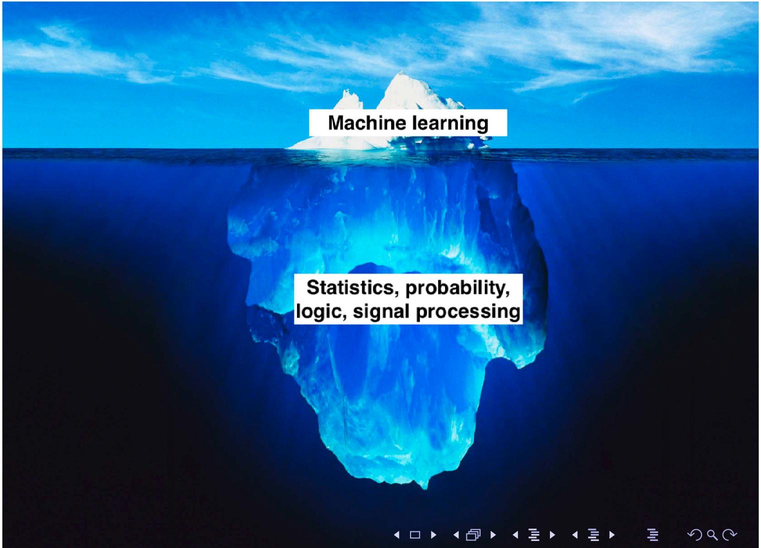
INFO-F-422: Statistical foundations of machine learning

Introduction

Gianluca Bontempi

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

Statistical foundations of machine learning



Cool ML applications

https://www.edureka.co/blog/machine-learning-applications/

- Traffic Alerts
- Social Media
- Transportation and Commuting
- Products Recommendations
- Virtual Personal Assistants
- Self Driving Cars
- Dynamic Pricing
- Google Translate
- Online Video Streaming
- Fraud Detection

Too many to mention...

What have they in common?

- Large and increasing availability of big data (samples, 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 complex masses of 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 R notebooks) (10 points).
- TP:
  - introduction to the R language but not to programming !
  - Hands-on
  - Real case studies, introduction to the project
- Free handbook in english (on UV page): use the form for comments/remarks.

Interaction with the professor

- Teams
- UV page: https://uv.ulb.ac.be/course/view.php?id=101031
- email
- Student representative

ML in the AI scenario

Systems that think like humans	Systems that think rationally
"The exciting new effort to make computers think . . . *machines with minds*, in the full and literal sense." (Haugeland, 1985) "[The automation of] activities that we associate with human thinking, activities such as decision-making, problem solving, learning . . ." (Bellman, 1978)	"The study of mental faculties through the use of computational models." (Charniak and McDermott, 1985) "The study of the computations that make it possible to perceive, reason, and act." (Winston, 1992)
Systems that act like humans	**Systems that act rationally**
"The art of creating machines that perform functions that require intelligence when performed by people." (Kurzweil, 1990) "The study of how to make computers do things at which, at the moment, people are better." (Rich and Knight, 1991)	"Computational Intelligence is the study of the design of intelligent agents." (Poole *et al.*, 1998) "AI . . . is concerned with intelligent behavior in artifacts." (Nilsson, 1998)

Rational agent implementation

Given a set of inputs (perceptions) and outputs (actions), loop over:

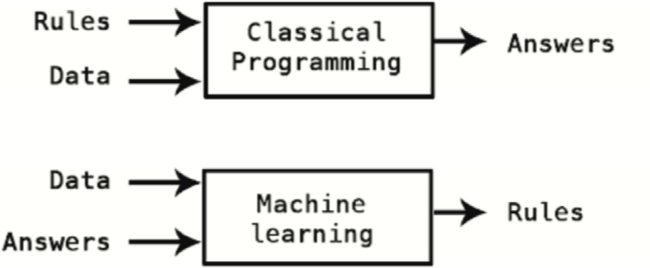
1. Update belief on state based on actions and perceptions
2. Predict outcome possible actions based on belief
3. Decide action with highest expected utility

All this requires knowledge about an **uncertain world** but much of this knowledge cannot be formalized (e.g. detect an object in an image). Attempts to encode human knowledge in a formal knowledge base typically failed.

Machine Learning: a definition

The field of machine learning is concerned with the question of how to construct computer programs that automatically improve with experience.

ML paradigm



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

Machine learning and statistics

- **Reductionist attitude:** *ML is a modern buzzword which equates to statistics plus marketing*
- **Positive attitude:** ML paved the way to the treatment of challenging problems, sometimes overlooked by statisticians (nonlinearity, classification, pattern recognition, missing variables, adaptivity, optimization, massive datasets, data management, causality, representation of knowledge, non stationarity, high dimensionality, parallelisation)
- **Interdisciplinary attitude:** ML *should* have its roots on statistics and complements it by focusing on: algorithmic issues, computational efficiency, data engineering.

Who is a data scientist ?

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

What is statistics?

- Early definition: ".. describes tabulated numerical facts relating to the **state**".
- ".. refers to the methodology for the collection, presentation and **analysis of data**, and for the uses of such data"
- "use of data to make intelligent, rigorous, statements about a **much larger phenomenon** from which the data were selected"
- "...aid the interpretation of data that are subject to appreciable haphazard **variability**"
- "... builds on the use of **probability** to describe variation".
- "...is applied philosophy of science.. ".

What is statistics?

