Elements of Statistical Machine Learning

Alessandro Rudi, Pierre Gaillard

April 28, 2020

1 Introduction

In this class we will introduce the main elements of PAC-Learning. PAC-Learning (Probably Approximately Correct Learning) is a theoretical framework for analysing machine learning algorithms. It was introduced by Leslie Valiant in 1984.

Notation and reminder

- Training set: $D_n = \{(X_i, Y_i)\}_{1 \le i \le n}$. The data points (X_i, Y_i) are i.i.d. random variables in $\mathcal{X} \times \mathcal{Y}$ and follow a distribution $\mathcal{P}$. $\mathcal{X}$ is the input set (typically $\mathbb{R}^d$) and $\mathcal{Y}$ the output set (typically $\{0,1\}$ for regression or $\mathbb{R}$ for classification).
- A learning algorithm is a function $\mathcal{A}$ that maps a training set D_n to an estimator $\hat{f}_n : \mathcal{X} \to \mathcal{Y}$:

$$\mathcal{A} : \underbrace{\cup_{n \ge 0} (\mathcal{X} \times \mathcal{Y})^n}_{\text{training set}} \mapsto \underbrace{\mathcal{Y}^{\mathcal{X}}}_{\text{estimator}} .$$

  We denote $\hat{f}_n = \mathcal{A}(D_n)$ the estimator (which is a random variable in $\mathcal{Y}^{\mathcal{X}}$). Sometimes the prediction set can differ from the output set. For instance, in classification in $\{0,1\}$ we might want to predict probability in $[0,1]$ for the output to belong to class 1.
- Loss function to measure the performance: $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$
- Risk of an estimator (statistical risk)

$$R(f) := \mathbb{E}_{(X,Y)\sim\mathcal{P}} \Big[\ell(f(X), Y) \Big] = \mathbb{E}\big[\ell(f(X), Y) | f \big]$$

- Target function: any function f^* such that $R(f^*) = \inf_{f : X \to Y} R(f)$, i.e.,

$$f^* \in \operatorname*{arg\,min}_{f : X \to Y} R(f),$$

  where $f : X \to Y$ is the set of measurable functions.

Definition 1 (Fundamental problem of Supervised Learning)**.** *The goal of supervised learning is to estimate f^* given only D_n and ℓ.*

To quatify the goal above we introduce the *excess risk* $\mathcal{R}(\hat{f}_n)$ which measures how close is a given predictor $\hat{f}_n$, to the best possible f^*, in terms of expected risk R, i.e., in terms of *average error on new examples*:

$$\mathcal{R}(\hat{f}_n) := R(\hat{f}_n) - R(f^*).$$

Remark 1. *Note that $R(\hat{f}_n)$ and then $\mathcal{R}(\hat{f}_n)$ are random variables, since $\hat{f}_n$ depends on the dataset D_n*

Definition 2 (Consistency)**.** *Let $\delta \in (0,1]$. The algorithm $\mathcal{A}$ is consistent (i.e., it is a proper* learning algorithm*)*

$$\lim_{n\to\infty} \mathbb{E}_{D_n}\ \mathcal{R}(\hat{f}_n) = 0.$$

We can ask also more: The algorithm $\mathcal{A}$ is strongly consistent*, i.e., the following holds with probability* 1*,*

$$\lim_{n\to\infty} \mathcal{R}(\hat{f}_n) = 0.$$

We can define more quantitative versions of the requirements above, that are useful to characterize how precise are the predictions

Definition 3 (Learning Rates)**.** *The sequence $(e_n)_{n\in\mathbb{N}}$ is a learning rate in expectation, if*

$$\mathbb{E}_{D_n}\ \mathcal{R}(\hat{f}_n) \leq e_n, \quad \forall\ n \in \mathbb{N}.$$

Given $\delta \in (0,1]$, a sequence $(p_{n,\delta})_{n\in\mathbb{N}}$ is a learning rate in probability, if

$$\mathbb{P}_{D_n}(\mathcal{R}(\hat{f}_n) > p_{n,\delta}) \leq \delta, \quad \forall\ n \in \mathbb{N}.$$

