Machine learning lecture I: Machine learning frameworks

B. Błaszczyszyn[(1)], L. Darlavoix[(2)], M.K. Karray[(2)]

(1) INRIA; (2) Orange Labs

September 25, 2023

You may download the PDF of the lecture from the website https://mohamedkadhem.com/data-science/.

Acknowledgement: We are particularly grateful to the authors [4], [5], and [3], who are our first source of inspiration for the present work.

Outline

Introduction

Introduction

- Data-science (machine and deep learning) is becoming a widespread tool to solve problems in many domains including telecommunications.
- Data-science has solid mathematical foundations inherited from statistics.
- Mathematics and softwares are oftenly imbricated in data-science writings.

  We believe that it is much more clear to present mathematics and softwares separately.
- We aim here to present the essential mathematical concepts and results of machine and deep learning.

  This lecture relies on [2].
  For useful softwares to write codes for machine and deep learning the reader may refer e.g. to [1].

Basic learning framework

Definition (Basic learning framework)

A *basic learning framework* is composed of the following elements:

- Two measurable spaces $(\mathcal{X}, \mathcal{F}_{\mathcal{X}})$ and $(\mathcal{Y}, \mathcal{F}_{\mathcal{Y}})$ called *input space* and *output space* respectively.
- A *learner* (or *learning algorithm*) is a measurable mapping

$$\mathbb{A} : \bigcup_{n \in \mathbb{N}^*} (\mathcal{X} \times \mathcal{Y})^n \to \mathcal{M},$$

where $\mathcal{M}$ is the set of measurable functions $\mathcal{X} \to \mathcal{Y}$; that is the mapping $\mathbb{A}$

  - takes as *input* a finite sequence $s^{(n)} = ((x_1, y_1), \ldots, (x_n, y_n)) \in (\mathcal{X} \times \mathcal{Y})^n$ called *training sequence*;
  - and returns as *output* a measurable function $h : \mathcal{X} \to \mathcal{Y}$ called *hypothesis*.

Definition (Basic learning framework; cont'd)

- *The learner's objective* is to minimize the loss of the hypothesis $h = \mathbb{A}\left(s^{(n)}\right)$ defined as follows. We assume given the following two elements which are not known to the learner:
  - A probability measure q over $\mathcal{X}$ representing the *input's distribution*.
  - A measurable function $f : \mathcal{X} \to \mathcal{Y}$, called *target function*, representing the correct function that the learner is trying to figure out. In particular, we assume for the training sequence that $y_i = f(x_i)$ for all $i \in \{1, \ldots, n\}$.

  The *true loss* of any $h \in \mathcal{M}$ is defined by

$$\mathcal{L}_{q,f}(h) := \mathbf{P}\left(h(X) \neq f(X)\right), \tag{1}$$

  where $X \overset{\text{dist.}}{\sim} q$.

Definition (Basic learning framework; cont'd)

- The *empirical loss* of $h \in \mathcal{M}$ with respect to a training sequence $s^{(n)} = ((x_1, y_1), \ldots, (x_n, y_n)) \in (\mathcal{X} \times \mathcal{Y})^n$ is defined by

$$L_{s^{(n)}}(h) := \frac{\sum_{i=1}^{n} \mathbf{1}\{h(x_i) \neq y_i\}}{n}. \tag{2}$$

- The training sequence $S^{(n)} = ((X_1, Y_1), \ldots, (X_n, Y_n))$ is sometimes assumed random i.i.d; i.e., $X_1, \ldots, X_n$ are i.i.d random variables with common distribution q and $Y_i = f(X_i)$ for all $i \in \{1, \ldots, n\}$. This is denoted $S^{(n)} \overset{\text{dist.}}{\sim} (q, f)^n$. Moreover, we denote

$$\mathcal{L}_{q,f}\left(\mathbb{A}\left(S^{(n)}\right)\right) := \mathbf{P}\left(\left(\mathbb{A}\left(S^{(n)}\right)\right)(X) \neq f(X) \middle| S^{(n)}\right), \tag{3}$$

  where $X \overset{\text{dist.}}{\sim} q$ is independent of $S^{(n)}$.

- We assume given some *hypothesis class* $\mathcal{H} \subset \mathcal{M}$ which is known to the learner.

