Fundaments of Machine learning for and with engineering applications

Enrico Riccardi[1]

Department of Mathematics and Physics, University of Stavanger (UiS).[1]

Sep 3, 2025

University of Stavanger

Let me first introduce US

Enrico Riccardi
email: enrico.riccardi@uis.no

Pål Østebø Andersen
email: pal.andersen@uis.no

Let me first introduce myself

Enrico Riccardi email: enrico.riccardi@uis.no office: KE-E-542

Before to be here...

- Chemical engineer graduated from the Politecnico di Torino
- PhD at M&ST university of Rolla, Missouri, USA
- Post Doc at TUD Darmstadt, Germany
- Researcher at NTNU, Trondheim, Norway
- Post Doc UiO, Oslo, Norway
- Assoc. Prof. at Uis, Stavanger, Norway
- Prof. at Uis, Stavanger, Norway from 28.8.2025!!

Course objectives

Understanding of

- data sources and consequent data properties.
- data analysis/machine learning approaches outcomes.
- sensitivity analysis
- predictive modeling
- multivariate data analysis
- machine learning techniques application
- ensemble methods
- Bayes approach
- visualization and reporting

Metalearning

- Pedagogic empiricism teaches us that good learning happens in social contexts where the students are actively participating in the classroom dialogue
- In general, the number of questions, and the dialogue in general, drops significantly when lectures are recorded
- Streaming tends to lead to empty classrooms and higher failure rate.

Teaching method

!**Active Learning**!

- There will be a combination of lectures and tutorials.
- Tutorial and hands-on will be presented during the course.
- Study groups are strongly encouraged.
- In class discussions are encouraged at any stage.
- Flip classroom approach: problem first (when possible).
- Feedback expected from you!
- Note: each teacher, even if coordinated, will have a different approach/expectations.

Python

The tutorials and hands-on will be mostly based on python.

- An introduction to python will be provided during the course. Yet, this is not a Python programming course!
- You are **strogly advised** to construct your own Git repository (and to keep it in order).
- Group projects will have different roles.

A personal considerations on generative AI

With great power comes great responsibility (Spider-Man)

LLMs can and shall be used. As their development is surging in the last years, their help in writing code and reports is undeniable. Students shall learn how to master these tools. Yet, while their usage is encouraged, the risk of excessively rely on them **DO** hinder learning.

If you want to be a student (i.e. whom is learning), you are encouraged to take the responsibility in delivering original output.

Why are we here? Why this course?

- Decisions are the hinge. What influences your decisions has value.

- Better results come from better decisions.

STATISTICAL argument

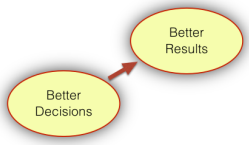
- Machine learning can be a useful ONLY when it helps making better decisions.

Decisions, decisions, decisions...

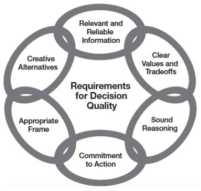
Life lesson?!

Life is a sum of all your choices (Albert Camus)

The only way you can purposefully influence your life, your family, your organization, your country or your world is through the decisions you make.



Uncertainty is fed to decisions



- In this course we will learn to use data and models (ML) to make better decisions.
- Each application will be a mere exercise of the concepts we will introduce here.

Do not try this at home!

- Please Do not use a technical approach for love matters.

Uncertainty

It is present in every stage of life (at each decision).

You would thus need to:

1. Rationalize uncertainty (qualify)
2. Quantify uncertainty
3. Make decisions under uncertainty
4. Operate under uncertainty

Uncertainty or Probability?

Our aim here is to provide good guidance on how to link data, models and output to value creation.

- First we need to understand uncertainty and probability, and the difference between the two.

- Spoiler: Probability is the language of uncertainty!

- We will analyse, quantify and structure models as a function of uncertainty.
- Clarity of language in probability is one of the hallmarks of decision analysis and of modeling.

Data properties

- All starts from data: what are data-properties?

- Are there such things as good data and bad data?

Life lesson (or exam question, same thing ;))

- Data DO NOT always have value.

- TRASH in TRASH out

Spatial and Temporal Data

Statistics is collecting, organizing, and interpreting data

//

Spatial and temporal statistics is a branch of applied statistics that emphasizes

1. the context of the data,
2. the spatial and time dependent relationship between data
3. the different relative value and precision of the data.

Quantifying uncertainty

On the data sources side

- Confidence intervals
- Relevance
- Significance
- Correlation
- Causation
- Data Filters
- Biases identification

Quantifying uncertainty

On the modelling side

- Regression
- Principal components
- Decision tree (random forests)
- Neural network
- Clustering
- Performance metrics

Fields of application

- Spatial estimation of energy and mineral resources
- Weather modeling: from aviation to agriculture
- Maintenance forecasting
- Commodity, currency, stock and financial markets
- Market analysis
- Risk analysis
- ... and much much more!

Hard and soft modeling

Models allow us to predict 'the future', or describe the past and present (what is the present...?)

Last life lesson for today

Models are always wrong, but some are useful. (George Box)

Three main families:

1. Hard models (physics)
2. Soft models (statistic)
3. Machine learning

Hard modeling

- Based on an accurate physical description of the system and mathematical modeling (e.g. differential equations). Hard models are often deterministic.
- Hard modeling methods usually use optimization methods to find out the best values for the parameters of the model.
- Hard modeling is preferable in laboratory experiments, where all the variables are controlled and the physicochemical nature of the dynamic model is known and can be fully described using a known mathematical model.
- Hard modeling, if successful, usually gives better understanding of a system and better extrapolations. Wrong assumptions often lead to non-sense results.

Soft modeling

- Soft-modeling describes systems without the need of an *a priori* physical or (bio)chemical model postulation. They are **data driven** models.
- Soft models are much easier to make than hard models.
- Soft modeling can be used to understand complex relationships.
- Soft modeling needs (much) more data than hard-modeling.
- Soft models have a poor extrapolating capabilities (compared with hard-modeling)

How to create hard models?

After understanding the problem to be solved we need to:

1. Link mathematics to physics.
2. Define boundary conditions and constitutive equations.
3. Make tons of assumptions.
4. Solve the constitutive equation in space and time.
5. Check solution stability and sensitivity analysis.
6. A long set of judicious approximations have to be taken.
7. It is hard (but we are engineers!).
8. Get quite some money for the awesome job.

How to create soft models?

After understanding the general problem to be solved we need to:

- Determine a suitable **numerical description** .
- Choose a suitable **model** to which parameters are fitted.
- Train, test, validate the model.
- Perform **data analysis** with chosen method(s).
- Link predictions with expectations.

Hard models vs Soft models

Requirements

Deterministic:

- physics and expert knowledge
- integration of various information sources
- very complicated

Statistical:

- quick
- uncertainty assessment
- data driven approach
- physics can be included
- stochastic modeling

Hard models vs Soft models

Behaviour

Deterministic:

- predictable
- defined error

Statistical:

- outcome uncertainty
- undefined error
- sampling resolution issues

Spatial and Temporal Modeling

It is a branch of statistical analysis and model that uses spatial and time dependent data.

- Only a subset of statistical models can be fed with time dependent data

(most standard statistical method assume independent, identically distributed, data)

- Spatial and time related data come at a different range of scales

Frequency of data collection can be dependent of time and space, resulting in different representativity of a sample.

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