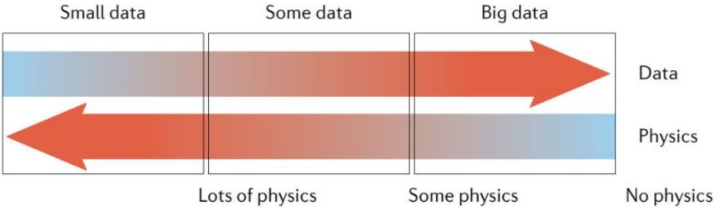
Fundaments of Machine learning for and with engineering applications

Enrico Riccardi[1]

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

Jan 15, 2025

Data vs Physics



Uncertainty

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

Probability is how Uncertainty is quantified!

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

Sentence also good for fortune cookies

Uncertainty is the only certainty

Uncertainty and Probability

Random quotes

- 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.

What is Statistics

Clarity test. Beer drinker?

Rain in Stavanger?

Data properties

1 D

logs

2 D: maps

Quite limited but great for visualization

3 D

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

4 D

Trajectories

x D

Data realm

Types of data

- Categorical / Nominal (classes)
- Categorical / Ordinal
- Continuous / Interval (e.g. Celsius)
- Continuous / Ratio
- Discrete: binned/grouped data
- Hard data: direct measurements
- Soft data: indirect measurements, very uncertain
- Primary data: variable(s) of interest
- Secondary data: descriptors
- Collective variables
- Latent variables

Descriptive and Predictive statistics

Estimation

- Process of obtaining the best value or range of a property in an unsampled location
- Local accuracy takes precedence over global spatial variability
- Not appropriate for forecasting

Inference

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

Variables and Features, Labels and Instances

Population

Exhaustive, 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 numbers.

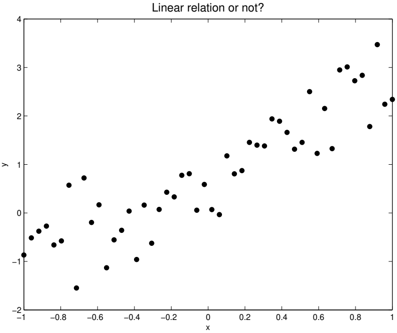
Predictor and Response Features

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

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

Finding a suitable model

Soft modeling is in most cases based on **multivariate statistical methods**. Many of these methods may be viewed as sophisticated ways of performing curve fitting to data.



What would be the best model?

- Straight?: $y(x) = ax + b$
- Parabolic?: $y(x) = ax^2 + bx + c$
- Trigonometric?: $y(x) = asin(x) + bcos(x)$

Uncertainty Modeling

Given a model, Generate multiple simulation to represent uncertainty

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

Sampling representative.

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 (arbitrarily) distorted

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

Cognitive biases

- anchoring: The first bits are over-considered
- 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
- conservatism bias
- Recency bias
- Supervision bias
- Many many more!

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

Simulations

Process of obtaining one or more values of a property

- Improved Global accuracy
- Better property distributions

Why simulations then?

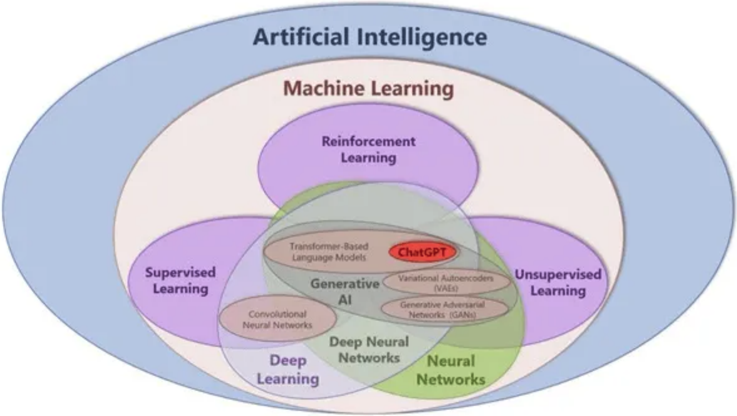
- We need to capture the full distribution of properties, extremes matter!
- We need more realistic models.

Why not?