Basic learning framework

- $\mathcal{H}$ represents some *prior knowledge*. Here are some examples:
  - We know that the target function f is q-almost everywhere equal to a function in $\mathcal{H}$ (or, more simply, $f \in \mathcal{H}$).
  - The class $\mathcal{H}$ is a class of benchmark hypotheses the learner's is competing with (in a sense that we will explicit later).

  If there is no prior knowledge; consider $\mathcal{H} = \mathcal{M}$.

Lemma (Empirical versus true loss)

If $S^{(n)} \overset{\text{dist.}}{\sim} (q, f)^n$, *then for any* $h \in \mathcal{M}$,

- $L_{S^{(n)}}(h) \to \mathcal{L}_{q,f}(h)$ *as* $n \to \infty$, $\mathbf{P}$*-a.s.*
- $\mathcal{L}_{q,f}(h) = 0$ $(\Leftrightarrow h = f$, q*-a.e.)* *implies that* $L_{S^{(n)}}(h) = 0$, $\mathbf{P}$*-a.s.*

Definition (Realizable hypothesis class)

Hypothesis class $\mathcal{H} \subset \mathcal{M}$ is *realizable* with respect to (q, f) if there exists $h \in \mathcal{H}$ such that $h = f$, q-a.e.

Empirical loss minimization (ELM)

- It is generally difficult to minimize the true loss (1) since the learner does not know q and f.

Definition (Empirical loss minimization (ELM))

The *empirical loss minimization* (ELM) *with respect to* $\mathcal{H}$ is a learner, denoted $\mathrm{ELM}_{\mathcal{H}}$, such that

$$\mathrm{ELM}_{\mathcal{H}}\left(s^{(n)}\right) \in \arg\min_{h \in \mathcal{H}} L_{s^{(n)}}(h), \quad s^{(n)} \in (\mathcal{X} \times \mathcal{Y})^n,$$

provided the above $\arg\min$ exists. If the mapping $(\mathcal{X} \times \mathcal{Y})^n \to \mathcal{M}; s^{(n)} \mapsto \mathrm{ELM}_{\mathcal{H}}\left(s^{(n)}\right)$ is measurable, we shall say that $\mathrm{ELM}_{\mathcal{H}}$ is *well defined*.

Empirical loss minimization (ELM)

Theorem (Finite hypothesis class)

Consider a finite hypothesis class $\mathcal{H} \subset \mathcal{M}$ realizable with respect to (q, f) such that $\mathrm{ELM}_{\mathcal{H}}$ is well defined. Let $\delta \in]0, 1[$, $\varepsilon \in \mathbb{R}_+^*$, and $n \in \mathbb{N}$ such that

$$n \geq \frac{\log(|\mathcal{H}|/\delta)}{\varepsilon}.$$

If the training sequence $S^{(n)} \overset{\text{dist.}}{\sim} (q, f)^n$, then

$$\mathbf{P}\left(\mathcal{L}_{q,f}\left(\mathrm{ELM}_{\mathcal{H}}\left(S^{(n)}\right)\right) \leq \varepsilon\right) \geq 1 - \delta. \tag{4}$$

PAC-learnability

Definition (PAC-learnability)

$\mathcal{H} \subset \mathcal{M}$ is *PAC-learnable* (*PAC*: '*probably approximately correct*') if there exist a learner $\mathbb{A} : \bigcup_{n \in \mathbb{N}^*} (\mathcal{X} \times \mathcal{Y})^n \to \mathcal{M}$ and a function $n_{\mathcal{H}} : \mathbb{R}_+^* \times]0,1[\to \mathbb{R}_+$ (called *sample-complexity*) such that:
For every $\varepsilon \in \mathbb{R}_+^*, \delta \in]0,1[$, for every distribution q over $\mathcal{X}$, and for every target function $f : \mathcal{X} \to \mathcal{Y}$, if $\mathcal{H}$ is realizable with respect to (q, f), and if the training sequence $S^{(n)} \overset{\text{dist.}}{\sim} (q, f)^n$ with $n \geq n_{\mathcal{H}}(\varepsilon, \delta)$, then

$$\mathbf{P}\left(\mathcal{L}_{q,f}\left(\mathbb{A}\left(S^{(n)}\right)\right) \leq \varepsilon\right) \geq 1 - \delta.$$

PAC-learnability

Corollary (Finite hypothesis class)

Let $\mathcal{H} \subset \mathcal{M}$ be a finite hypothesis class with well defined $\mathrm{ELM}_{\mathcal{H}}$. Then $\mathcal{H}$ is PAC-learnable using learner $\mathrm{ELM}_{\mathcal{H}}$ with sample-complexity

$$n_{\mathcal{H}}(\varepsilon, \delta) = \frac{\log(|\mathcal{H}|/\delta)}{\varepsilon}. \tag{5}$$

- Infinite hypothesis classes may also be PAC-learnable.

