T3A: Machine Learning Algorithms

Master of Science in AI and Master of Science in Data Science
@ UPSaclay
2024/2025.

Faïcel Chamroukhi

chamroukhi.com

T3A: Machine learning algorithms

Objective: master the core concept of algorithmic design in ML, from an optimization or a probabilistic point-of-view, using supervised and unsupervised algorithms.

Programme

- Week1: Regression/classification seen in optimization and probabilistic frameworks, implication on batch and stochastic gradient descent
- Week2: Learning theory and Vapnick-Charvonenkis dimension
- Week3: Evaluating performances of ML algorithms in different contexts (imbalanced, small-sized, etc)
- Week4: Probabilistic framework for machine learning: Discriminative vs Generative learning, Empirical Risk Minimization, Risk Decomposition, Bias-Variance Tradeoff; Maximum Likelihood Estimation (MLE), MLE and OLS in regression, MLE and IRLS in softmax classification
- Week5: Unsupervised Learning and Clustering: K-means, Mixture Models, EM algorithms,..)
- Week6: Unsupervised Learning and Dimensionality reduction: PCA, Probabilistic PCA & EM, ICA,…
- Week7: Final exam. An intermediate exam (contrôle continu) will be scheduled during the course

Schedule

- **When & where:** Thursdays, 1:30-5:00 pm, from January 09 till February 27, in Building PUIO
- All sessions should be attended in person unless you're informed otherwise about an exceptional online session
- Little coding sessions might be organized during the course for illustrations

Contact : Inquires about the course to be sent to Sylvain Chevallier and Faicel Chamroukhi <firstame.lastname@universite-paris-saclay.fr>

Teaching Schedule :

- Sylvain Chevallier is teaching in weeks 1, 2, and 3.
- Faïcel Chamroukhi is teaching in weeks 4, 5, and 6.

Outline

Introduction

université PARIS-SACLAY

- What does Data Science mean ?
- What about Statistics in the Data Science "area" ?

CONTRIBUTED ARTICLES

Data Science and Prediction

By Vasant Dhar
Communications of the ACM, Vol. 56 No. 12, Pages 64-73
10.1145/2500499
Comments (2)

VIEW AS: SHARE:

Use of the term "data science" is increasingly common, as is "big data." But what does it mean? Is there something unique about it? What skills do "data scientists" need to be productive in a world deluged by data? What are the implications for scientific inquiry? Here, I address these questions from the perspective of predictive modeling.

Back to Top

Key Insights

- Data science is the study of the generalizable extraction of knowledge from data.
- A common epistemic requirement in assessing whether new knowledge is actionable for decision making is its predictive power, not just its ability to explain the past.
- A data scientist requires an integrated skill set spanning mathematics, machine learning, artificial intelligence, statistics, databases, and optimization, along with a deep understanding of the craft of problem formulation to engineer effective solutions.

ASA | Amstat News | ASA Community | The World of Stati

AMSTATNEWS
The Membership Magazine of the American Statistical Association

HOME | ABOUT | EDITORIAL CALENDAR | PDF ARCHIVES | ADVERTISE | STATISTICIANS IN HISTORY

Home » Featured

ASA Statement on the Role of Statistics in Data Science

1 OCTOBER 2015 6,956 VIEWS 13 COMMENTS

Statement Contributors
David van Dyk, *Imperial College* (chair)
Montse Fuentes, *NCSU*
Michael I. Jordan, *UC Berkeley*
Michael Newton, *University of Wisconsin*
Bonnie K. Ray, *Pegged Software*
Duncan Temple Lang, *UC Davis*
Hadley Wickham, *RStudio*

The rise of data science, including Big Data and data analytics, has recently attracted enormous attention in the popular press for its spectacular contributions in a wide range of scholarly disciplines and commercial endeavors. These successes are largely the fruit of the innovative and entrepreneurial spirit that characterize this burgeoning field. Nonetheless, its interdisciplinary nature means that a substantial collaborative effort is needed for it to realize its full potential for productivity and innovation. While there is not yet a consensus on what precisely constitutes data science, three professional communities, all within computer science and/or statistics, are emerging as foundational to data science: (i) Database Management enables transformation, conglomeration, and organization of data resources, (ii) Statistics and Machine Learning convert data into knowledge, and (iii) Distributed and Parallel Systems provide the computational infrastructure to carry out data analysis.

- For a review, see the report of D. Donoho (2015) : "50 years of Data Science"
- There is not yet a consensus on what precisely constitutes Data Science, but

- Data Science can be seen (defined ?) as [a] :

↪ the study of the generalizable extraction of knowledge from data.