2 Empirical Risk Minimization

A classical way to estimate f^* is via *empirical risk minimization*. Let $\mathcal{F}$ be a set of functions called *hypothesis space* containing some candidate estimators of choice, the estimator is defined as

$$\hat{f}_n := \arg\min_{f\in\mathcal{F}} R_n(f), \quad R_n(f) = \frac{1}{n}\sum_{i=1}^{n} \ell(y_i, f(x_i)),$$

where $R_n(f)$ is the *empirical risk* and measures the average error performed by f on the training set. The intuition is that $R_n(f)$ approximates $R(f)$ (the expected error) increasingly better, when n go to infinity.
A crucial question we need to address is to understand under which conditions *empirical risk minimization* is a learning algorithm and has learning rates.
We first recall some results that will be useful for the proof.

3 Preliminary results

Lemma 1 (Union bound)**.** *Let $\mathcal{F}$ be a finite set indexing a family of sets $(A_f)_{f\in\mathcal{F}}$, then*

$$\mathbb{P}(\bigcup_{f\in\mathcal{F}} A_f) \leq \sum_{f\in\mathcal{F}} \mathbb{P}(A_f)$$

Proof. Let A, B be two sets, we have $\mathbb{P}(A\cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A\cap B)$, then

$$\mathbb{P}(A\cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A\cap B) \leq \mathbb{P}(A) + \mathbb{P}(B).$$

The result is obtained by iterating the formula above over $(A_f)_{f\in\mathcal{F}}$. □

Lemma 2 (supremum of random variables)**.** *Let $t > 0$ and $\mathcal{F}$ be a finite set indexing real random variables $(u_f)_{f\in\mathcal{F}}$, we have*

$$\mathbb{P}(\sup_{f\in\mathcal{F}} |u_f| > t) \leq \sum_{f\in\mathcal{F}} \mathbb{P}(|u_f| > t).$$

Proof. Since $\sup_{f\in\mathcal{F}}|u_f| \leq t$ is equivalent to $|u_f| \leq t \ \ \forall f \in \mathcal{F}$, we have

$$\{\sup_{f\in\mathcal{F}}|u_f| \leq t\} = \cap_{f\in\mathcal{F}}\{|u_f| \leq t\},$$

where we denote by $\{condition\}$ the event (subset of the sample space) satisfying *condition*. Since $\mathbb{P}(A) = 1 - \mathbb{P}(A^c)$, for any set A, with A^c the complement of A, we have

$$\mathbb{P}(\sup_{f\in\mathcal{F}}|u_f| > t) = 1 - \mathbb{P}(\sup_{f\in\mathcal{F}}|u_f| \leq t) = 1 - \mathbb{P}(\cap_{f\in\mathcal{F}}\{|u_f| \leq t\}) = \mathbb{P}(\cup_{f\in\mathcal{F}}\{|u_f| > t\}).$$

Finally by using the union bound from Lemma 1

$$\mathbb{P}(\cup_{f\in\mathcal{F}}\{|u_f| > t\}) \leq \sum_{f\in\mathcal{F}}\mathbb{P}(|u_f| > t),$$

□

Proposition 1. *Let x be a random variable and let f, g be real functions, with $t \in \mathbb{R}$ such that $f(x) \geq g(x) > t$ almost surely, then*

$$\mathbb{P}(g(x) > t) \leq \mathbb{P}(f(x) > t)$$

Proof. The result is obtained considering that the event $G := \{g(x) > t\}$ satisfies $G \subset F$ with F the event $F := \{f(x) > t\}$. So $\mathbb{P}(G) \leq \mathbb{P}(F)$. □

