PAC-Learning and Concentration inequalities

Pierre Gaillard

November 2018

1 Introduction

Pac-Learning (Probably Approximately Correct Learning) is a theoretical framework for analysing machine learning algorithms. It was introduced by Lesli Valiant in 1984.

Notation and reminder

- Training set: $D_n = \{(X_i, Y_i)\}_{1\leq i\leq n}$. 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 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 $\widehat{f}_n : \mathcal{X}\to\mathcal{Y}$:

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

  We denote $\widehat{f}_n = \mathcal{A}(D_n)$ 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)

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

- Since an estimator $\widehat{f}_n = \mathcal{A}(D_n)$ is random, we often consider the *expected risk* or *frequentist risk*:

$$\mathbb{E}[R(\widehat{f}_n)] = \mathbb{E}_{D_n\sim\mathcal{P}^{\otimes n}}\big[R(\widehat{f}_n)\big] = \mathbb{E}\big[\ell(f(X),Y)\big]$$

- An estimator is consistent for $\mathcal{P}$ if

$$\lim_{n\to\infty}\mathbb{E}[R(\widehat{f}_n)] = R(f^*) \qquad \text{where} \quad f^*\in\operatorname*{arg\,min}_{f\in\mathcal{Y}^{\mathcal{X}}} R(f)\,.$$

  We denote by $R^* = R(f^*)$ the optimal risk. An estimator is universally consistent if this is valid for all $\mathcal{P}$.
- Example: in classification $\mathcal{X}=\mathbb{R}^d$, $\mathcal{Y}=\{0,1\}$, $\ell(y,\widehat{y}) = \mathbb{1}\{y\neq\widehat{y}\}$, $R(f) = \mathbb{P}_{(X,Y)\sim\mathcal{P}}(f(X)\neq Y)$. kNN when $k\to\infty$ and $k/n\to 0$ is universally consistent.

Ultimate goal Minimize the risk $R(\widehat{f}_n)$ with high probability or in expectation. We can decompose the excess risk into two terms:

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

The *approximation error* depends on $\mathcal{P}$ and $\mathcal{F} \subseteq \mathcal{Y}^{\mathcal{X}}$ but not on $\widehat{f}_n$. To control it, we must make some assumption on $\mathcal{P}$. It is possible to have asymptotic result on it without assumptions (remember KNN) but assumptions are needed to get rates of convergence.

The *estimation error* depends on $\mathcal{P}, \mathcal{F}$, and $\widehat{f}_n$. We can bound this term without making any assumption on the data distribution $\mathcal{P}$. These are the type of results we are going to prove in this lecture.

1.1 PAC bounds

As noted above, the estimator $\widehat{f}_n$ is a random variable. A way to deal with this randomness is to consider the *expected risk*. 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 1. *We say that $\widehat{f}_n$ is ε-accurate with confidence $1-\delta$ of (ε,δ)-PAC if*

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

From bounds in high-probability to bounds in expectation It is worth to notice that if the risk is bounded by L and $\widehat{f}_n$ is (ε,δ)-PAC, this implies:

$$\mathbb{E}\big[\underbrace{R(\widehat{f}_n) - \inf_{f\in\mathcal{F}} R(f)}_{=:\Delta_n}\big] \le \mathbb{E}\big[\Delta_n > \varepsilon\big]\mathbb{P}\big\{\Delta_n > \varepsilon\big\} + \mathbb{E}\big[\Delta_n | \Delta_n \le \varepsilon\big]\mathbb{P}\big\{\Delta_n \le \varepsilon\big\} \le L\delta + \varepsilon\,. \tag{1}$$

Therefore, a result in high probability is stronger than a result in expectation. Tighter bound on the expected risk can be obtained from (ε,δ)-PAC bounds by using the equality for any non-negative random variable X:

$$\mathbb{E}[X] = \int_0^\infty \mathbb{P}(X > \varepsilon) d\varepsilon\,. \tag{2}$$

1.2 A simple PAC-bound in the binary classification setting

We consider the binary classification setting with $\mathcal{X} = \mathbb{R}^d$, $\mathcal{Y} = \{0,1\}$, $\ell(y,\widehat{y}) = \mathbb{1}\{y \neq \widehat{y}\}$, $R(f) = \mathbb{P}_{(X,Y)\sim\mathcal{P}}(f(X) \neq Y)$. Let $\mathcal{F} \subset \mathcal{Y}^{\mathcal{X}}$ be a finite class of models such that one of the models is perfect: $\inf_{f\in\mathcal{F}} R(f) = 0$. This is a very strong assumption.