Noisy learning framework

Noisy learning framework; Motivation

- In many situations the output is not related to the input through a deterministic function. Here are some examples:
  - *Additive noise communication channel*: The input is the emitted signal, say x, and the output is the received signal, say Y:

    $$Y = f(x) + W,$$

    for some function f, and some random variable W called noise.
  - *General communication channel*. The channel is characterized by a probability kernel κ from $\mathcal{X}$ to $\mathcal{Y}$; so that $\mathbf{P}_{Y|X}(\mathrm{d}y|x) = \kappa(\mathrm{d}y|x)$.
- We shall now account for randomness in the relation between the input and the output.
  - We replace the target function f by a probability kernel κ from $\mathcal{X}$ to $\mathcal{Y}$;
  - this means that the output is a random variable Y with conditional probability distribution $\mathbf{P}_{Y|X}(\mathrm{d}y|x) = \kappa(\mathrm{d}y|x)$.

Definition (Noisy learning framework)

A *noisy learning framework* is composed of the same elements as the basic framework; except for the following points.

- The target function is replaced by a *probability kernel* κ from $\mathcal{X}$ to $\mathcal{Y}$, representing the conditional probability of output given the input. In particular, we assume for the training sequence $s^{(n)} = ((x_1, y_1), \ldots, (x_n, y_n))$ that $(y_1, \ldots, y_n)$ are sampled according to the distribution $\prod_{i=1}^{n} \kappa(\mathrm{d}y_i | x_i)$. The kernel κ as well as the input distribution q are not known to the learner.
- The *true loss* of a hypothesis h is defined by

$$\mathcal{L}_{q\kappa}(h) := \mathbf{P}(h(X) \neq Y), \tag{6}$$

where $X \overset{\text{dist.}}{\sim} q$ and $\mathbf{P}_{Y|X} = \kappa$. We shall denote by $q\kappa$ the probability distribution of (X, Y), which is called *mixture* of κ with respect to q.

Definition (Noisy learning framework; cont'd)

- We shall say that the *training sequence* $S^{(n)} = ((X_1, Y_1), \ldots, (X_n, Y_n))$ *is i.i.d* if $X_1, \ldots, X_n$ are i.i.d random variables with common distribution q and the random vector $Y^{(n)} = (Y_1, \ldots, Y_n)$ has conditional distribution $\mathbf{P}_{Y^{(n)}|X^{(n)}}\left(\mathrm{d}y^{(n)}|x^{(n)}\right) = \prod_{i=1}^{n} \kappa\left(\mathrm{d}y_i|x_i\right)$. This is denoted $S^{(n)} \overset{\text{dist.}}{\sim} (q\kappa)^n$. Moreover, for a learner $\mathbb{A}$, we denote

$$\mathcal{L}_{q\kappa}\left(\mathbb{A}\left(S^{(n)}\right)\right) := \mathbf{P}\left(\left(\mathbb{A}\left(S^{(n)}\right)\right)(X) \neq Y \middle| S^{(n)}\right), \tag{7}$$

  where $X \overset{\text{dist.}}{\sim} q$, $\mathbf{P}_{Y|X} = \kappa$, and (X, Y) is independent of $S^{(n)}$.
- We shall assume that for any $y \in \mathcal{Y}$, the singleton $\{y\}$ is $\mathcal{F}_{\mathcal{Y}}$-mesurable.
- If $\kappa\left(\mathrm{d}y|x\right) = \mathbf{1}\left\{y = f\left(x\right)\right\}\delta_y$ for some measurable function $f : \mathcal{X} \to \mathcal{Y}$, then f is called *target function*.