↪ requires an integrated skill set spanning maths/statistics, machine learning, optimization, databases..

a. Vasant Dhar (2013) : Communications of the ACM, Vol. 56 No. 12 : 64-73

- Foundations : Databases, statistics and machine learning, and distributed systems [1]

(i) Databases : organization of data resources,

(ii) **Statistics** and **Machine Learning** : convert data into knowledge,

(iii) Distributed and Parallel Systems : computational infrastructure

Statistics play a central role in data science

- Allow to quantify the randomness component in the data
- A well-established background to deal with uncertainty (probabilistic framework) and to establish generalizable methods for estimation and prediction
- allow soft decision : e.g. confidence intervals (error bars)

1. ASA Statement on the Role of Statistics in Data Science, oct. 2015

Introduction

- We assume that we have a set of data collected in some way (e.g., independent or not (i.e sequential), complete or not, etc.),

- to analyze, in some sense (e.g., for prediction, exploration, selection, visualisation, etc.), some scenario or system, in a broad sense.

  prediction, clustering, dimensionality reduction, visualisation, etc

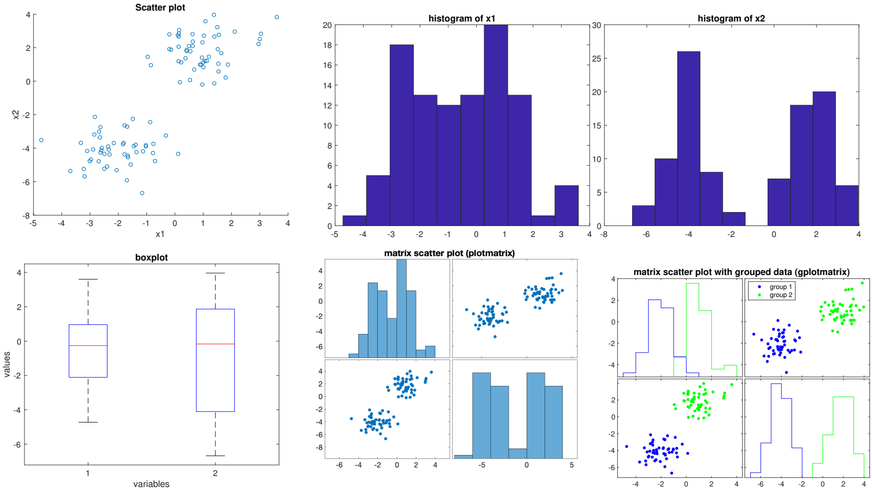
- We assume that the data are represented by random variables $\hookrightarrow$ statistical learning framework

Introduction

universite PARIS-SACLAY

- We assume that we have a set of data collected in some way (e.g., independent or sequential, complete or incomplete).
- We analyze this data for various purposes, such as :
  - Prediction
  - Exploration
  - Selection
  - Visualization
  - Clustering
  - Dimensionality Reduction
- We assume that the data are represented by random variables, leading to the **statistical learning framework**.

Descriptive Analysis



Predictive Analysis

Regression

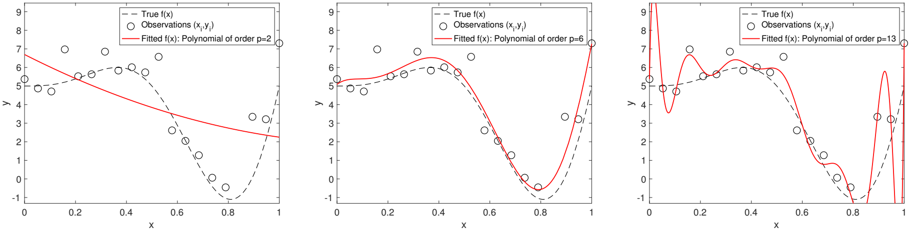


FIGURE – Scatter plot (∘), Target function (--), fitted function (—)

Regression

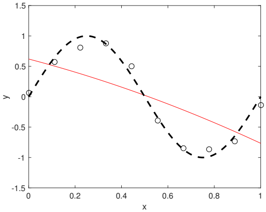


FIGURE – Scatter plot (∘), Target function (--), fitted function (—)