Theorem 1. *Let $\widehat{f}_n = \arg\min_{f\in\mathcal{F}} \widehat{R}_n(f) = \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{f(X_i)\neq Y_i}$ the empirical risk minimizer. Then, for every $n \ge 1$ and $\varepsilon > 0$*

$$\mathbb{P}\big\{R(\widehat{f}_n) \ge \varepsilon\big\} \le |\mathcal{F}| e^{-n\varepsilon} \qquad (=: \delta).$$

Therefore, the empirical risk minimizer is $(\varepsilon, |\mathcal{F}|e^{-n\varepsilon})$-PAC.

Proof. Let $f^* \in \arg\min_{f\in\mathcal{F}} R(f)$. First, remark that $\widehat{R}_n(\widehat{f}_n) = 0$ since $0 \le \mathbb{E}[\widehat{R}_n(\widehat{f}_n)] \le \mathbb{E}[\widehat{R}_n(f^*)] = R(f^*) = 0$ by assumption. Let $\mathcal{G} := \{f \in \mathcal{F} : \widehat{R}_n(f) = 0\}$ then $\widehat{f}_n \in \mathcal{G}$ therefore,

$$\begin{aligned}
\mathbb{P}\big\{R(\widehat{f}_n) \ge \varepsilon\big\} &\le \mathbb{P}\big\{ \cup_{f\in\mathcal{G}} \{R(f) \ge \varepsilon\}\big\} \\
&= \mathbb{P}\big\{ \cup_{f\in\mathcal{F}} \{R(f) \ge \varepsilon, \widehat{R}_n(f) = 0\}\big\} \\
&= \mathbb{P}\big\{ \cup_{f\in\mathcal{F}:R(f)\ge\varepsilon} \{\widehat{R}_n(f) = 0\}\big\} \\
&\le \sum_{f\in\mathcal{F}:R(f)\ge\varepsilon} \mathbb{P}\big\{\widehat{R}_n(f) = 0\big\} \\
&= \sum_{f\in\mathcal{F}:R(f)\ge\varepsilon} \mathbb{P}\big\{\forall i = 1,\dots,n, \quad \mathbb{1}_{f(X_i)\neq Y_i} = 0\big\} \qquad \text{by definition of } \widehat{R}_n \\
&= \sum_{f\in\mathcal{F}:R(f)\ge\varepsilon} \mathbb{P}_{(X,Y)\sim\mathcal{P}}\big\{\mathbb{1}_{f(X)\neq Y} = 0\big\}^n \qquad \text{because data is i.i.d. with distribution } \mathcal{P} \\
&= \sum_{f\in\mathcal{F}:R(f)\ge\varepsilon} \big(1 - R(f)\big)^n \\
&\le \sum_{f\in\mathcal{F}:R(f)\ge\varepsilon} (1-\varepsilon)^n \\
&\le |F|(1-\varepsilon)^n \le |F|e^{-n\varepsilon}\,,
\end{aligned}$$

where the last inequality is because $1 - x \le e^{-x}$. □

Note that for n large, $\delta = e^{\log|\mathcal{F}| - n\varepsilon}$ can be made arbitrailly small. For a given δ, we need

$$n = \frac{\log|F| - \log\delta}{\varepsilon}$$

training samples.

Corollary 1. $\mathbb{E}[R(\widehat{f}_n)] \le \frac{1+\log|F|+\log n}{n}$.

Proof. Since the risk is bounded by $L = 1$, applying inequality (1), we have for any $\varepsilon > 0$

$$\mathbb{E}\big[R(\widehat{f}_n)\big] \le \varepsilon + \delta \le \varepsilon + |F|e^{-n\varepsilon}$$

The choice $\varepsilon = (\log|F| + \log n)/n$ concludes. □

Exercise: get rid of the $\log n$ term in the corollary by using Inequality (2) instead of Inequality (1).

