Introduction to Machine Learning (67577)

Recitation 8
PAC Learnability & VC-Dimension

Second Semester, 2021

Contents

1 The PAC Model

The basic setup discussed when considering the PAC framework of learnability is as follows. Given a domain set $\mathcal{X}$ and a response/label set $\mathcal{Y}$ that consists of item receiving one of two items (i.e $\{\pm 1\}$ or $\{0,1\}$) we want to learn some mapping function $\mathcal{X} \mapsto \mathcal{Y}$. We then restrict ourselves to specific hypothesis classes which are sets of function $\mathcal{X} \mapsto \mathcal{Y}$: $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$.

Then, given a training set $S = \{(\mathbf{x}_i, y_i)\}_{i=1}^m$, which is assumed to be sampled *i.i.d* (independently and identically distributed) from some unknown distribution $\mathcal{D}$, we want to select an hypothesis $h \in \mathcal{H}$. We refer to this hypothesis as the prediction rule. To determine which hypothesis to retrieve we need to define some learning algorithm $\mathcal{A} : (\mathcal{X} \times \mathcal{Y})^m \to \mathcal{H}$ that given the training set S will choose some hypothesis $\mathcal{A}(S)$.

In order to choose an hypothesis we supply the algorithm some way to evaluate each hypothesis. This is done in the form of some loss function $L : ((\mathcal{X} \times \mathcal{Y})^m, \mathcal{H}) \to \mathbb{R}_{\geq 0}$. In the case of classification problems we often use the mis-classification error, for which we then define:

- The **empirical risk** of the hypothesis/classifier h as:

$$L_S(h) := \frac{1}{m} \left|\{i \in [m] : h(\mathbf{x}_i) \neq y_i\}\right|$$

- The **generalization error** of the hypothesis/classifier h as:

$$L_{\mathcal{D}}(h) := \mathbb{E}_{\mathbf{x} \sim \mathcal{D}}[h(\mathbf{x}) \neq f(\mathbf{x})]$$

  where f is the true labeling function that maps $\mathcal{X}$ to $\mathcal{Y}$.

For the true labeling function f we say that the realizability assumption holds if f is indeed in the hypothesis class $\mathcal{H}$ that we have decided to work with.

1.1 PAC Learnability

Then, using the above setup we can define what does Probably Approximately Correct Learning means.

Definition 1.1 — PAC learnability. An hypothesis class $\mathcal{H}$ is *PAC learnable* if there exists a function $m_{\mathcal{H}} : (0,1)^2 \to \mathbb{N}$ and a learning algorithm $\mathcal{A}$ such that:

- For every $\varepsilon, \delta \in (0,1)$
- For every distribution $\mathcal{D}$ over $\mathcal{X}$
- For every labeling function $f : \mathcal{X} \to \{0,1\}$

if the realizable assumption holds with respect to $\mathcal{H}, \mathcal{D}, f$, when running the learning algorithm $\mathcal{A}$ on $m \geq m_{\mathcal{H}}(\varepsilon, \delta)$ *i.i.d* samples drawn from $\mathcal{D}$ and labeled by f, the algorithm returns a hypothesis $h_S = \mathcal{A}(S)$ such that:

$$\mathbb{P}_{S \sim \mathcal{D}^m}[L_{\mathcal{D}}(h_S) \leq \varepsilon] \geq 1 - \delta$$

We refer to:

- ε as the *accuracy parameter*, which corresponds to the “approximately correct“ in PAC.

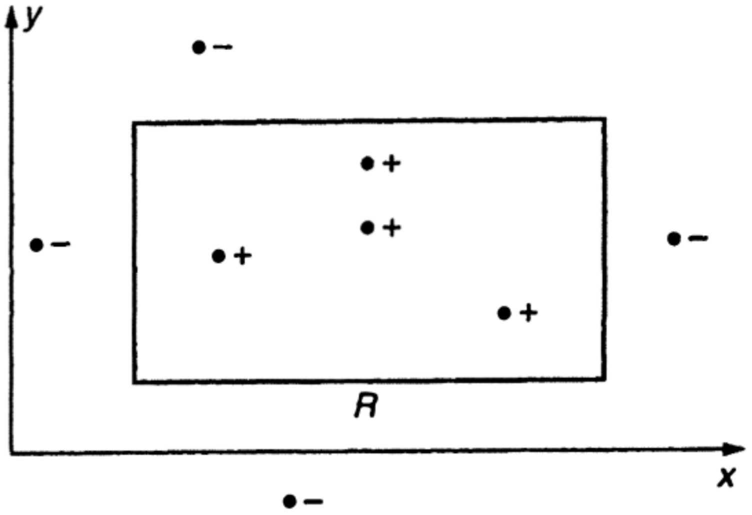
- δ as the *confidence parameter*, which corresponds to the “probably“ in PAC.
- $m_{\mathcal{H}}$ as the *sample complexity*, which for every ε, δ specifies the number of samples required.

