Statistical Learning

Chapter 0. Introduction

Pedro Delicado and Alex Sanchez

Universitat Politécnica de Catalunya and Universitat de Barcelona

Statistics and Machine Learning. The two Cultures

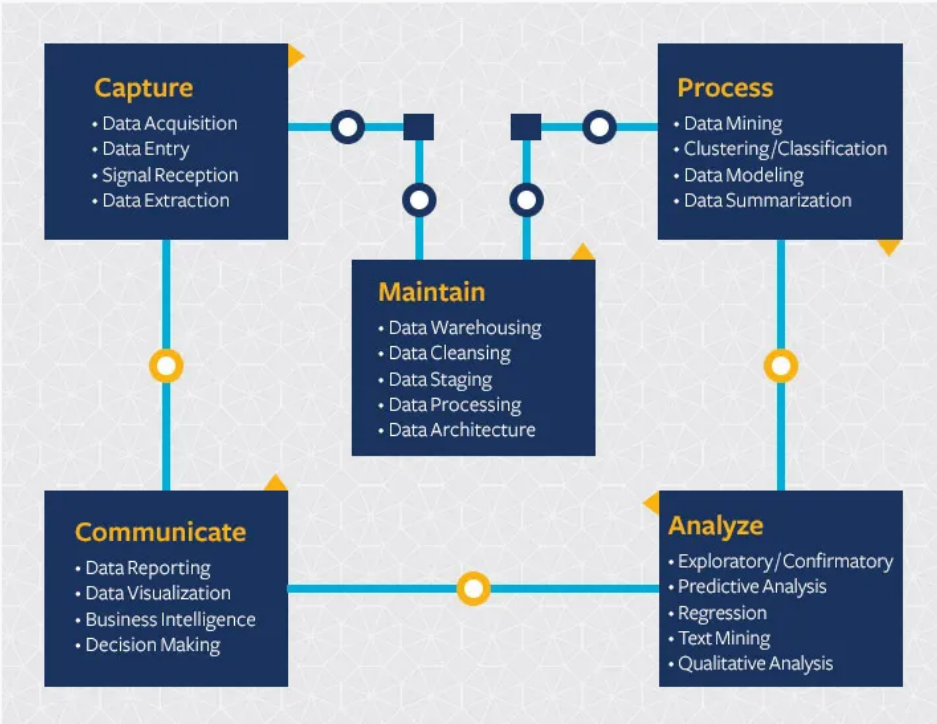
- Statistics is the science of collecting, analyzing, interpreting, and presenting data to extract meaningful insights, quantify uncertainty, and support decision-making under uncertainty (Efron and Hastie 2016).
- We define Machine Learning as a set of methods that can automatically detect patterns in data, and then use the uncovered patterns to predict future data (Murphy 2012).
- While Statistics and Machine Learning have many elements in common they are considered to have appeared (more or less) independently.
- Leo Breiman wrote a famous paper about it Breiman (2001).

Statistical Learning

- Statistical Learning refers to a framework for understanding and modeling complex datasets by using statistical methods to estimate relationships between variables, make predictions, and assess the reliability of conclusions.
- It bridges traditional statistics and machine learning by incorporating model interpretability, regularization techniques, and a focus on uncertainty quantification.
- It became popular after the 1st edition of Trevor Hastie, Tibshirani, and Friedman (2009).

Data Science

- Data Science: an interdisciplinary field combining techniques from statistics, machine learning, computer science, and domain-specific knowledge to extract insights and value from data.



Statistics, ML. Statistical Learning and Data Science

- Statistics provides the theoretical foundation for analyzing data, quantifying uncertainty, and drawing valid inferences. It ensures the rigor and interpretability of results.
- Machine Learning contributes computational and algorithmic tools that enable automated pattern detection and predictive modeling, often with a focus on scalability and performance.
- Statistical Learning serves as the bridge between statistics and machine learning, emphasizing interpretable models, regularization techniques, and uncertainty quantification.
- Data Science integrates all these components while also incorporating data engineering, visualization, and domain expertise to address real-world data-driven problems.

Different focus of Statistics and Machine Learning (1)

- Much of statistical technique was originally developed in an environment where data were scarce and difficult or expensive to collect, so statisticians focused on creating methods that would maximize the strength of inference one is able to make, given the least amount of data Baumer, Kaplan, and Horton (2017).
- Much of the development of statistical theory was to find mathematical approximations for things that we couldn't yet compute Baumer, Kaplan, and Horton (2017).
- Mathematics was the best computer T. Hastie and Efron (2016).

Different focus of Statistics and Machine Learning (2)

- From the 1950s to the present is the "computer age" of Statistics, the time when computation, the traditional bottleneck of statistical applications, became faster and easier by a factor of a million T. Hastie and Efron (2016).
- ML as well as SL and DS found in this increase in computation capabilities the perfect field to develop and strengthen algorithmic approaches to problem solving that were nhard to model or solve analytically.

	Asymptotics, optimality	Interpretability	Accurate prediction	Scalability
Statistics	XXXXX	XXXXX	XX	X
Machine Learning	X	XX	XXXXX	XXXXX

Data Science, learning and prediction

- A particularly energetic brand of the statistical enterprise has flourished in the new century, *data science*, emphasizing algorithmic thinking rather than its inferential justification.
- Similarly to ML, Data Science seems to represent a statistics discipline without parametric probability models or formal inference.
- Why have they taken center stage?
  - Prediction is commercially valuable.
  - Prediction is the simplest use of regression theory.
  - It can be carried out successfully without probability models, perhaps with the assistance of cross-validation, permutations, bootstrap.

Statistical Learning is (also) about *learning from data*

- In statistical learning...
  - We study models and tools originally developed by statisticians (sparse estimation or nonparametric versions of linear and generalized linear models, classification and regression trees, ...)
  - Many of these methods are now part of the toolkit of Machine Learning practitioners, usually coming from Computer Science, and Data Scientists in general.
  - We also learn algorithms and procedures proposed by researchers in Machine Learning (neural networks, support vector machines, boosting, random forests, ...) now part of the statisticians' prediction toolkit.
- So we can also don't care much about the labels and agree that the course is about *learning from data*, and specifically focused on the prediction problem.

Examples of Statistical Learning Problems

- Identify the risk factors for prostate cancer.
- Predict whether someone will have a heart attack on the basis of demographic, diet and clinical measurements.
- Classify a recorded phoneme based on a log-periodogram.
- Customize an email spam detection system.
- Identify the numbers in a handwritten zip code.
- Classify a tissue sample into one of several cancer classes, based on a gene expression profle.
- Establish the relationship between salary and demographic variables in population survey data.

Source *James et al. (2023)*

References

Baumer, B. S., D. T. Kaplan, and N. J. Horton. 2017. *Modern Data Science with r*. CRC Press.

Breiman, Leo. 2001. "Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author)." *Statistical Science* 16 (3): 199–231. https://doi.org/10.1214/ss/1009213726.

Hastie, T., and B. Efron. 2016. *Computer Age Statistical Inference: Algorithms, Evidence, and Data Science*. Cambridge University Press.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. 2009. "The Elements of Statistical Learning." https://doi.org/10.1007/978-0-387-84858-7.

James, Gareth, Daniela Witten, T. Hastie, R. Tibshirani, and Jonathan Taylor. 2023. *An Introduction to Statistical Learning with Applications in Python*. Springer.