2 General PAC-bounds for ERM

Goals and tools In the previous section, we made a very strong assumption $R(f^*) = 0$. Basically, this means both that $\mathbb{E}_{(X,Y)\sim\mathcal{P}}[\ell(\mathbb{E}[X|Y], Y)] = 0$ and that the optimal predictor belongs to our class of models $\mathbb{E}[Y|X] \in \mathcal{F}$. Here we want similar result without assumption on $\mathcal{P}$. We focus on (penalized) empirical risk minimization (ERM):

$$\widehat{f}_n \in \arg\min_{f\in\mathcal{F}} \widehat{R}_n(f)\,.$$

How good is ERM? Assume that we have a result which says that the empirical risk is close to the true risk with high probability uniformly over all models in $\mathcal{F}$: i.e.,

$$\mathbb{P}\Big\{\forall f \in \mathcal{F},\ \big|\widehat{R}_n(f) - R(f)\big| \leq \varepsilon\Big\} \geq 1 - \delta\,. \tag{$*$}$$

In this case ERM is a good choice, since with probability $1-\delta$

$$R(\widehat{f}_n) \leq \widehat{R}_n(f_n) + \varepsilon \leq \widehat{R}_n(f^*) + \varepsilon \leq R(f^*) + 2\varepsilon\,. \tag{3}$$

Therefore, with high probability, the risk of ERM is close to the best risk in $\mathcal{F}$. Now, the question is how to get Inequality $(*)$.

2.1 Tools needed to obtain $(*)$: concentration inequalities

Let us first better understand what $(*)$ means. By the strong law of large number

$$\widehat{R}_n(f) = \frac{1}{n}\sum_{i=1}^{n} \ell(f(X_i), Y_i) \xrightarrow[n\to\infty]{R} (f)$$

almost surely. This is unfortunately not enough to get $(*)$. We still need two properties:

a) Get the speed of convergence $\rightarrow$ Obtained thanks to concentration inequalities (Chernoff bound)

b) Get this result simultaneously for all $f \in \mathcal{F} \rightarrow$ Union bound

a) Concentration inequality The first property is partially answered by the Central Limit Theorem (CLT) if the variance of the loss exists:

$$\sqrt{n}\big(\widehat{R}_n(f) - R(f)\big) \xrightarrow[n\to\infty]{\text{law}} \mathcal{N}\big(0, \mathrm{Var}(\ell(f(X), Y)\big)\,.$$

But this is only valid asymptotically when n goes to ∞. To have the result for finite n, we use concentration inequality such as Chernoff's bound. The proof will be done in practical session.

Proposition 1 (Chernoff's inequality)**.** *Let $(Z_i)_{1\leq i\leq n}$ i.i.d. from a Bernoulli distribution with parameter $p \in [0,1]$. Then, for all $\varepsilon > 0$*

$$\mathbb{P}\left\{\frac{1}{n}\sum_{i=1}^{n} Z_i \geq p + \varepsilon\right\} \leq e^{-2n\varepsilon^2}\,.$$

b) Union bound To get the second property, if $\mathcal{F}$ is finite or countable, it is possible to use a union bound: for any family of events $(A_f)_{f\in\mathcal{F}}$

$$\mathbb{P}\big\{\cup_{f\in\mathcal{F}} A_f\big\} \leq \sum_{f\in\mathcal{F}} \mathbb{P}\{A_f\}\,. \tag{4}$$

Otherwise, we need more sophisticated tools coming from empirical process theory (such as chaining).