3.1 Bernstein inequality

Lemma 3 (exponential moments of bounded random variables)**.** *Let u be a random variable such that $|u| \leq B$, for $B > 0$ almost surely and $\mathbb{E}u = 0$. Define $\sigma^2 := \mathbb{E}u^2$. Let $0 \leq \theta < 1/B$, then*

$$\mathbb{E}e^{\theta u} \leq e^{\frac{\theta^2\sigma^2}{2(1-\theta B)}}.$$

Proof. First note that for $k \geq 2$

$$\mathbb{E}u^k \leq B^{k-2}\mathbb{E}u^2 = B^{k-2}\sigma^2.$$

By Taylor expansion, we have

$$\mathbb{E}e^{\theta u} = 1 + \underbrace{\mathbb{E}\theta u}_{:=0} + \sum_{k=2}^{\infty}\frac{\theta^k\mathbb{E}u^k}{k!} \leq 1 + \frac{\theta^2\sigma^2}{2}\sum_{k=2}^{\infty}\frac{\theta^{k-2}B^{k-2}}{k!/2} \leq 1 + \frac{\theta^2\sigma^2}{2}\sum_{k=2}^{\infty}\theta^{k-2}B^{k-2} = 1 + \frac{\theta^2\sigma^2}{2(1-\theta B)}.$$

Finally, since $1 + x \leq e^x$ for any x we have

$$\mathbb{E}e^{\theta u} \leq 1 + \frac{\theta^2\sigma^2}{2(1-\theta B)} \leq e^{\frac{\theta^2\sigma^2}{2(1-\theta B)}}.$$

□

Lemma 4 (Bernstein inequality for random variables)**.** *Let $u, u_1, \ldots, u_n$ be independently and identically distributed random variables, such that $\mathbb{E}u = 0$, and $|u| \leq B$ almost surely, for $B > 0$. Define $\sigma^2 := \mathbb{E}u^2$. Let $t > 0$, the following holds*

$$\mathbb{P}(\frac{1}{n}\sum_{i=1^n}u_i > t) \leq e^{-\frac{t^2n/2}{\sigma^2+Bt}},$$

Proof. Let $0 < \theta < 1/B$,

$$\mathbb{P}(\frac{1}{n}\sum_{i=1}^{n} u_i > t) = \mathbb{P}(\theta \sum_{i=1}^{n} u_i > \theta n t) = \mathbb{P}(e^{\theta \sum_{i=1}^{n} u_i} > e^{\theta n t}) \leq \frac{\mathbb{E} e^{\theta \sum_{i=1}^{n} u_i}}{e^{\theta n t}},$$

where in the last step we used the Markov inequality. Now since $u_1, \ldots, u_n$ are independent, we have

$$\mathbb{E} e^{\theta \sum_{i=1}^{n} u_i} = \prod_{i=1}^{n} \mathbb{E} e^{\theta u_i} \leq e^{\frac{\theta^2 \sigma^2 n}{2(1-\theta B)}},$$

where in the last step we used Lemma 3. So we have

$$\mathbb{P}(\frac{1}{n}\sum_{i=1}^{n} u_i > t) \leq e^{\frac{\theta^2 \sigma^2 n}{2(1-\theta B)} \; - \; \theta n t}.$$

Since the inequality above holds for any θ in $[0, B)$, we will take a θ that makes the exponent negative, in particular the one satisfying $\theta n t = \theta^2 \sigma^2 n/(1 - \theta B)$, that is $\theta := t/(\sigma^2 + Bt) < 1/B$ obtaining the desired result. □

Corollary 1. *Under the same assumptions of Lemma 4*

$$\mathbb{P}(|\frac{1}{n}\sum_{i=1}^{n} u_i| > t) \leq 2e^{-\frac{t^2 n/2}{\sigma^2 + Bt}},$$

Proof. For a random variable z, by applying Lemma 1, we have

$$\mathbb{P}(|z| > t) = \mathbb{P}(\{z > t\} \cup \{-z > t\}) \leq \mathbb{P}(z > t) + \mathbb{P}(-z > t).$$

The final result is obtained by setting $z = \frac{1}{n}\sum_{i=1}^{n} u_i$ and using two times the Bernstein inequality. □

