

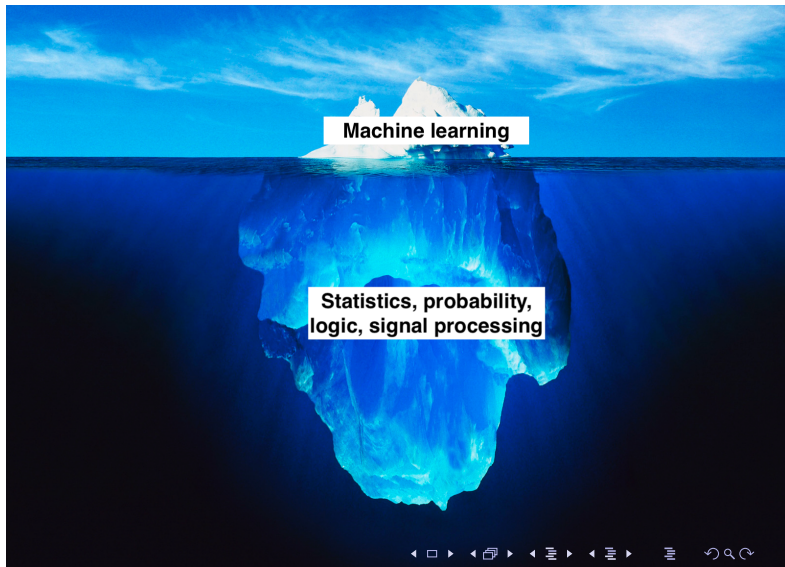
INFO-F-422: Statistical foundations of machine learning

Introduction

Gianluca Bontempi

Machine Learning Group
ULB Computer Science Department
`mlg.ulb.ac.be`

Statistical foundations of machine learning



Cool ML applications

Do you know some?

What have they in common?

- ▶ Large and increasing availability of big data (observations, variables) made possible by ubiquitous measurement systems (web, communications, gps, bank, sensors).
- ▶ Desire of extracting value from data
- ▶ Growing computing power.
- ▶ Awareness of the existence of some information pattern hidden somewhere within massive data.
- ▶ New adaptive, automated, intelligent, smart, applications.
- ▶ Prediction.
- ▶ Solving hard problems in a new manner.
- ▶ Decision support.

About the course

- ▶ Why is it a course for computer scientists?
 - ▶ Information.
 - ▶ Automatic improvement of computer capabilities.
 - ▶ Models, algorithms.
 - ▶ Simulations, programs.
- ▶ Requirements: Preliminary course on statistics and probability, programming skills.
- ▶ Exam: project (10 points) + UV assessment on theory questions (use of Python) (10 points).
- ▶ TP (Cedric Simar and Pascal Tribel):
 - ▶ introduction to the Python language but not to programming !
 - ▶ Hands-on
 - ▶ Real case studies, introduction to the project
- ▶ Slides on UV page:
<https://uv.ulb.ac.be/course/view.php?id=122910>
- ▶ Free handbook in English (on UV page): use the form for comments/remarks.

Prof to students: UV messages

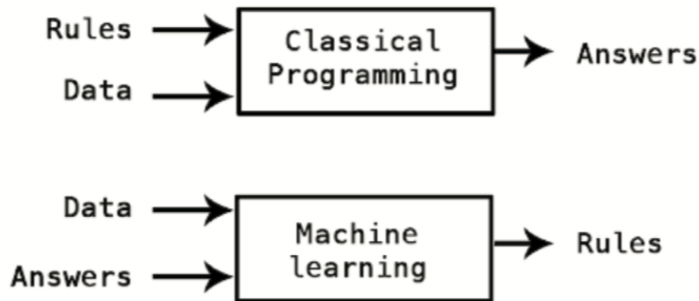
Students to professor:

- ▶ Teams
- ▶ email
- ▶ Student representative

Machine Learning: a definition

The process of converting experience (e.g. data) into expertise or knowledge (e.g. a model)

ML paradigm



Excerpt from *Deep Learning from R* book from F. Chollet

Who is a data scientist ?

Someone who knows statistics better than a computer scientist and codes better than a statistician...

Definition

Statistics is the study of how information should be employed to reflect on, and give guidance for action, in a practical situation involving uncertainty. (V. Barnett, "Comparative statistical inference", 1999)

This definition requires some clarification, specifically

- ▶ What is meant by uncertainty?
- ▶ What is meant by *situation involving uncertainty*?
- ▶ What is meant by *information*?

Some examples of uncertain situations

- ▶ A student tackling an exam.
- ▶ A doctor prescribing a drug.
- ▶ A wind farmer deciding where to situate the wind turbine.
- ▶ A football trainer deciding who will shoot the decisive penalty.
- ▶ A financial investor in the NASDAQ trade market.
- ▶ An employee facing the offer of a new job.

What is typical to uncertain situations?