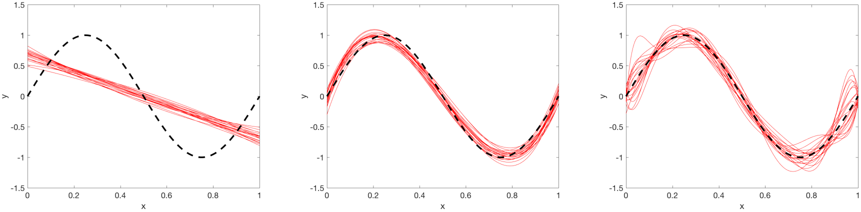
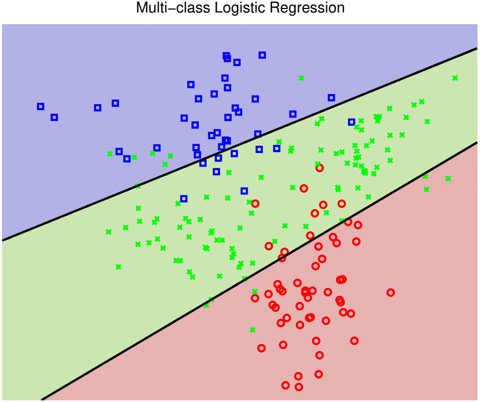


FIGURE – True model (--), realizations from the fitted model (—)

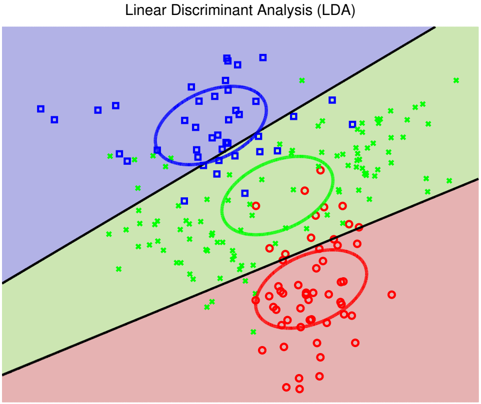
Predictive Analysis

Classification



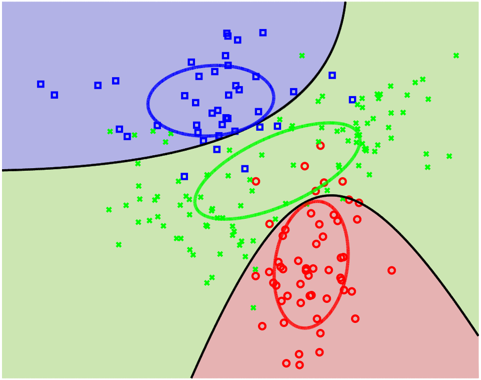
Predictive Analysis

Classification



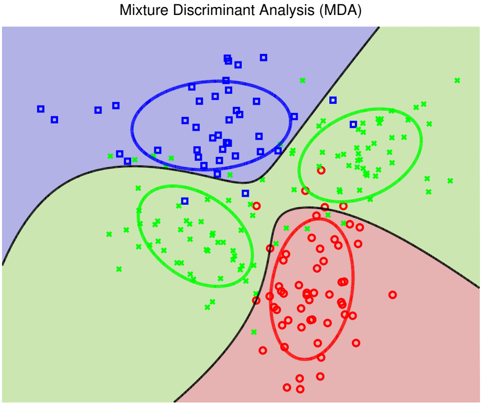
Classification

Quadratic Discriminant Analysis (QDA)