4 Consistency and Learning rates for Empirical Risk Minimization

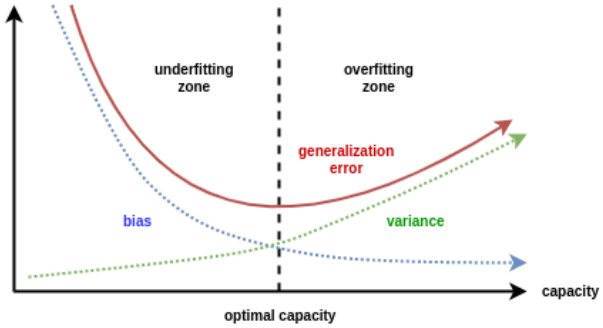
The goal of this section is to prove consistency and learning rates for empirical risk minimization. Key step here is to decompose the excess risk as:

$$\mathbb{E}\big[R(\hat{f}_n)\big] - R^* = \underbrace{\left(\mathbb{E}\big[R(\hat{f}_n)\big] - \inf_{f \in \mathcal{F}} R(f)\right)}_{\text{Estimation error}} + \underbrace{\left(\inf_{f \in \mathcal{F}} R(f) - R^*\right)}_{\text{Approximation error}}.$$

The *bias term* (or *approximation error*) depends on f^* and $\mathcal{F} \subseteq \mathcal{Y}^{\mathcal{X}}$ but not on $\hat{f}_n, D_n$. To control it, we must make some assumption on ρ. It is possible to prove consistency it without assumptions, but assumptions are needed to get rates of convergence.

The *variance term* (or *estimation error*) depends on $D_n, \mathcal{F}$, and $\hat{f}_n$. We can bound this term making very mild or no assumption on the data distribution $\mathcal{P}$. These are the type of results we are going to prove in this lecture.

In the picture below the effect of the capacity $|\mathcal{F}|$ is analyzed on the variance term, bias term and total excess risk, for a fixed dataset. In particular there are two important regimes *underfitting* and *overfitting*.



The algorithm is in a regime of underfitting when the capacity of the hypothesis space is too small, i.e., the chosen functions are not able to approximate well f^*, but at the same time there are only few function in $|\mathcal{F}|$ and then the variance is small. On the other side, when we select a very large hypothesis space, then likely there exists a function $f \in \mathcal{F}$ very close to f^*, inducing a small bias, but at the same time since $|\mathcal{F}|$ is big, the variance is large.
Obviously there are cases that are not represented by the graph above, consider for example the very lucky case $\mathcal{F} = \{f^*\}$, or more generally $\mathcal{F} = \{f^*\} \cup \mathcal{F}_0$.

4.1 Bounding the variance term: PAC bounds

As noted above, the estimator $\hat{f}_n$ is a random variable. A way to deal with this randomness is to consider the expectation of $R(\hat{f}_n)$. But this is limited: it makes statements about the risk on average. A finer control over the excess risk can be stated in terms of a probabilistic statement: a PAC bound (probably approximately correct).

Definition 4. *We say that $\hat{f}_n$ is ε-accurate with confidence $1-\delta$ of (ε, δ)-PAC if*

$$\mathbb{P}_{D_n}\left\{R(\hat{f}_n) - \inf_{f\in\mathcal{F}} R(f) > \varepsilon\right\} < \delta\,.$$

Now we have

$$R(\hat{f}_n) - \inf_{f\in\mathcal{F}} R(f) \quad = \quad [R(\hat{f}_n) - \inf_{f\in\mathcal{F}} R_n(f)] \quad + \quad [\inf_{f\in\mathcal{F}} R_n(f) - \inf_{f\in\mathcal{F}} R(f)].$$

Noting that the definition of $\hat{f}_n$ is to be the minimum of $R_n(f)$ over $\mathcal{F}$, then for the first term on the right hand side of the equation above, we have

$$R(\hat{f}_n) - \inf_{f\in\mathcal{F}} R_n(f) = R(\hat{f}_n) - R_n(\hat{f}_n) \leq \sup_{f\in\mathcal{F}} |R(f) - R_n(f)|.$$