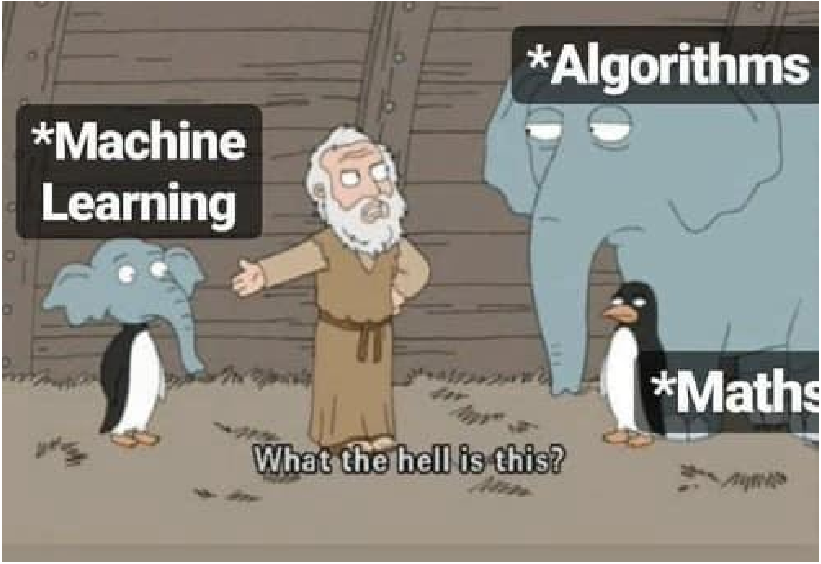
- High dimensionality level
- Computationally expensive
- Convergence limitations
- Constitutive equations need to be rather accurate.

Statistics, Machine learning or Artificial intelligence?

What is the main difference between the three fields?



How Machine Learning Started?



Statistics

Let’s start from the definition

- Statistics (origin "description of a state/country") is the discipline that concerns the collection, organization, analysis, interpretation, and presentation of data.
- It is conventional to begin with a statistical population or a statistical model to be studied. Populations can be diverse groups of people or objects such as "all people living in a country" or "every atom composing a crystal".
- Statistics deals with every aspect of data, including the planning of data collection in terms of the design of surveys and experiments.[Wikipedia]

Machine learning

Definitions:

- Machine learning is a branch of artificial intelligence (AI) and computer science which focuses on the use of data and algorithms to imitate the way that humans learn, gradually improving its accuracy. [IBM]
- Machine learning (ML) is a field of study in artificial intelligence concerned with the development and study of statistical algorithms that can learn from data and generalize to unseen data, and thus perform tasks without explicit instructions. [WIKI]
- Machine learning is a subfield of artificial intelligence that uses algorithms trained on data sets to create models that enable machines to perform tasks that would otherwise only be possible for humans, such as categorizing images, analyzing data, or predicting price fluctuations. [Coursera]

Machine learning

One technical definition

Machine learning is a set of computer based statistical approaches that aim to minimise the loss function to maximise inference accuracy. [Enrico, 5.2.2024]

The loss function is the actual engine in machine learning.

Loss function

It quantifies the difference between the predicted outputs of a machine learning algorithm and the actual target values.

Artificial intelligences

And more definitions:

- Artificial intelligence is the intelligence of machines or software, as opposed to the intelligence of humans or other animals. It is a field of study in computer science that develops and studies intelligent machines. [WIKI]
- Artificial intelligence (AI) is the theory and development of computer systems capable of performing tasks that historically required human intelligence, such as recognizing speech, making decisions, and identifying patterns [Coursera]
- It is the science and engineering of making intelligent machines, especially intelligent computer programs. It is related to the similar task of using computers to understand human intelligence, but AI does not have to confine itself to methods that are biologically observable. [IBM]

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!

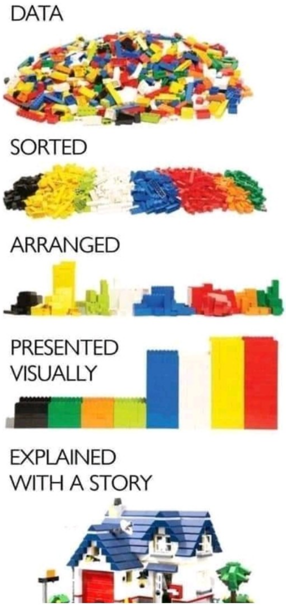
Super controversial

- 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

Data


DATA
SORTED
ARRANGED
PRESENTED VISUALLY
EXPLAINED WITH A STORY

Representation

A representation should **capture** the nature of the subject being studied.