- ▶ There is **more than one** possible outcome (e.g. success or failure).
- ▶ The actual outcome is **unknown** to us in advance: it is indeterminate and variable.
- ▶ Uncertainty is pervasive in our world.
- ▶ *...in this world there is nothing certain but death and taxes...(Franklin)*

Why are we interested to uncertainty?

- ▶ We would like to know what the outcome will be (e.g. will the drug be successful in curing the patient?)
- ▶ We want to decide on a course of action relevant to, and affected by, that outcome (e.g. where is the oil company going to drill?, how much should I study to pass the exam?).
- ▶ We will have to **take a decision anyway**.

Why probabilistic modeling?

- ▶ Construct a theory to guide behavior in a situation involving uncertainty requires a **formal model** of such situations.
- ▶ This requires a formal notion of uncertainty.
- ▶ In this course we will use the formalism of **probability** to represent uncertainty.
- ▶ A stochastic model is made of
 1. a **set of possible outcomes** of a phenomenon and
 2. a **probability distribution**, e.g. a mathematical function that gives the probabilities of occurrence/generation of events (sets of possible outcomes).
- ▶ Notions like (in)dependence, randomness, etc., has to be defined for distinguishing and characterising the different situations in terms of their degree of uncertainty.

Models and reality

- ▶ A model is a formal (mathematical, logical, probabilistic ...) description of a real phenomenon.
- ▶ A model is an **idealization** of a real situation. No mathematical model is perfect.
- ▶ A model makes **assumptions**.
- ▶ The adequacy of a model depends on how valid and appropriate are the assumptions on which it is based.
- ▶ A model can be used to make **deductions** and take **decisions**.
- ▶ The biggest concern is how to define an adequate model, either as a description of the real situation or to suggest a reasonable action.
- ▶ Different aspects of the same phenomenon may be described by different models (e.g. physical, chemical, biological...)

Truth is a model..

The most common misunderstanding about science is that scientists seek and find truth. They don't. They make and test models.. Making sense of anything means making models that can predict outcomes and accommodate observations. Truth is a model. (Neil Gershenfeld, American physicist, 2011)

but do not slip into relativism...

All models are wrong...

Some are useful ...

and above all...

You need one, you can't help it.

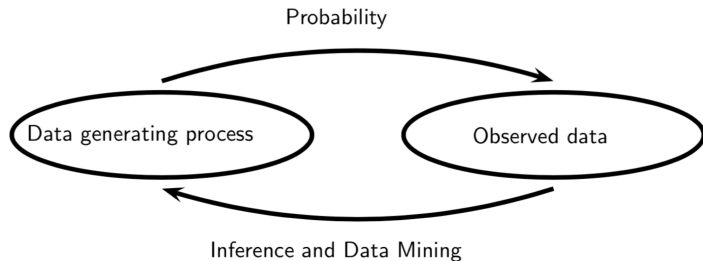
You 'd better choose one that is not too bad...

Two configurations involving a model

Consider

1. a real phenomenon P , e.g. epidemics spread.
 2. a probabilistic model M describing the number of expected infected per days.
- ▶ **Deductive or probabilistic configuration.** We use the model M to predict the number of hospitalised persons in the time interval $[t, t + \Delta t]$.
 - ▶ **Inductive or statistic configuration.** We collect measures in order to estimate the model M starting from the real observations.

Statistics and machine learning look backward in time (e.g. what model generated the data), probability looks forward (e.g. it simulates future behaviour).



Excerpt from Wasserman book "All of statistics"

- ▶ **Under the assumption that a probabilistic model is correct**, logical deduction leads to a description of the properties of data that might arise from the real situation.
- ▶ Statistics **reverses the deductive process**. It takes data from a practical situation and uses them
 - ▶ estimate the parameters of a parametric model,
 - ▶ suggest a model,
 - ▶ validate a guessed model,
 - ▶ generalize to similar settings.

Deduction and induction

- ▶ Deduction is a form of reasoning that works from the general to the specific (**top-down**), drawing **necessary conclusions** from the premises.
- ▶ Induction is a form of reasoning that works from the specific to the general (**bottom-up**), drawing **probable conclusions** from the premises.

Example rule of deduction

An example of deductive rule is

All university professors are smart. I am listening to an university professor.

So this professor is smart

- ▶ There is no possible way in which the premise can be true without the corresponding conclusion to be true. This rule is always valid since it **never leads from true premises to false conclusions**.
- ▶ Logic guarantees that if premises are true and the argument is valid then the conclusion is true.
- ▶ Logic is **not concerned with the truth of the premises**, it is concerned with the validity of the reasoning mechanism leading from the premises to the conclusions.

- ▶ **Logic is truth preserving but is not a source of new truths.**
- ▶ Does it mean that logical results are mere tautologies, i.e. already contained in the premises?
- ▶ Yes, but in the same sense that the intricate equations of the motion of SpaceX rocket are included in Newton's laws.