For the second term, recalling that $\inf_{z\in Z} a(z) - \inf_{z\in Z} b(z) \leq \sup_{z\in Z} |a(z) - b(z)|$, we have

$$\inf_{f\in\mathcal{F}} R_n(f) - \inf_{f\in\mathcal{F}} R(f) \leq \sup_{f\in\mathcal{F}} |R(f) - R_n(f)|.$$

So finally,

$$R(\hat{f}_n) - \inf_{f\in\mathcal{F}} R(f) \leq 2 \sup_{f\in\mathcal{F}} |R(f) - R_n(f)|. \tag{1}$$

4.2 Bounds in probability

Here we assume

- **Assumption 1** $\mathcal{F}$ is a set of finite cardinality
- **Assumption 2** there exists $B > 0$ such that $\ell(y, y') \leq B$ for any $y, y' \in Y$.

Note that for any $f \in \mathcal{F}$, $R(f) - R_n(f)$ is a random variable since it depends on D_n. Let $t > 0$, by Prop. 1, we have

$$\mathbb{P}(R(\hat{f}_n) - \inf_{f \in \mathcal{F}} R(f) > 2t) \leq \mathbb{P}(2 \sup_{f \in \mathcal{F}} |R(f) - R_n(f)| > 2t) \tag{2}$$

$$\leq \sum_{f \in \mathcal{F}} \mathbb{P}(2|R(f) - R_n(f)| > 2t) \tag{3}$$

$$= \sum_{f \in \mathcal{F}} \mathbb{P}(|R(f) - R_n(f)| > t) \tag{4}$$

where the last step is due to Lemma 2. To further bound the inequality above we need to study the probability of the event $\{|R(f) - R_n(f)| > t\}$. Given $f \in \mathcal{F}$, denote by v_i the random variable defined as $v_i := \ell(y_i, f(x_i)) - R(f)$ for $i \in \{1, \ldots, n\}$. Then we have

$$|R_n(f) - R(f)| = |\frac{1}{n} \sum_{i=1}^{n} v_i|.$$

Now note that $\mathbb{E}v_i = 0$, and that $v_1, \ldots, v_n$ are independent and identically distributed. Moreover, by Assumption 2, $|v_i| \leq B$ almost surely and that $\mathbb{E}v_i^2 \leq B^2$. So, by applying the Bernstein inequality in Lemma 4 (see also the subsequent corollaries), we have

$$\mathbb{P}(|\frac{1}{n} \sum_{i=1}^{n} v_i| > s) \leq 2e^{-\frac{t^2 n/2}{B^2 + Bt}}. \tag{5}$$

To conclude, by Eq. (1), (4), (5) and Prop. 1

$$\mathbb{P}_{D_n}\left\{R(\hat{f}_n) - \inf_{f \in \mathcal{F}} R(f) > 2t\right\} \leq \mathbb{P}_{D_n}\left\{2 \sup_{f \in \mathcal{F}} |R(f) - R_n(f)| > 2t\right\}$$

$$\leq \sum_{f \in \mathcal{F}} \mathbb{P}_{D_n}\left\{2 \sup_{f \in \mathcal{F}} |R(f) - R_n(f)| > 2t\right\}$$

$$\leq \sum_{f \in \mathcal{F}} 2e^{-\frac{t^2 n/2}{B^2 + Bt}} = 2|\mathcal{F}|e^{-\frac{t^2 n/2}{B^2 + Bt}}.$$

Now note that when $t \leq B$, then $2|\mathcal{F}|e^{-\frac{t^2 n/2}{B^2 + Bt}} \leq 2|\mathcal{F}|e^{-\frac{t^2 n}{B^2}}$. Let $\delta \in (0, 1]$, when $n \geq \log \frac{2|\mathcal{F}|}{\delta}$, select $t = \sqrt{\frac{B^2 \log \frac{2|\mathcal{F}|}{\delta}}{n}}$, then $t \leq B$ and so we have the (t, δ)-PAC bound

$$\mathbb{P}_{D_n}\left\{R(\hat{f}_n) - \inf_{f \in \mathcal{F}} R(f) > \sqrt{\frac{4B^2 \log \frac{2|\mathcal{F}|}{\delta}}{n}}\right\} \leq \delta,$$

or equivalently: the following holds with probability at least $1 - \delta$

$$R(\hat{f}_n) - \inf_{f \in \mathcal{F}} R(f) \leq \sqrt{\frac{4B^2 \log \frac{2|\mathcal{F}|}{\delta}}{n}}.$$

5 Bounding the Bias

Bias term depends on the specific properties of the function $f^\star$ we are going to learn and the function space $\mathcal{F}$ we have chosen. Assumptions must be done on such two objects, to quantify the bias term.