2.2 Finite class $\mathcal{F}$

Theorem 2. *If $\mathcal{F}$ is finite, the ERM satisfies with probability $1-\delta$*

$$\widehat{R}_n(f) \leq \inf_{f\in\mathcal{F}} R(f) + \sqrt{\frac{2(\log|\mathcal{F}| + \log(2/\delta)}{n}}\,.$$

Proof. We first show $(*)$:

$$\mathbb{P}\left\{\forall f \in \mathcal{F},\ \left|R(f) - \widehat{R}_n(f)\right| \leq \sqrt{\frac{\log|\mathcal{F}| + \log(2/\delta)}{2n}}\right\} \geq 1 - \delta\,. \tag{5}$$

Indeed,

$$\mathbb{P}\Big\{\exists f \in \mathcal{F},\ \widehat{R}_n(f) \geq R(f) + \varepsilon\Big\} \overset{(4)}{\leq} \sum_{f\in\mathcal{F}} \mathbb{P}\Big\{\widehat{R}_n(f) \geq R(f) + \varepsilon\Big\} \overset{\text{Prop. 2}}{\leq} |\mathcal{F}|e^{-2n\varepsilon^2}\,,$$

where the second inequality is by applying Chernoff's inequality with $Z_i = \ell(f(X_i), Y_i)$ and $p = R(f)$. Similarly, we can show the other way

$$\mathbb{P}\Big\{\exists f \in \mathcal{F},\ \widehat{R}_n(f) \leq R(f) - \varepsilon\Big\} \leq |\mathcal{F}|e^{-2n\varepsilon^2}\,.$$

Choosing $\delta := 2|\mathcal{F}|e^{-2n\varepsilon^2}$ i.e., $\varepsilon = (\log|\mathcal{F}| + \log(2/\delta))/(2n)$, we get using again an union bound

$$\begin{aligned}
\mathbb{P}\Big\{\forall f \in \mathcal{F}, |\widehat{R}_n - R(f)| \leq \varepsilon\Big\} &= 1 - \mathbb{P}\Big\{\exists f \in \mathcal{F}, |\widehat{R}_n - R(f)| \geq \varepsilon\Big\} \\
&\geq 1 - \mathbb{P}\Big\{\exists f \in \mathcal{F},\ \widehat{R}_n(f) \geq R(f) + \varepsilon\Big\} - \mathbb{P}\Big\{\exists f \in \mathcal{F},\ \widehat{R}_n(f) \leq R(f) - \varepsilon\Big\} \\
&\geq 1 - \delta\,,
\end{aligned}$$

which concludes the proof of (5) by substituting ε. The proof of the theorem is obtained similarly to (3): for any $f \in cF$

$$R(\widehat{f}_n) \leq \widehat{R}_n(f_n) + \varepsilon \leq \widehat{R}_n(f) + \varepsilon \leq R(f) + 2\varepsilon = \sqrt{\frac{2(\log|\mathcal{F}| + \log(2/\delta)}{n}}\,.$$

□

Example: the histogram classifier Consider the classification setting $\mathcal{Y} = \{0,1\}$, $\mathcal{X} = [0,1]^d$ and $\mathcal{Q} = (Q_j)_{1\leq j\leq m}$ a partition of $\mathcal{X}$. Let

$$\mathcal{F}_m := \big\{f : \mathcal{X} \to \{0,1\} : f(x) = \sum_{j=1}^m c_j \mathbb{1}_{x\in Q_j},\quad c_j \in \{0,1\} \forall 1 \leq j \leq m\big\}$$

be the class of classification rules that take either 0 or 1 in each of the cells of the partition $\mathcal{Q}$. Let $\widehat{f}_n$ be the estimator that predicts in each cell Q_j by doing a majority vote in cell Q_j. More formally, for all $x \in \mathcal{X}$

$$\widehat{f}_n(x) = \sum_{j=1}^m \widehat{c}_j \mathbb{1}_{x\in Q_j} \quad \text{where} \quad \widehat{c}_j = \begin{cases} 1 & \text{if} \frac{\sum_{i:X_i\in Q_j} Y_i}{\sum_{i:X_i\in Q_j} 1} \geq 1/2 \\ 0 & \text{otherwise} \end{cases}$$

Exercise: show that $\widehat{f}_n \in \arg\min_{f\in\mathcal{F}} \left\{\frac{1}{n}\sum_{i=1}^n \mathbb{1}_{Y_i \neq f(X_i)}\right\}$ is an ERM in $\mathcal{F}_m$.