Bayes optimal hypothesis

Proposition (Bayes optimal hypothesis)

Assume that the mapping $\varphi : \mathcal{X} \to \mathbb{R}; x \mapsto \sup_{y \in \mathcal{Y}} \kappa\left(\{y\} | x\right)$ *is measurable, and that the mapping*

$$\tilde{h} : \mathcal{X} \to \mathcal{Y}; x \mapsto y \in \arg\max_{z \in \mathcal{Y}} \kappa\left(\{z\} | x\right) \tag{8}$$

(with some arbitrary tie-breaking rule) is well defined and measurable. Then for any hypothesis $h \in \mathcal{M}$*,*

$$\mathcal{L}_{q\kappa}(h) \geq \mathcal{L}_{q\kappa}(\tilde{h}) = 1 - \int_{\mathcal{X}} \varphi\left(x\right) q\left(\mathrm{d}x\right). \tag{9}$$

Such mapping $\tilde{h}$ *is called Bayes optimal hypothesis. For any* $h \in \mathcal{M}$*, the following quantity is called the excess loss of* h*:*

$$\mathcal{L}_{q\kappa}(h) - \mathcal{L}_{q\kappa}(\tilde{h}). \tag{10}$$

Proposition (Bayes optimal hypothesis for binary classification)

Assume that $\mathcal{Y} = \{0, 1\}$, *let* $\tilde{h}$ *be a Bayes optimal hypothesis, let* $X \overset{\text{dist.}}{\sim} q$, *and let* $\eta(x) := \kappa(\{1\}|x)$ *(for* $x \in \mathcal{X}$*).*

- $$\mathcal{L}_{q\kappa}(\tilde{h}) = \mathbf{E}\left[\min\left(\eta(X), 1 - \eta(X)\right)\right] \leq 1/2. \tag{11}$$
- *For any* $h \in \mathcal{M}$,

$$\mathcal{L}_{q\kappa}(h) - \mathcal{L}_{q\kappa}(\tilde{h}) = \mathbf{E}\left[|2\eta(X) - 1| \times \mathbf{1}\left\{h(X) \neq \tilde{h}(X)\right\}\right]. \tag{12}$$

Observe that the excess loss (12) of h *weights the disagreement between* h *and* $\tilde{h}$ *according to how far* η *is from* $1/2$.

Agnostic PAC-learnability

- Definition of *PAC-learnability* extends verbatim to the noisy learning framework.

Definition (Agnostic PAC-learnability)

A hypothesis class $\mathcal{H} \subset \mathcal{M}$ is called *agnostic PAC-learnable* if there exist a learner $\mathbb{A} : \bigcup_{n \in \mathbb{N}^*} (\mathcal{X} \times \mathcal{Y})^n \to \mathcal{M}$ and a function $n_{\mathcal{H}} : \mathbb{R}_+^* \times]0,1[\to \mathbb{R}_+$ (called *sample-complexity*) such that:
For every $\varepsilon \in \mathbb{R}_+^*, \delta \in]0,1[$, for every distribution q over $\mathcal{X}$, and for every probability kernel κ from $\mathcal{X}$ to $\mathcal{Y}$, if the training sequence $S^{(n)} \overset{\text{dist.}}{\sim} (q\kappa)^n$ where $n \geq n_{\mathcal{H}}(\varepsilon, \delta)$, then

$$\mathbf{P}\left(\mathcal{L}_{q\kappa}\left(\mathbb{A}\left(S^{(n)}\right)\right) \leq \inf_{h \in \mathcal{H}} \mathcal{L}_{q\kappa}(h) + \varepsilon\right) \geq 1 - \delta.$$

- An agnostic PAC-learnable class is PAC-learnable.

General learning framework

General learning framework

- Consider a noisy learning framework. Let $\mathcal{Z} := \mathcal{X} \times \mathcal{Y}$ and let $Q := q\kappa$ be the mixture of κ with respect to q. Considering a random vector $(X, Y) \overset{\text{dist.}}{\sim} Q$, then the true loss of a hypothesis $h \in \mathcal{M}$ may be denoted as

$$\begin{aligned}\mathcal{L}_Q(h) &:= \mathcal{L}_{q\kappa}(h) \\ &= \mathbf{P}\left(h(X) \neq Y\right) \\ &= \mathbf{E}\left[\mathbf{1}\left\{h(X) \neq Y\right\}\right] = \mathbf{E}\left[\ell\left(h, (X, Y)\right)\right],\end{aligned}$$

where

$$\ell : \mathcal{M} \times (\mathcal{X} \times \mathcal{Y}) \to \mathbb{R}_+; (h, (x, y)) \mapsto \mathbf{1}\left\{h(x) \neq y\right\} \tag{13}$$

is called *0-1 loss function*.