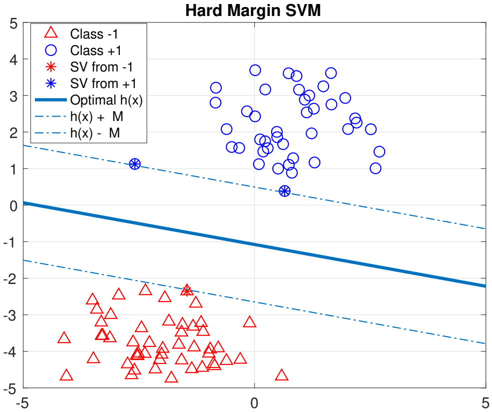
Predictive Analysis

Classification

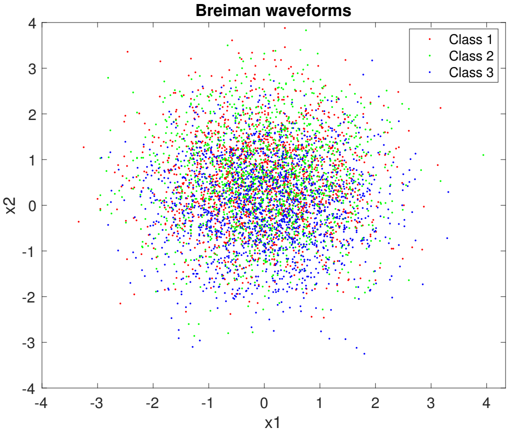


Predictive Analysis

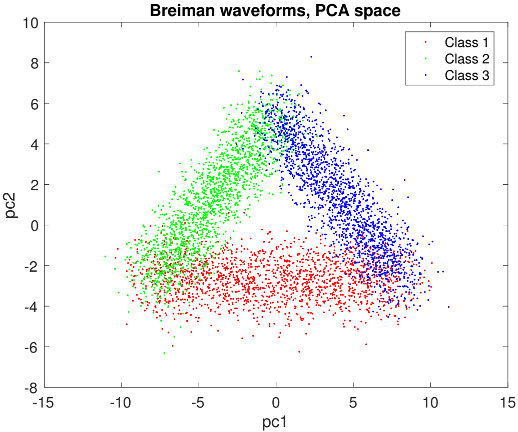
Classification



Representation



Representation

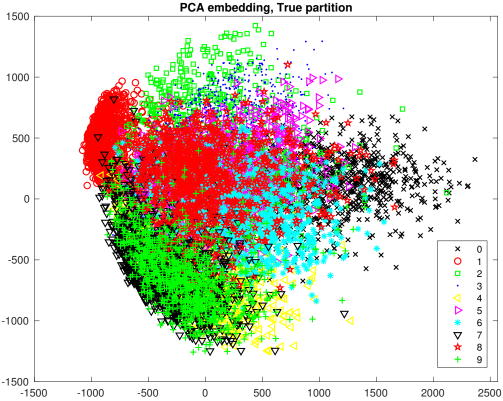


Unsupervised Learning

Clustering / Representation / Data viz / Dimensionality reduction

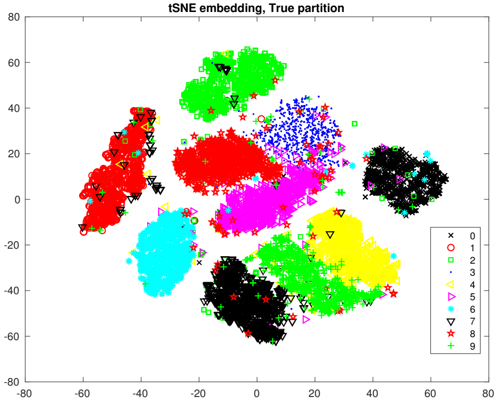
Unsupervised Analysis

Representation / Data viz / Dimensionality reduction

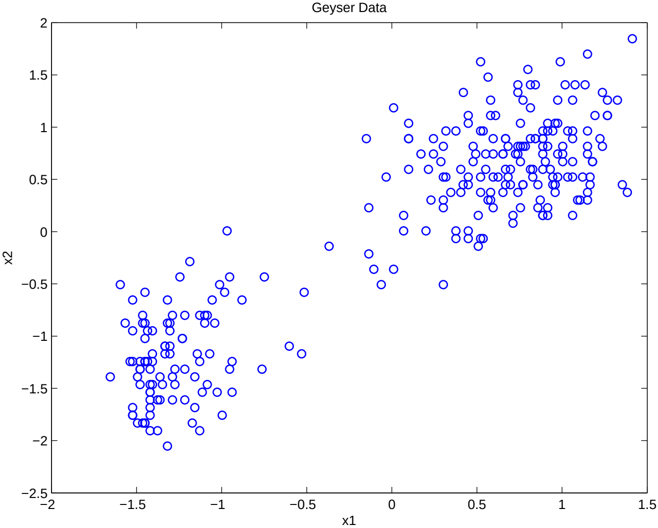


Unsupervised Analysis

Representation / Data viz / Dimensionality reduction

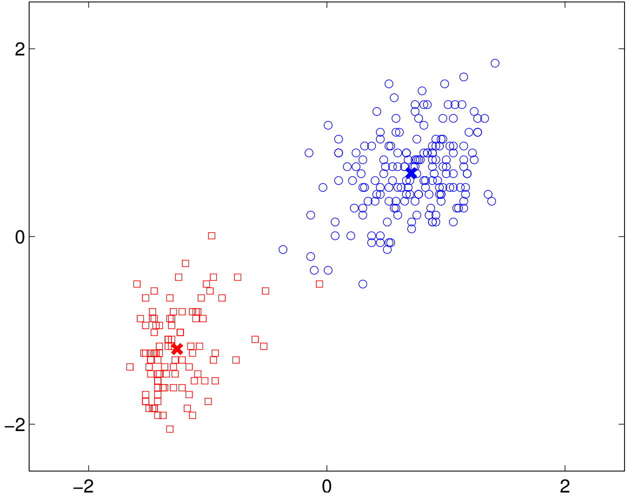


Unsupervised analysis



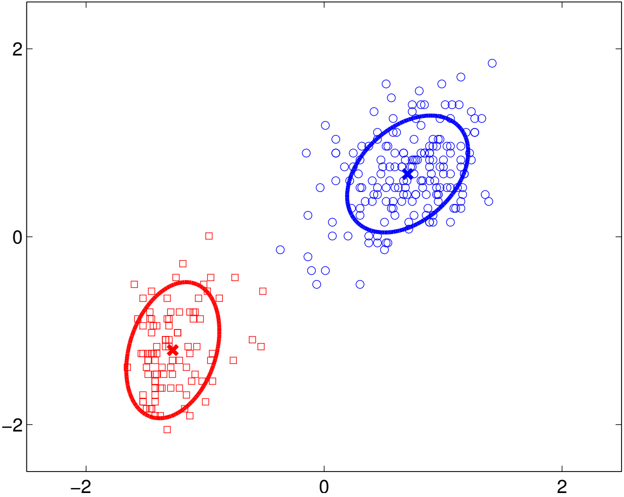
Unsupervised analysis

Clustering



Unsupervised analysis

Clustering



Unsupervised analysis

Clustering

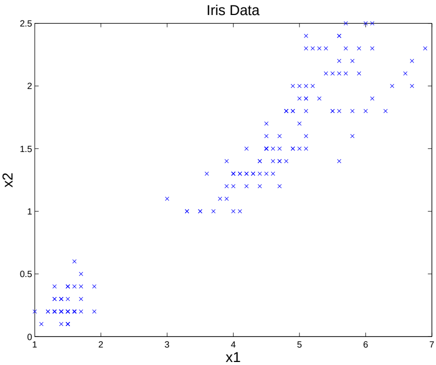


FIGURE – A three-class example of a real data set : Iris data of Fisher.

Unsupervised analysis

Clustering

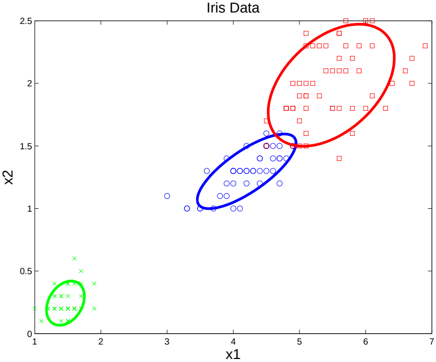


FIGURE – Iris data : Clustering results with EM for a GMM and AIC.

Machine Learning

- Field of study that gives computers the ability to learn without being explicitly programmed ; "Programming computers to learn from experience should eventually eliminate the need for much of this detailed programming erfort." (A. Samuel, 1959).
- "Machine learning is concerned with the question of how to construct computer programs that automatically improve with experience". Tom M. Mitchell
- **Definition.** A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E. Tom M. Mitchell
- Example : A handwriting recognition learning problem :
