (Findable, Accessible, interoperable, Reusable)
- New applications, business, discoveries can be thus enabled.
- ChatGPT, Bard, Gemini, and all the LLMs are functional only thanks to this!

Super controversial

- Who would be responsible for them then?
- What is the advantage for who releases the data?
- Who gets the money for what?
- Copyright for data and/or for data processing?

Good examples

- Norwegian offshore directorate
- Norway Statistics
- World statistics
- Code repositories
- Data repositories

Git

Git is a distributed version control system that tracks changes in any set of computer files, usually used for coordinating work among programmers who are collaboratively developing source code during software development. Its goals include speed, data integrity, and support for distributed, non-linear workflows (thousands of parallel branches running on different computers). [Wiki]

Let's try to be more accessible.

Git is a computer program/tool to save and download files on a hosting server (e.g. GitHub and GitLab).

A distributed version control system

GIT

- Git facilitates users to track the various versions of files. It is not a necessary tool, but it can be very very helpful. Generally, the time spent to learn its syntax is well paid off

(do you remember to save some file like *manuscript_draft_v4.02_final_definitive_forreal_lastcomments_edite*
Exactly! Imagine to do that for a repository of files...)

- It permits to save and share the intermediate stages of a work in progress (which software is complete and always up to date?) in an accessible, consistent and structured way, allowing an effective version tracking. It allows retrieval of previous working versions, limiting the risk to overwrite useful files.

What is git actually for

> The tool is particularly useful for programmers working in teams or in projects whose outcomes can be used by others.

- Git helps to co-develop a code, test its functions and the compatibility of the various code sections.
- A long list of further possibilities became possible by git.
- Different software integration on development platforms, based on git, will help you to develop and co-develop your code.
- The platform GitLab and GitHub have a large set of functionalities to further support code documentation and public releases.
- Files can be disclosed to the public, becoming a great integration of your CV, showing what you are able to do in an open and accessible way.

Why should I care?

As the open libraries are exploding in numbers, you might need some criteria to assert the reliability of a project.

Unit test driven development!

That is taking full advantage of python object oriented structure.

Community

Good project are not only used by communities, but also **supported**

Git allows the development of projects without a clear lead. The community engagement is generally a desirable target to help developed to directly integrate feedbacks by users (and fix bugs).

Developing approaches

Different code editors are available to interpret python language.

- jupyter notebooks are mostly dedicated to learning (Markdown)
- ipython is for interactive coding (similar to R, Matlab, etc)
- python packages (.py) developing suites (debug possibilities and git integration)

Introducing code standards

When developing code, there are **guidelines** and best practices aimed at improving the quality, readability, and maintainability of a code.

There are different levels of coding quality, mostly depending on the code intended usage (and developer skills).

- Private codes can be whatever (Cpt. Obvious)
- Public packages shall use a 'Golden code standards' such to be used and eventually supported by communities.

Golden code standard

Principle of 'clean coding':

1. Readability and Clarity

A good code shall be possible to read as when reading a book

1. Structure and object oriented

A code shall be composed by objects, each of them connected in the less redundant way possible.

1. Consistency and Style

Variable naming, function naming and classes naming has to be consistent.

1. Documentation

Each file, each function and each class shall contain the relative description of its aim and its usage

1. Testing

[illegible]