- This leads to the following general learning framework.

Definition (General learning framework)

A *general learning framework* is composed of the following elements.

- Two measurable spaces $(\mathcal{Z}, \mathcal{F}_{\mathcal{Z}})$ and $(\mathcal{M}, \mathcal{F}_{\mathcal{M}})$ called *data space* and *hypothesis space* respectively.
- A *learner* (or *learning algorithm*) is a measurable mapping

$$\mathbb{A} : \bigcup_{n \in \mathbb{N}^*} \mathcal{Z}^n \to \mathcal{M},$$

that is the mapping $\mathbb{A}$

- takes as *input* a finite sequence $s^{(n)} = (s_1, \ldots, s_n) \in \mathcal{Z}^n$ called *training sequence*;
- returns as *output* some $h \in \mathcal{M}$ called *hypothesis*.

Definition (General learning framework; cont'd)

- *The learner's objective* is to minimize the loss of the hypothesis $h = \mathbb{A}\left(s^{(n)}\right)$ defined as follows. We assume given
  - a probability measure Q over $\mathcal{Z}$ representing the *data distribution*, assumed not known to the learner;
  - a measurable function $\ell : \mathcal{M} \times \mathcal{Z} \to \mathbb{R}_+$ called *loss function* assumed known to the learner.

  The *true loss* of any $h \in \mathcal{M}$ is defined by

$$\mathcal{L}_Q(h) := \mathbf{E}\left[\ell\left(h, Z\right)\right], \quad \text{where } Z \overset{\text{dist.}}{\sim} Q. \tag{14}$$

- The *empirical loss* of $h \in \mathcal{M}$ with respect to a training sequence $s^{(n)} = (s_1, \ldots, s_n) \in \mathcal{Z}^n$ is defined by

$$L_{s^{(n)}}(h) := \frac{\sum_{i=1}^n \ell\left(h, s_i\right)}{n}. \tag{15}$$

Definition (General learning framework; cont'd)

- The training sequence $S^{(n)} = (S_1, \ldots, S_n)$ will sometimes be assumed random; in this case, we assume that $S_1, \ldots, S_n$ are i.i.d with common distribution Q. This is denoted $S^{(n)} \overset{\text{dist.}}{\sim} Q^n$. Moreover, we denote

$$\mathcal{L}_Q\left(\mathbb{A}\left(S^{(n)}\right)\right) := \mathbf{E}\left[\ell\left(\mathbb{A}\left(S^{(n)}\right), Z\right) \middle| S^{(n)}\right],$$

  where $Z \overset{\text{dist.}}{\sim} Q$ is independent of $S^{(n)}$.
- We assume given some *hypothesis class* $\mathcal{H} \subset \mathcal{M}$ which is known to the learner.
- We shall say that the data space has a *product-form* if $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ for some measurable spaces $(\mathcal{X}, \mathcal{F}_{\mathcal{X}})$ and $(\mathcal{Y}, \mathcal{F}_{\mathcal{Y}})$ and $\mathcal{M}$ is the set of measurable functions $\mathcal{X} \to \mathcal{Y}$.

- The extension of the notion of '*agnostic PAC-learnability*' to the general learning framework is straightforward.

Definition (Agnostic PAC-learnability: General framework)

$\mathcal{H} \subset \mathcal{M}$ is called *agnostic PAC-learnable* if there exist a learner $\mathbb{A} : \bigcup_{n \in \mathbb{N}^*} \mathcal{Z}^n \to \mathcal{M}$ and a function $n_{\mathcal{H}} : \mathbb{R}_+^* \times]0,1[\to \mathbb{R}_+$ (called *sample-complexity*) such that:
For every $\varepsilon \in \mathbb{R}_+^*, \delta \in]0,1[$, for every distribution Q over $\mathcal{Z}$, if the training sequence $S^{(n)} \overset{\text{dist.}}{\sim} Q^n$ where $n \geq n_{\mathcal{H}}(\varepsilon, \delta)$, then

$$\mathbf{P}\left(\mathcal{L}_Q\left(\mathbb{A}\left(S^{(n)}\right)\right) \leq \inf_{h \in \mathcal{H}} \mathcal{L}_Q(h) + \varepsilon\right) \geq 1 - \delta. \tag{16}$$