  - Task T : recognizing and classifying handwritten words within images
  - Performance measure P : percent of words correctly classified
  - Training experience E : a database of handwritten words with given classifications

Notation

General used notation (deviation from this notation will be mentioned prior to use) :

- $x, y, z, t, ...$ small letter for scalars
- $\alpha, \beta, \gamma, \theta, \ldots$ Greek letters for scalar parameters
- $\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}, ...$ boldface letters and $\mathbf{x}, \mathbf{y}, \mathbf{z}, \ldots$ upright bold for vectors
- $\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\gamma}, ...$ boldface Greek letters for vector parameters
- $X, Y, Z, ...,$ Capitalized for random variables
- $\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z}, \boldsymbol{T}$ Capitalized boldface letters for random vectors
- $\boldsymbol{A}, \mathbf{B}, \mathbf{X}, \mathbf{Y}, \ldots$ Capitalized upright bold for Matrices
- $\Gamma, \Sigma, \Lambda, \Upsilon$ Capital Greek letters for matrix parameters
- $\mathcal{X}, \mathcal{Y}, \mathcal{Z}, \ldots$ Calligraphic capital letters for sets (except for standard sets $\mathbb{N}$, $\mathbb{R}$,..)
- $\Theta, \Omega, \boldsymbol{\Theta}, \boldsymbol{\Omega}, \ldots$ (boldface) capital Greek for sets of (vector) parameters
- $\mathbb{P}$ probability, $\mathbb{E}$ expectation, $\mathbb{V}$ or Var variance, Cov co-variance
- $\mathbf{A}^\top$, $\mathbf{A}^{-1}$, trace$(\mathbf{A})$, $|\mathbf{A}|$, diag$(\mathbf{A})$: transpose, inverse, trace, determinant, and diagonal of $\mathbf{A}$
- All vectors are assumed to be column vectors

