Fundaments of Machine learning for and with engineering applications

Enrico Riccardi[1]

Department of Energy Resources, University of Stavanger (UiS).[1]

Jan 14, 2025

University of Stavanger

Let me first introduce US

Enrico Riccardi
email: enrico.riccardi@uis.no

Reidar Bratvold
email: reidar.bratvol@uis.no

Remus Gabriel Hanea
email: rhane@equinor.com

Let me first introduce myself

Enrico Riccardi email: enrico.riccardi@uis.no office: KE-E-301B

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

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
- Byes approach
- visualization and reporting

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 apprach: problem first (when possible).
- Feedback expected from you!
- Note: each teachers, 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.

For the next 3 weeks, you are welcome to join BIO510, where I will be providing an intro to Python.

A personal considerations on generative AI

With great power comes great responsability (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. Studends shall learn how to master these tools. Yet, while their usage is encouraged, the risk of excessevily rely on them **DO** hinder learning.

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

Why are we here? Why this course?

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

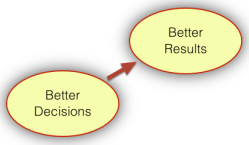
- Better results come from better decisions.
  STATISTICAL argument
- Machine learning can be useful ONLY when it helps taking better decisions.

Decions, decisions, decisions...

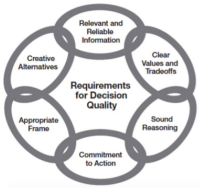
Life lesson?!

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.



Uncertainity is fed to decions



- 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 here introduce.

Do not try this at home!

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

Uncertainity

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

You woud thus need to:

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

Uncertainity or Probability?

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

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

- Spoiler: Probability is the language of uncertainity!

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

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 userful. (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 leads 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
- uncertainity assessment
- data driven approach
- phisics can be included
- stochastic modeling

Hard models vs Soft models

Behaviour

Deterministic:

- predictable
- defined error

Statistical:

- outcome uncertainity
- 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 into different rappresentativity of a sample.

Data concepts, Data types and Models

Population

Exaustive, finite list of properties of interest over area of interest.

Generally the entire population is not accessible

Samples/experiments/instances

The set of values and location that have been measured.

How many experiments are needed?

Features

The values to be measured for each sample/experiment/instance.

How many features are needed?

Variables and Features, Labels and Instances

Predictors = input variables, X_1, ..., X_M

Response = output variables

Error

Deviation from ... exact value (or expected value, mean value, trend...?)

Errors without definitions are just number.

Predictor and Response Features

Given a model $Y = f(X_1, ..., X_M) + e$

!Here and error! But is it even an error?

Descriptive and Predictive statistics

Estimation

- Process of obtaining the best value or range of a propery in an unsampled location
- Local accuracy takes precendence over global spatial variabiltiy
- Not appropiate for forecasting

Inference

- Predict unseen samples given assumptions about the population
- Test with a pre-trained model (ML definition)
- Generality versus Accuracy

Simulations

Process of obtaining one or more values of a property

- Improved Global accuracy
- Better property distributions

Why simulations then?

- We need to capture the full distribution of properties, extremes matter!
- We need more realistic models.

Why not?

- High dimensionality level
- Computationally expensive
- Convergence limitations
- Costitutive equations need to be rather accurate.

Uncertainity Modeling

Given a model, Generate multiple simulation to represent uncertainity

- Realizations: for the same input parameters, different random numbers.
- Scenarios: different input parameters.

Sampling rapresentative.

Random sampling

Each item of the population has an equal chance of being chosen.

- Very expensive
- Mostly not interesting
- Gives some global properties

Bias sampling

Selection of data is (arbitrarely) distorted

- Sample probability bias has to be corrected for
- Might not capture the global picture

Cognitive biases

- anchoring: The first bits are overconsidered
- availability: over-estimating the importance of info
- bandwagon: P increases with the number of people holding a belief
- blind spot: not seen biases
- choice supporting: commitment/decision dependent
- clustering illusion: seeing patterns in random events
- confirmation bias
- concervatism bias
- Recency bias
- Supervision bias
- Many many more!

Bias DO NOT cancel out! They sum up (or multiply?)

Data dimensionality

1 D

logs

2 D: maps

Quite limited but great for visualization

3 D

3d maps, sysmic cubes. More informative mostly ok in digital formats.

4 D

Trajectories

x D

Data realm

Types of data

- Categorical / Nominal (classes)
- Categorical / Ordinal
- Continous / Interval (e.g. Celcius)
- Continous / Ratio
- Discrete: binned/groupped data
- Hard data: direct measurements
- Soft data: indirect measurements, very uncertain come from model
- Primary data: variable(s) of interest
- Secondary data: descriptors
- Collective variables
- Latent variables uncertain effect on modelling outcome

Uncertainity and Probability

Probability: there is not science more worthy in out contemplations nor a more useful one for admission to our system of public education

The theory of probabilities is at the bottom of nothing but common sense reduced to calculus.

Clarity test. Beer drinker?

Rain in Stavanger?

Uncertainity

- Def 1: Not knowing if an event is true or false. (Useful)
- Def 2: Things that cannot be measured. (Not useful)

Probability is how Uncertainity is quantified!

- Clarity test
- Assign a number between 0 and 1 to our degree of belief
- Error definition

Sentence also good for fortune cookies

Uncertainity is the only certainity