Definition (Empirical loss minimization (ELM))

The *ELM with respect to* $\mathcal{H}$ is a learner, denoted $\mathrm{ELM}_{\mathcal{H}}$, such that for any $s^{(n)} \in \mathcal{Z}^n$,

$$\mathrm{ELM}_{\mathcal{H}}\left(s^{(n)}\right) \in \arg\min_{h \in \mathcal{H}} L_{s^{(n)}}(h),$$

whenever the '$\arg\min$' set is not empty; in which case we say that $\mathrm{ELM}_{\mathcal{H}}\left(s^{(n)}\right)$ is *well defined*. We shall say that $\mathrm{ELM}_{\mathcal{H}}$ is *well defined* if the above $\arg\min$ is not empty for any $s^{(n)} \in \mathcal{Z}^n$, and the mapping $\mathcal{Z}^n \to \mathcal{M}; s^{(n)} \mapsto \mathrm{ELM}_{\mathcal{H}}\left(s^{(n)}\right)$ is measurable.

Lemma

The minimum $\min_{h \in \mathcal{H}} L_{s^{(n)}}(h)$ *is achievable in either of the following cases:*

- *The hypothesis class* $\mathcal{H}$ *is a finite set.*
- *The loss function* ℓ *has a finite range. (This is in particular the case for the noisy learning framework.)*

Example (Linear fitting as learning problem)

The linear fitting problem may be described as a learning problem as follows:

- The input space is $\mathcal{X} = \mathbb{R}$, the output space is $\mathcal{Y} = \mathbb{R}$, a hypothesis is a measurable function $h : \mathbb{R} \to \mathbb{R}$, and the target function is an affine function $f : \mathbb{R} \to \mathbb{R}$; i.e. f is in the hypothesis class

$$\mathcal{H}_{\mathrm{a}} = \{\text{functions } h : \mathbb{R} \to \mathbb{R}; x \mapsto \beta_0 + \beta_1 x \text{ for some reals } \beta_0, \beta_1\}.$$

- Consider the loss function ℓ defined by $\ell\left(h, (x, y)\right) = \left(y - h\left(x\right)\right)^2$, for any hypothesis h and any reals x, y.
- Empirical loss of a hypothesis $h : x \mapsto \beta_0 + \beta_1 x$ with respect to a training sequence $s^{(n)} = \left((x_1, y_1), \ldots, (x_n, y_n)\right)$:

$$L_{s^{(n)}}(h) = \frac{\sum_{i=1}^{n} \ell\left(h, (x_i, y_i)\right)}{n} = \frac{\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2}{n}.$$

It follows that $\mathrm{ELM}_{\mathcal{H}_{\mathrm{a}}}\left(s^{(n)}\right)$ leads to the least-squares fitting.

Lemma (ELM versus optimal)

Let $s^{(n)} \in \mathcal{Z}^n$ be such that $\hat{h} := \mathrm{ELM}_{\mathcal{H}}\left(s^{(n)}\right)$ is well defined, then

$$\mathcal{L}_Q(\hat{h}) - \inf_{h \in \mathcal{H}} \mathcal{L}_Q(h) \leq 2 \sup_{h \in \mathcal{H}} \left| L_{s^{(n)}}(h) - \mathcal{L}_Q(h) \right|.$$

Definition (ε-representative)

We say that a training sequence $s^{(n)} \in \mathcal{Z}^n$ is *ε-representative* if

$$\sup_{h \in \mathcal{H}} \left| L_{s^{(n)}}(h) - \mathcal{L}_Q(h) \right| \leq \varepsilon.$$

Corollary

Let $s^{(n)} \in \mathcal{Z}^n$ be such that $\hat{h} := \mathrm{ELM}_{\mathcal{H}}\left(s^{(n)}\right)$ is well defined. If $s^{(n)}$ is $\frac{\varepsilon}{2}$-representative , then

$$\mathcal{L}_Q(\hat{h}) \leq \inf_{h \in \mathcal{H}} \mathcal{L}_Q(h) + \varepsilon.$$