Some analysis notions for optimization

Argmax/Argmin, Max/Min, and Supremum/Infimum of a function f defined on a set D_f

$$\begin{aligned}
\textbf{arg max :} \quad \arg\max_{x \in D_f} f(x) &= \{x : f(x) \geq f(y), \forall y \in D_f\} \\
\textbf{max :} \quad \max_{x \in D_f} f(x) &= f(x^*), \text{ for any } x^* \in \arg\max_{x \in D_f} f(x) \\
\textbf{sup :} \quad \sup f(x) &= \min_{y : y \geq f(x), \forall x \in D_f} y.
\end{aligned}$$

If g is strictly monotonic, meaning that $\alpha > \beta$ implies $g(\alpha) > g(\beta)$, then

$$\arg\max g(f(x)) = \arg\max f(x) \quad \text{and} \quad \max g(f(x)) = g(\max f(x)).$$

$$\begin{aligned}
\textbf{arg min :} \quad \arg\min_{x \in D_f} f(x) &= \{x : f(x) \leq f(y), \forall y \in D_f\} \\
\textbf{min :} \quad \min_{x \in D_f} f(x) &= f(x^*), \text{ for any } x^* \in \arg\min_{x \in D_f} f(x) \\
\textbf{inf :} \quad \inf f(x) &= \max_{y : y \leq f(x), \forall x \in D_f} y.
\end{aligned}$$

$$\begin{aligned}
\arg\min_{x \in D_f} f(x) &= \arg\max_{x \in D_f} -f(x) \\
\min_{x \in D_f} f(x) &= -\max_{x \in D_f} -f(x) \\
\inf_{x \in D_f} f(x) &= -\sup_{x \in D_f} -f(x).
\end{aligned}$$

Some analysis notions for optimization

Argmax/Argmin, Max/Min, and Supremum/Infimum of a function f defined on a set D_f

$$\begin{aligned} \textbf{arg max :} \quad \arg\max_{x \in D_f} f(x) &= \{x : f(x) \geq f(y), \forall y \in D_f\} \\ \textbf{max :} \quad \max_{x \in D_f} f(x) &= f(x^*), \text{ for any } x^* \in \arg\max_{x \in D_f} f(x) \\ \textbf{sup :} \quad \sup f(x) &= \min_{y : y \geq f(x), \forall x \in D_f} y. \end{aligned}$$

If g is strictly monotonic, meaning that $\alpha > \beta$ implies $g(\alpha) > g(\beta)$, then

$$\arg\max g(f(x)) = \arg\max f(x) \quad \text{and} \quad \max g(f(x)) = g(\max f(x)).$$

$$\begin{aligned} \textbf{arg min :} \quad \arg\min_{x \in D_f} f(x) &= \{x : f(x) \leq f(y), \forall y \in D_f\} \\ \textbf{min :} \quad \min_{x \in D_f} f(x) &= f(x^*), \text{ for any } x^* \in \arg\min_{x \in D_f} f(x) \\ \textbf{inf :} \quad \inf f(x) &= \max_{y : y \leq f(x), \forall x \in D_f} y. \end{aligned}$$

$$\begin{aligned} \arg\min_{x \in D_f} f(x) &= \arg\max_{x \in D_f} -f(x) \\ \min_{x \in D_f} f(x) &= -\max_{x \in D_f} -f(x) \\ \inf_{x \in D_f} f(x) &= -\sup_{x \in D_f} -f(x). \end{aligned}$$

Some analysis notions for optimization

Argmax/Argmin, Max/Min, and Supremum/Infimum of a function f defined on a set D_f

$$
\begin{aligned}
\textbf{arg max :} \quad \arg\max_{x \in D_f} f(x) &= \{x : f(x) \geq f(y), \forall y \in D_f\} \\
\textbf{max :} \quad \max_{x \in D_f} f(x) &= f(x^*), \text{ for any } x^* \in \arg\max_{x \in D_f} f(x) \\
\textbf{sup :} \quad \sup f(x) &= \min_{y : y \geq f(x), \forall x \in D_f} y.
\end{aligned}
$$