Then, from Theorem 2 together with $|\mathcal{F}| = 2^m$ and (1) choosing properly δ, we have

$$\mathbb{E}\big[\widehat{R}_n(f)\big] \leq \min_{f\in\mathcal{F}_m} R(f) + 2\sqrt{\frac{m\log 2 + 2 + \log(n/2)}{n}}\,.$$

This gives a way to choose m. If we want a vanishing estimation error, we need $m/n \xrightarrow[n\to\infty]{} 0$. But m should be large to approximate the optimal Bayes classifier well.

Theorem 3. *If the cells are of uniform size,* $m \xrightarrow[n\to\infty]{} \infty$ *and* $n/m \xrightarrow[n\to\infty]{} 0$*, then*

$$\mathbb{E}\big[R(\widehat{f}_n)\big] \xrightarrow[n\to\infty]{} R^*$$

for all distribution $\mathcal{P}$*.*

This result can also be obtained by Stoke's Theorem that we saw in the KNN lecture.

2.3 Countably infinite $\mathcal{F}$

If the class of model is infinite, we need to regularize (otherwise the union bound is infinite). To do so, we need to assign a positive number $c(f)$ to each $f \in \mathcal{F}$ such that

$$\sum_{f\in\mathcal{F}} e^{-c(f)} = 1\,.$$

This $c(f)$ can be interpreted as:
- a measure of complexity;
- negative log of prior probability of f: $c(f) = \log(1/\pi(f))$ where π is a probability distribution on $\mathcal{F}$;
- length of a codeword describing f in a binary langage.

Theorem 4. *Let* $\delta > 0$*. With probability* $1-\delta$*, for all* $f \in \mathcal{F}$

$$R(f) \le \widehat{R}_n(f) + \sqrt{\frac{c(f) + \log(1/\delta)}{2n}}\,.$$

The proof left as an exercise is thanks to a union bound together with Chernoff's inequality. This theorem provides an inequality similar to $(*)$, however it is not uniform over all $f \in \mathcal{F}$. This motivates the choice of penalized ERM:

$$\widehat{f}_n \in \arg\min_{f\in\mathcal{F}} \left\{ \widehat{R}_n(f) + \sqrt{\frac{c(f) + \log(1/\delta)}{2n}} \right\}$$

in order to choose the model with the best upper-bound on the risk. With a similar analysis as previously we can show that with probability $1-\delta$:

$$R(\widehat{f}_n) \le R(f^*) + \sqrt{\frac{2(c(f^*) + \log(2/\delta))}{n}}\,.$$

where $f^* \in \arg\min_{f\in\mathcal{F}} R(f)$. This bound is useful if we were able to assign a small complexity to f^*. There is no free lunch! Furthermore, an important point to note here is that it can be NP hard to find $\widehat{f}_n$. To find it efficiently, good properties on the loss (such as convexity) are needed.

How to choose c(f)? Usually we give more weight $\pi(f)$ (or smaller complexity $c(f)$) to simpler functions $f \in \mathcal{F}$. Suppose that we encode the elements of $\mathcal{F}$ by using a binary code, we can choose $c(f) =$ codelength(f). From Kraft's inequality: $\sum_{i=1}^{\infty} e^{-c(f)} \le 1$.

Example: histogram Let us go back to the histogram example of the previous section and consider $\mathcal{F} = \cup_{m\ge1}\mathcal{F}_m$, where each $\mathcal{F}_m$ are the class of 2^m classification rules obtained from a partition of size m. We can encode any function $f \in \mathcal{F}$ using a binary code as follows:
- use m bits to encode the smallest k such that $f \in \mathcal{F}_m$
- use $m = \log|\mathcal{F}_m|$ bits to encode which of the 2^m histograms it corresponds to in $\mathcal{F}_m$.

Then, from Kraft's inequality together with Theorem 4, with probability $1-\delta$ for all $m \ge 1$ and all $f \in \mathcal{F}_m$, we have

$$R(f) \le \widehat{R}_n(f) + \sqrt{\frac{2m\log 2 + \log(1/\delta)}{2n}}\,.$$