Example: If you want to evaluate the 3D structure of a wind turbine, a set of descriptors an be:

1. Blade length
2. Turbine height
3. Geographical position
4. Output power
5. Wind direction

which are two decimal numbers, a 2d tuple, a 1D time series and a 2D time series (or 3D even).

Comparability

Same meaning **represenations** for different objects (inputs).

Discussion point!

How do we compare two wind turbines accounting for the 5 variables previously introduced?

Data properties

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

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

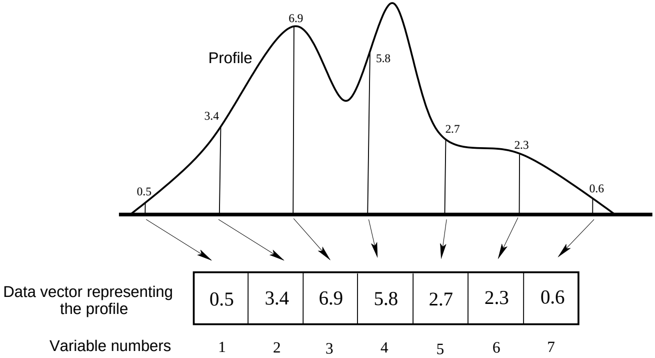
Life lesson (or exam question, same thing ;))

- Data DO NOT always have value.

- TRASH in TRASH out

Sampling point representation (SPR)

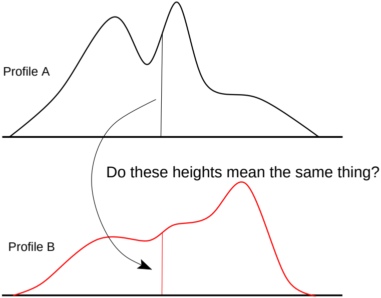
- An intuitive way to represent curves and spectra is the **sampling point representation**.
- We sample at regular intervals where each sample point is represented by a variable



Data vector representing the profile	0.5	3.4	6.9	5.8	2.7	2.3	0.6
Variable numbers	1	2	3	4	5	6	7

Sampling point representation (SPR)

- SPR is useful until point i in a curve has the same meaning of the point i in another curve.



- Which parts of the profiles or shapes are comparable, i.e. have the same meaning?

Data structures

Given a representation, it is then needed to decide on a suitable **data structure** for the problem.

Definition

A data structure is a way of storing and organising data in a computer so that it can be used effectively.

Typical data structures used in data analysis are:

- Data points
- Arrays (vectors, matrices, N-mode (way) arrays)
- Graphs (trees)
- Databases

Workflow

Data has to be prepared with these steps in mind

1. Plan experiments: Use experimental design to set up experiments in a *systematic* way
2. Pre-processing: Is there systematic variation in the data which should be removed Can cross-checking/validation procedures be designed?
3. Examine the data: Look at data (tables and plots). Strange behaviours? Smooth behaviour? WARNING!
4. Define desired model outcomes (speed, accuracy, false positive/negatives rate)
5. Estimate and validate model: What do the results tell us? Is the generated model general (valid for future sampling)?
6. Apply model to unknown samples

Spatial and Temporal Data

Statistics is collecting, organising, and interpreting data

Spatial and temporal statistics is a branch of applied statistics that emphasises:

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

Actual data

The data matrix is an extremely common data structure.

$$X = \begin{bmatrix} 95 & 89 & 82 \\ 23 & 76 & 44 \\ 61 & 46 & 62 \\ 49 & 2 & 79 \end{bmatrix}$$

In python these can be saved as

- lists (vanilla python)
- numpy.arrays
- pandas dataframes

Nomenclature

There are different conventions. Commonly we will construct data matrix such that:

- Rows are called instances, objects or samples.
- Columns are called features, variables.

One can think of each row to be an experiment, and the rows its properties. Each row (experiment, object, sample, ...) is thus a list of values, one for property.

Note

Mathematically speaking, this is just a notation. As long as one keeps track and is consistent, columns can be used as rows and vice versa.

A quick example

Environmental measurements of rivers. The features (properties) can be:

- pH
- Temperature
- Concentration of pollutants
- Flow rate
- water speed

The experiments/observations/sample can be:

- Po
- Danube
- Rio delle Amazzoni
- Sjoa
- Atna