If g is strictly monotonic, meaning that $\alpha > \beta$ implies $g(\alpha) > g(\beta)$, then

$$
\arg\max g(f(x)) = \arg\max f(x) \quad \text{and} \quad \max g(f(x)) = g(\max f(x)).
$$

$$
\begin{aligned}
\textbf{arg min :} \quad \arg\min_{x \in D_f} f(x) &= \{x : f(x) \leq f(y), \forall y \in D_f\} \\
\textbf{min :} \quad \min_{x \in D_f} f(x) &= f(x^*), \text{ for any } x^* \in \arg\min_{x \in D_f} f(x) \\
\textbf{inf :} \quad \inf f(x) &= \max_{y : y \leq f(x), \forall x \in D_f} y.
\end{aligned}
$$

$$
\begin{aligned}
\arg\min_{x \in D_f} f(x) &= \arg\max_{x \in D_f} -f(x) \\
\min_{x \in D_f} f(x) &= -\max_{x \in D_f} -f(x) \\
\inf_{x \in D_f} f(x) &= -\sup_{x \in D_f} -f(x).
\end{aligned}
$$

Some analysis notions for optimization

Argmax/Argmin, Max/Min, and Supremum/Infimum of a function f defined on a set D_f

$$\begin{aligned}
\textbf{arg max :} \quad \arg\max_{x \in D_f} f(x) &= \{x : f(x) \geq f(y), \forall y \in D_f\} \\
\textbf{max :} \quad \max_{x \in D_f} f(x) &= f(x^*), \text{ for any } x^* \in \arg\max_{x \in D_f} f(x) \\
\textbf{sup :} \quad \sup f(x) &= \min_{y : y \geq f(x), \forall x \in D_f} y.
\end{aligned}$$

If g is strictly monotonic, meaning that $\alpha > \beta$ implies $g(\alpha) > g(\beta)$, then

$$\arg\max g(f(x)) = \arg\max f(x) \quad \text{and} \quad \max g(f(x)) = g(\max f(x)).$$

$$\begin{aligned}
\textbf{arg min :} \quad \arg\min_{x \in D_f} f(x) &= \{x : f(x) \leq f(y), \forall y \in D_f\} \\
\textbf{min :} \quad \min_{x \in D_f} f(x) &= f(x^*), \text{ for any } x^* \in \arg\min_{x \in D_f} f(x) \\
\textbf{inf :} \quad \inf f(x) &= \max_{y : y \leq f(x), \forall x \in D_f} y.
\end{aligned}$$

$$\begin{aligned}
\arg\min_{x \in D_f} f(x) &= \arg\max_{x \in D_f} -f(x) \\
\min_{x \in D_f} f(x) &= -\max_{x \in D_f} -f(x) \\
\inf_{x \in D_f} f(x) &= -\sup_{x \in D_f} -f(x).
\end{aligned}$$

Statistical learning framework

- **Statistical learning** "=" Machine Learning '+' Statistics : the data are assumed to be realizations of random variables $\Rightarrow$ infer probabilistic models from the data
- Let (X, Y) be a pair of random variables distributed on a sample space $\mathcal{X} \times \mathcal{Y}$
- A joint probability distribution on $\mathcal{X} \times \mathcal{Y}$ is denoted as $P_{X,Y}$
- Let (X, Y) be a pair of random variables distributed according to $P_{X,Y}$. We use
- P_X (resp. P_Y) to denote the marginal distribution of X (resp. Y).
- $P_{Y|X}$ (resp. $P_{X|Y}$) to denote the conditional distribution of Y given X (resp. X given Y).

- **Supervised learning** : We are given a set of observed pairs (input, output), and the objective is the **prediction** of the outputs of new inputs. Classification, Regression
- **Unsupervised learning** : The objective is to explore a set of inputs to restore or reveal hidden information. Clustering, Dimensionality reduction (Representation)

Statistical learning framework

université PARIS-SACLAY

- **Statistical learning** "=" Machine Learning '+' Statistics : the data are assumed to be realizations of random variables $\Rightarrow$ infer probabilistic models from the data
- Let (X, Y) be a pair of random variables distributed on a sample space $\mathcal{X} \times \mathcal{Y}$
- A joint probability distribution on $\mathcal{X} \times \mathcal{Y}$ is denoted as $P_{X,Y}$
- Let (X, Y) be a pair of random variables distributed according to $P_{X,Y}$. We use
- P_X (resp. P_Y) to denote the marginal distribution of X (resp. Y).
- $P_{Y|X}$ (resp. $P_{X|Y}$) to denote the conditional distribution of Y given X (resp. X given Y).