- "The objective of statistics is the understanding of **information** contained in data"
- "...is the study of how degrees of **belief** are altered by data".
- "Statistics concerns the **optimal** methods of treating, analyzing data generated from some chance mechanism".
- "... presents management with quantitative facts which may assist in making **decisions**".
- "...permits the decision maker to evaluate the magnitude of the **risk** in the light of possible gains to be achieved".

An integrated definition

We will adopt the integrated definition proposed by Vic Barnett in the book "Comparative statistical inference" (1999)

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.

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.
- We are interested in knowing what that outcome will be.
- We will have to take a decision anyway.

Why are we interested to uncertainty?

- Uncertainty is pervasive in our world.
- *...in this world there is nothing certain but death and taxes...(Franklin)*
- 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?).

Why probabilistic modeling?

- Any attempt to construct a theory to guide behavior in a situation involving uncertainty must depend on the construction of a formal model of such situations.
- This requires a formal notion of uncertainty.
- In this course we will recur to the formalism of probability in order to represent uncertainty.
- In very general terms, a stochastic model is made of
  1. a set of possible outcomes of a phenomenon and
  2. a a probability distribution, e.g. a mathematical function that gives the probabilities of occurrence 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 course of action relevant to that situation.
- 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

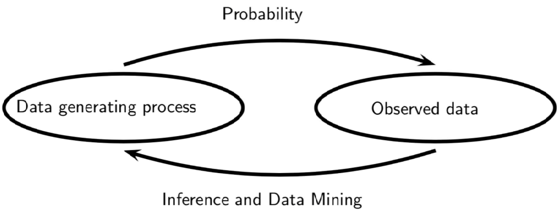
Consider

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 is useful for deriving statements about the behavior of a phenomenon described by a probabilistic model.

Probability and statistics



Excerpt from Wasserman book "All of statistics"

Deduction and induction

- **Under the assumption that a probabilistic model is correct**, logical deduction through the ideas of mathematical probability leads to a description of the properties of data that might arise from the real situation.
- The theory of statistics is designed to **reverse the deductive process**. It takes data that have arisen from a practical situation and uses the data to
  - 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.**

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*

Note that in this case 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

The two main phases of human reasoning are the acquisition of true knowledge and its manipulation in a truth-preserving manner.

- Deduction is *truth-preserving*, i.e. conclusion is a logical consequence of (it follows from) the premises, in other terms conclusion is implicitly contained in the premises
- Induction is *ampliative*, 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 as there is no correspondent to the notion of deductive validity
- Models or universal statements (as well as knowledge) are not logical or **necessary** consequence of data or observational statements (or facts).

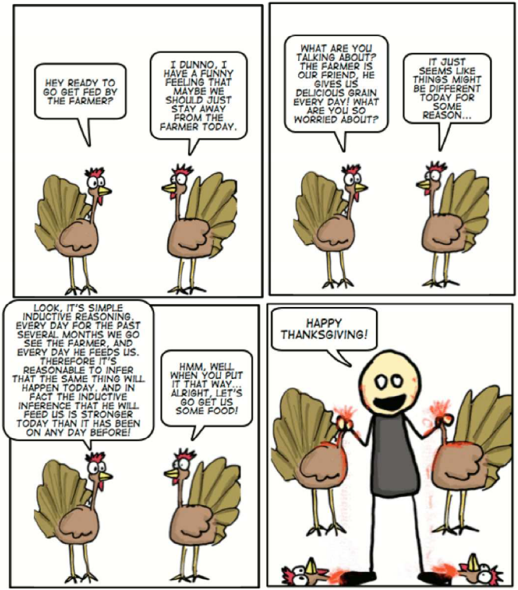
Hume

- David Hume (1711-1776) is a Scottish philosopher who studied the problem of induction from a philosophic perspective.
- In 1739 (he was 28 years old) he published *A treatise of human nature*, one of the most influential books of Western philosophy
- According to Hume *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



Popper

- Karl Popper (1902-1994) is generally regarded as one of the greatest philosophers of science of the 20th century. In 1935 he published *The Logic of Scientific Discovery*.
- He says that Hume is right and there is no justification of inductive inference.
- According to Popper, humans or scientists do not make inductions, they make conjectures (or hypothesis) 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 observational statements
- However the falsity of universal statements ("All swans are white") can be deduced from suitable observations

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.
- 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 and goes further by proposing a set of strategies for automatically generating hypothesis and validating them on the basis of empirical evidence.
- We will show that there is 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*?

Inductive bias

There are plenty of methods to introduce a bias (or an assumption) in the learning process. Some of them are explicit but mostly of them are implicit. Common examples of inductive biases are

- Parametric family.
- Family of algorithms (e.g. based on similar tasks).
- Regularization strategy.
- Prior distributions on parameters or hyperparameters.
- Smoothness (e.g. $f(x) \approx f(x + \epsilon)$).
- Locality (e.g. $f(x_1) \approx f(x_2)$ if x_1 close to x_2).
- Existence of low-dimensional manifolds (data concentrated in low dimensional spaces).
- Sparsity.
- Independence.
- Composition hypothesis (e.g. underlying deep learning)
- Priors (in Bayesian statistics)

Interpretations of probability

Probability is the language of stochastic modeling and statistical machine learning. However, a variety of philosophical interpretations of the probability concept can exist.

- **Frequentist**: statistical analysis must be based on the use of sample data evaluated through a frequency concept of probability. Information comes only from repeated observations.
- **Bayesian**: wider concept of information than that of just sample data. Earlier experience or prior information must be also taken into consideration. This approach addresses the issue of combining prior information with sample data.

In this course, probability has to be considered as a useful theory of uncertainty with its own rules and models.