Definition (Uniform-convergence property)

We say that a hypothesis class $\mathcal{H}$ has the *uniform-convergence property* if there exists a function $n_{\mathcal{H}}^{\mathrm{UC}} : \mathbb{R}_+^* \times]0,1[\to \mathbb{R}_+$ (called *uniform-convergence sample-complexity*) such that:
For every $\varepsilon \in \mathbb{R}_+^*, \delta \in]0,1[$, for every data distribution Q over $\mathcal{Z}$, if the training sequence $S^{(n)} \overset{\text{dist.}}{\sim} Q^n$ where $n \geq n_{\mathcal{H}}^{\mathrm{UC}}(\varepsilon, \delta)$, then

$$\mathbf{P}_* \left(\sup_{h \in \mathcal{H}} |L_{S^{(n)}}(h) - \mathcal{L}_Q(h)| \leq \varepsilon \right) \geq 1 - \delta, \tag{18}$$

where $\mathbf{P}_*$ is the inner probability defined by $\mathbf{P}_*(B) = \sup\{\mathbf{P}(A) : A \subset B, A \in \mathcal{A}\}$.

Corollary (Learnability versus uniform-convergence)

Assume that a hypothesis class $\mathcal{H}$ has the uniform-convergence property with uniform-convergence sample-complexity $n_{\mathcal{H}}^{\mathrm{UC}}$, and that the $\mathrm{ELM}_{\mathcal{H}}$ learner is well defined.
Then $\mathcal{H}$ is agnostic PAC-learnable using $\mathrm{ELM}_{\mathcal{H}}$ learner with sample-complexity

$$n_{\mathcal{H}}(\varepsilon, \delta) := n_{\mathcal{H}}^{\mathrm{UC}}(\varepsilon/2, \delta). \tag{19}$$

Finite hypothesis class

Lemma

Assume that $\ell : \mathcal{M} \times \mathcal{Z} \to [0,1]$, *let* $S^{(n)} \overset{\text{dist.}}{\sim} Q^n$, *and* $\delta \in]0,1[$. *Then for any finite hypothesis class* $\mathcal{H} \subset \mathcal{M}$,

$$\mathbf{P}\left(\sup_{h\in\mathcal{H}} |L_{S^{(n)}}(h) - \mathcal{L}_Q(h)| \leq \sqrt{\frac{\log(2|\mathcal{H}|/\delta)}{2n}}\right) \geq 1-\delta. \tag{20}$$

Finite hypothesis class

Corollary

Assume that $\ell : \mathcal{M} \times \mathcal{Z} \to [0,1]$, *and let* $\mathcal{H} \subset \mathcal{M}$ *be a finite hypothesis class. Then*

- $\mathcal{H}$ *has the uniform-convergence property with uniform-convergence sample-complexity*

$$n_{\mathcal{H}}^{\mathrm{UC}}(\varepsilon, \delta) = \frac{\log(2|\mathcal{H}|/\delta)}{2\varepsilon^2}.$$

- *If moreover* $\mathrm{ELM}_{\mathcal{H}}$ *is well defined, then* $\mathcal{H}$ *is agnostic PAC-learnable using* $\mathrm{ELM}_{\mathcal{H}}$ *learner with sample-complexity* $n_{\mathcal{H}}(\varepsilon, \delta) := n_{\mathcal{H}}^{\mathrm{UC}}(\varepsilon/2, \delta)$.

Bibliography

Bibliography

[1] L. Berrah, B. Błaszczyszyn, L. Darlavoix, and M. K. Karray.
Python softwares for machine and deep learning - Tutorial, 2023.
https://mohamedkadhem.com/data-science/.

[2] B. Błaszczyszyn, L. Darlavoix, and M. K. Karray.
Data science: From multivariate statistics to machine and deep learning.
Book in preparation, 2023.

[3] S. Shalev-Shwartz and S. Ben-David.
Understanding machine learning: From theory to algorithms.
Cambridge University Press, 2014.

[4] M. Talagrand.
Sharper bounds for Gaussian and empirical processes.
The Annals of Probability, 22(1), 1994.

[5] A. W. van der Vaart and J. A. Wellner.
Weak convergence and empirical processes with applications to statistics.
Springer, 1996.