Example rule of induction

An induction rule looks like

Until now, all university professors I met, were smart.

I am listening to an university professor.

So this professor is smart

The premise could be true even though the conclusion is false. How to measure the reliability of this rule?

Black swan example: note that, *before discovery of Australia, people in the Old World were convinced that all swans were white.*

Deduction and induction

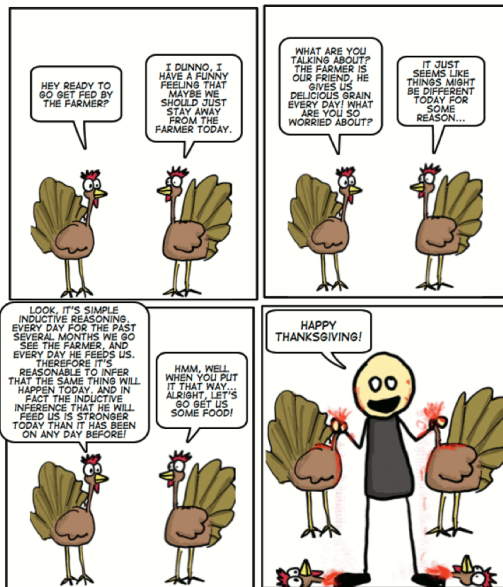
- ▶ Deduction is *truth (or content)-preserving*, i.e. conclusion is a logical consequence of (it follows from) the premises,
- ▶ Induction is *ampliative* (or content expanding), i.e. it has more content in the conclusion than in the premises. The truth of the premises is not enough to guarantee the truth of the conclusion.
- ▶ Models or universal statements (as well as knowledge) are not logical or **necessary** consequence of observational statements (or facts).

- ▶ David Hume (1711-1776) is a Scottish philosopher who studied the problem of induction from a philosophic perspective.
- ▶ *All reasonings concerning nature are founded on experience, and all reasonings from experience are founded on the **supposition** that the course of nature will continue uniformly the same* or in other terms that the future will be like the past.
- ▶ Any attempt to show, based on experience, that a regularity that has held in the past will or must continue to hold in the future will be circular.
- ▶ **All our science relies on a supposition!**

About inductive reasoning

- ▶ Though inductive reasoning is not logically valid, there is no doubt that we continuously carry out this kind of reasoning, and that our practical life would be impossible without it.
- ▶ The most obvious class of inductive claims are the ones concerning the future.
- ▶ Every human act relies on considerations about the future or estimations, though future is by definition not observable.
- ▶ Induction postulates what Hume calls the *uniformity of nature*: for the most part, if a regularity holds in my experience, then it holds in nature generally, or at least in the next instance.
- ▶ The truth of this principle cannot be logically demonstrated but the success of the human beings in using it is an empirical proof of its validity.
- ▶ *Could you logically prove that the sun will raise tomorrow?*

Russels's inductivist turkey



- ▶ Karl Popper (1902-1994) is generally regarded as one of the greatest philosophers of science of the 20th century.
- ▶ He agrees with Hume that there is no justification of inductive inference.
- ▶ Humans or scientists do not make inductions, they make conjectures (or hypotheses) and test them (or their logical consequences obtained by deduction).
- ▶ If the test is successful, the conjecture is corroborated but never verified or proven.
- ▶ Confirmation is a myth: no theory or belief about the world can be proven: it can be only submitted to test for falsification and, in the best case, be confirmed by the evidence for the time being.

Falsificationism

- ▶ It is not possible to logically deduce universal laws from **tons of observational statements**.
- ▶ However, the falsity of universal statements ("All swans are white") can be deduced from a **single observation**.



So what?

- ▶ The Hume skepticism and the Popper definition of science, suggest that a *tabula rasa* approach moving automatically from data to correct hypothesis is not possible: inverse ill-posed problem.
- ▶ Inductive bias (or assumptions) are needed to make the inductive leap.
- ▶ This is implemented by a process of hypothesis generation, validation and selection
- ▶ Machine learning complies with the Popper interpretation of science by proposing **a set of strategies for automatically generating hypothesis and validating them on the basis of empirical evidence.**
- ▶ No univocal (or optimal) way of proceeding from observations to models (no-free-lunch) and that any learning process relies on (implicit or explicit) assumptions.

May data alone speak for themselves?

Every time you rely your analysis on a dataset, you have already made a lot of assumptions:

- ▶ convenient number of observations.
- ▶ convenient number of variables (e.g. you did not forget any relevant variable).
- ▶ convenient way of collecting data (e.g. no sampling bias).
- ▶ data generation process (e.g. population) underlying the observations is invariant (identical distribution, i.e. no shift or nonstationarity).
- ▶ all observations are independent (e.g. the fact of observing something does not change the likelihood of observing something in the future).

Is it still a *tabula rasa*?