Statistical foundations of machine learning

INFO-F-422

Gianluca Bontempi

Département d'Informatique

Boulevard de Triomphe - CP 212

http://www.ulb.ac.be/di

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.
- Exam: oral questions and project.
- TP:
  - introduction to the R language.
  - Hands-on
  - Real case studies
- Web page:
  `http://www.ulb.ac.be/di/map/gbonte/InfoF422.html`
- Syllabus in english (on the web page)

Outline of the course

- Foundations of statistical inference.
- Estimation
- Hypothesis testing
- Nonparametric methods
- Statistical machine learning
- Linear models
  - Regression
  - Classification
- Nonlinear models
  - Regression
  - Classification

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.

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 real problems related to data analysis, 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.

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”.
- “ ... treats data that were obtained by some repetitive operation... probability as an idealization of the proportion of times that a certain result will occur in repeated trials of an experiment”

What is statistics?

- “The object of statistics is information. 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”.
- “Statistics is fundamentally concerned with the understanding of structure in data”.
- “... 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”.

A little history

- Statistics, as an organized and formal scientific discipline has a relative short history in comparison with the traditional deterministic sciences of chemistry, physics or mathematics.
- Colbert(1619-1683), Vauban (1633-1707): census in Canada and France.
- Petty (1623-1687): Political Arithmetic school (data modeling, forecasting laws in the economic and demographic context).
- Books of Cardano (1663), Galilei (1656) on games of chance.
- Pascal (1623-1662) and Fermat (1601-1665) work on probability.
- Gauss and Laplace apply the probability models to astronomy.
- Galton introduces the notions of correlation and regression in biometry.

Advances in the XX century

- 1890-1920: mathematical statistics (Pearson, Yule and Gosset (UK); Borel, Fréchet and Poincaré (France); Chuprov and Markov (Russia)).
- 1921-1932: estimation theory (Fisher).
- 1940-1945: hypothesis testing (Neyman and Pearson), sampling theory (Neyman), experimental design (Fisher).
- 1958: control theory, system identification (Kalman).
- 1958-: neural networks (Rosenblatt,Widrow, Hoff).
- See *A History of Mathematical Statistics from 1750 to 1930* by Anders Hald or *Against the Gods: the remarkable history of risk* by P. L. Bernstein.

An integrated definition

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

Definition 1. *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*?
- What is the difference between the *reflection* and the *guidance* function of statistics?

Some examples of uncertain situations

- A student tackling an exam.
- A doctor prescribing a drug.
- An oil company deciding where to drill.
- A football trainer deciding who will shoot the decisive penalty.
- A financial investor in the NASDAQ trade market.
- An employee in front of an 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 could be interested in knowing what that outcome will be.
- We could have to take a decision anyway.

Why are we interested to uncertainty?

- Uncertainty is pervasive in our world.
- 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 stochastic 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 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 statement of the set of possible outcomes of a phenomenon and
  2. a specification of the probabilistic mechanism governing the pattern of outcomes that might arise.
- Notions like independence, randomness, etc., has to be defined for distinguishing and characterizing 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...)

Two configurations involving a model

Consider

1. a real phenomenon P , e.g. the car traffic in a Brussels boulevard leading to Place Montgomery.
2. a probabilistic model M describing the number of cars per hour in a boulevard leading to Place Montgomery.

Deductive or probabilistic configuration. We use the model M to predict the number of cars passing through the square 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.

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 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.
- Note that the inductive process is possible only because the “language” of probability is available to form the deductive link.

Rules of deduction and induction

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.

An induction rule looks like

> *Until now, all university professors we met, were smart. I am listening to an university professor.*
> *So this professor will be smart*

Note that in this case the premise could be true even though the conclusion is false. How to mesure the reliability of this rule?

Note that, *before discovery of Australia, people in the Old World were convinced that all swans were white.*

About inductive reasoning

- Though inductive reasoning is not logically valid, there is no doubt that we continuosly 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.
- 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 logicaly demonstrated but the success of the human beings in using it is an empirical proof of its validity.
- *Could you prove that the snow is cold?*

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 (Fisher, Neyman, Pearson,von Mises, Bartlett).

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 (Ramsey, Lindley).

Decision theory: This approach is designed to provide rules for action in situations of uncertainty. It considers the notion of uncertainty as strictly related to the notion of action. The assessment of consequences of alternative actions and their formal quantification through the concept of utility is central to this approach (Wald, von Neumann).

Considerations

- The three approaches address three general forms of information that may, depending on circumstances, be relevant to a statistical study.
- Behind the frequentist approach there is the intention to produce a theory which should be universal, free of subjective assessments and based on quantifiable elements.
- The three forms of information refer to three different time horizons: the prior information (Bayesian) accumulated from *past* experiences, the sample data (frequentist) arising from the *current* situation, and assessment of consequences referring to potential *future* action (decision theory).
- Specific statistical tools have been developed to model, incorporate, use these types of information.

Inference and decision making

- Any statistical procedure that utilizes information to obtain a description of the practical situation in probabilistic terms is called a *statistical inference procedure*.
- Any statistical procedure with the aim of suggesting action to be taken is a *statistical decision making procedure*.
- Decision-making extends the descriptive aims of inference by incorporating assessments of consequences.

The doctor example

- Consider a doctor prescribing a drug for a patient.
- Let us consider two possible outcomes (drug works, drug does not work) and suppose that uncertainty exists about the effect of this drug.
- How to model the uncertain effect of the drug?
- We can use a simple probabilistic model that assigns a probability p to the success and a probability $1 - p$ to the failure.
- How do we interpret the probability value p?
- In the frequentist interpretation p is the probability of success that the same drug had on “similar” patients in the past. The value p is then related to the proportion of successes, or potential successes, in a large population.

The doctor example (II)

- In the Bayesian interpretation, the probability p is regarded as a measure of the personal doctor's *degree of belief* in the success of the treatment.
- In the decision theory framework, the probability p has to be combined with a certain measure of utility u (or cost) to take the best action.

	p	1−p
	success	failure
drug	0	c_1
no drug	c_2	0