- **Supervised learning** : We are given a set of observed pairs (input, output), and the objective is the **prediction** of the outputs of new inputs. Classification, Regression
- **Unsupervised learning** : The objective is to explore a set of inputs to restore or reveal hidden information. Clustering, Dimensionality reduction (Representation)

Statistical learning framework

- **Statistical learning** "=" Machine Learning '+' Statistics : the data are assumed to be realizations of random variables $\Rightarrow$ infer probabilistic models from the data
- Let (X, Y) be a pair of random variables distributed on a sample space $\mathcal{X} \times \mathcal{Y}$
- A joint probability distribution on $\mathcal{X} \times \mathcal{Y}$ is denoted as $P_{X,Y}$
- Let (X, Y) be a pair of random variables distributed according to $P_{X,Y}$. We use
- P_X (resp. P_Y) to denote the marginal distribution of X (resp. Y).
- $P_{Y|X}$ (resp. $P_{X|Y}$) to denote the conditional distribution of Y given X (resp. X given Y).

- **Supervised learning** : We are given a set of observed pairs (input, output), and the objective is the **prediction** of the outputs of new inputs. Classification, Regression
- **Unsupervised learning** : The objective is to explore a set of inputs to restore or reveal hidden information. Clustering, Dimensionality reduction (Representation)

Statistical learning framework

université PARIS-SACLAY

- **Statistical learning** "=" Machine Learning '+' Statistics : the data are assumed to be realizations of random variables $\Rightarrow$ infer probabilistic models from the data
- Let (X, Y) be a pair of random variables distributed on a sample space $\mathcal{X} \times \mathcal{Y}$
- A joint probability distribution on $\mathcal{X} \times \mathcal{Y}$ is denoted as $P_{X,Y}$
- Let (X, Y) be a pair of random variables distributed according to $P_{X,Y}$. We use
- P_X (resp. P_Y) to denote the marginal distribution of X (resp. Y).
- $P_{Y|X}$ (resp. $P_{X|Y}$) to denote the conditional distribution of Y given X (resp. X given Y).

- **Supervised learning** : We are given a set of observed pairs (input, output), and the objective is the **prediction** of the outputs of new inputs. Classification, Regression
- **Unsupervised learning** : The objective is to explore a set of inputs to restore or reveal hidden information. Clustering, Dimensionality reduction (Representation)

Generative vs Discriminative learning

universite PARIS-SACLAY

- **Bayes decision rule** : From the conditional distribution $p(y|x)$, we can make predictions of y for any new value of x by maximizing the conditional distribution given the learnt model :

$$\widehat{y} = \arg\max_{y \in \mathcal{Y}} p(y|x).$$

- **Discriminative** approaches directly learn a model of the conditional distribution

$$p(y|x)$$

  or learn a direct map from the input x to the output y.
  (especially used in supervised learning (classification, regression))

- **Generative** approaches learn a model of the joint distribution

$$p(x, y)$$

  They model the conditional distribution $p(x|y)$ together with the prior distribution $p(y)$. The required posterior distribution is then obtained using Bayes' theorem

$$p(y|x) \propto p(y)p(x|y)$$

Generative vs Discriminative learning

universite PARIS-SACLAY

- **Bayes decision rule** : From the conditional distribution $p(y|x)$, we can make predictions of y for any new value of x by maximizing the conditional distribution given the learnt model :

$$\widehat{y} = \arg\max_{y \in \mathcal{Y}} p(y|x).$$

- **Discriminative** approaches directly learn a model of the conditional distribution

$$p(y|x)$$

  or learn a direct map from the input x to the output y.
  (especially used in supervised learning (classification, regression))

- **Generative** approaches learn a model of the joint distribution

$$p(x, y)$$

  They model the conditional distribution $p(x|y)$ together with the prior distribution $p(y)$. The required posterior distribution is then obtained using Bayes' theorem

$$p(y|x) \propto p(y)p(x|y)$$

Generative vs Discriminative learning

universite PARIS-SACLAY

- **Bayes decision rule** : From the conditional distribution $p(y|x)$, we can make predictions of y for any new value of x by maximizing the conditional distribution given the learnt model :

$$\widehat{y} = \arg\max_{y \in \mathcal{Y}} p(y|x).$$

- Discriminative approaches directly learn a model of the conditional distribution

$$p(y|x)$$

or learn a direct map from the input x to the output y.
(especially used in supervised learning (classification, regression))

- **Generative** approaches learn a model of the joint distribution

$$p(x, y)$$

They model the conditional distribution $p(x|y)$ together with the prior distribution $p(y)$. The required posterior distribution is then obtained using Bayes' theorem

$$p(y|x) \propto p(y)p(x|y)$$