Example. Let $X = \mathbb{R}, Y = \mathbb{R}$. Let ℓ satisfy $|\ell(y', y) - \ell(y'', y)| \leq C|y' - y''|$ for $C > 0$. Assume that function $f^\star$ satisfies $f^\star : [-1, 1] \to \mathbb{R}$ with

$$f^\star(x) = \sum_{k=0}^{\infty} \beta_k x^k,$$

for a sequence $(\beta_k)_{k\in\mathbb{N}}$ such that $\sum_{k=0}^{\infty} |\beta_k| := S < \infty$. Let $R > 0, p \in \mathbb{N}$ and define

$$\mathcal{F} = \left\{ f(x) = \sum_{k=1}^{p} \alpha_k x^k \;\middle|\; \alpha_k \in [-R, R] \right\}.$$

Denote by $\tilde{f}_p$ the function $\tilde{f}_p(x) = \sum_{k=0}^{p} \beta_k x^k$ When $R \geq S$ $\tilde{f}_p \in \mathcal{F}$, then

$$\begin{aligned}
\inf_{f\in\mathcal{F}} R(f) - R(f^*) \leq R(\tilde{f}) - R(f^*) &= \mathbb{E}[\ell(\tilde{f}_p(x), y) - \ell(\tilde{f}(x), y)] && (6)\\
&\leq C\mathbb{E}|\tilde{f}_p(x) - f^\star(x)| && (7)\\
&\leq C\mathbb{E} \sum_{k=p+1}^{\infty} \beta_k x^k && (8)\\
&\leq C \sum_{k=p+1}^{\infty} |\beta_k|. && (9)
\end{aligned}$$

6 Optional exercises

Ex. 1. $X = \mathbb{R}$, $Y = \mathbb{R}$. Loss $\ell(y', y) = \min(|y' - y|, B)$, with $B > 0$. Let $p \in \mathbb{N}$ and $\varepsilon > 0$, $T = 1/\varepsilon$,

$$\mathcal{F}_{T,\varepsilon} = \{f(x) = \sum_{k=1}^{p} \alpha_k x^k \mid \alpha_k \in \{-T, -T + \varepsilon, -T + 2\varepsilon, \ldots, T - 2\varepsilon, T - \varepsilon, T\}\}.$$

Moreover assume that $f^\star(x) = \sum_{k=0}^{\infty} \beta_k x^k$, with $\sum_{k=r+1}^{\infty} \beta_k \leq Cr^{-\gamma}$, with $C, \gamma > 0$.

1. Compute an upper bound for the variance term,
2. for the bias term,
3. for the excess risk.
4. Given the number of examples n, which is a good choice for ε, p depeding on n, in order to guarantee the fastest rate possible for the upper bound?

Ex. 2. (Non-discrete hypothesis spaces) Can we generalize the analysis above, considering a non-discrete $\mathcal{F}_T$ defined as

$$\mathcal{F} = \{f(x) = \sum_{k=1}^{p} \alpha_k x^k \mid \alpha_k \in [-T, T]\}?$$

Suggestion. Start from the analysis of the discrete one (and perform a decomposition of the form

$$R(\hat{f}) - R(f^\star) \quad \leq \quad [R(\hat{f}) - \inf_{f\in\mathcal{F}_{T,\varepsilon}} R(f)] \quad + \quad [\inf_{f\in\mathcal{F}_{T,\varepsilon}} R(f) - \inf_{f\in\mathcal{F}_T} R(f)] \quad + \quad [\inf_{f\in\mathcal{F}_T} R(f) - R(f^\star)].$$