Definition 1.2 Let $\mathcal{H}$ be an hypothesis class. If there exists an algorithm $\mathcal{A}$ that satisfies 1.1 then $\mathcal{A}$ is said to be *a learning algorithm for $\mathcal{H}$*.

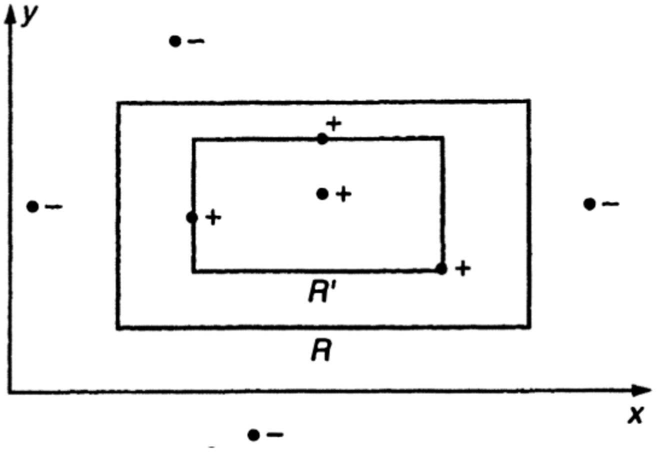
So, in order to show that a class is PAC learnable we need to show that there is (1) a learning algorithm and (2) a function $m_{\mathcal{H}}$ with the property specified in 1.1. As seen in class, an algorithm that returns an ERM hypothesis is a learning algorithm for $\mathcal{H}$ meaning that $\mathcal{H}$ is PAC learnable.

1.2 Rectangle Learning Game

The following example is taken from "An Introduction to Computational Learning Theory" by Kearns and Vazirani. Consider the scenario of learning an unknown axis-aligned rectangle R in $\mathbb{R}^2$: Points are scattered in $\mathbb{R}^2$ and are labeled as positive or as negative. There is some rectangle R such that all positive points are inside the rectangle and all negative points are outside it. Given a sample $S = \{(\mathbf{x}_i, y_i)\}_{i=1}^{m}$ where points are drawn according to some unknown (to us) probability distribution $\mathcal{D}$, the goal is to choose some **hypothesis** rectangle R' that is as similar as possible to the real rectangle R. In addition, we want to achieve this with as few training samples as possible. To measure how good a given R' is we will ask what is the probability of a randomly chosen sample, drawn according to $\mathcal{D}$, to fall in the area between R' and R.



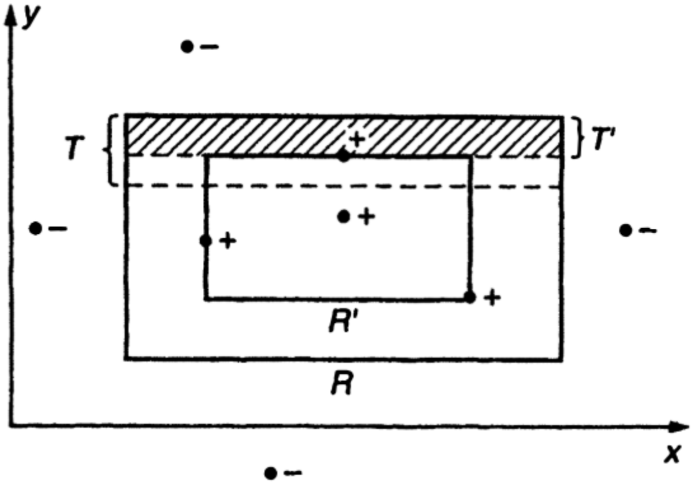
Let us show that this is a PAC learnable problem by describing a learning algorithm and a sample complexity function. The learning algorithm we will use is to simply return the axis-aligned rectangle R' which gives the **tightest fit to the positive examples**. That is, the rectangle with the smallest area that includes all of the $+1$ samples and none of the -1 samples. If there are no positive examples in S then $R' = \emptyset$.



We will show that for any true rectangle R, for any distribution $\mathcal{D}$ and for any $\varepsilon, \delta \in (0,1)$ if we correctly select m then with probability of at least $1-\delta$, the rectangle R' returned by the algorithm above has an error of at most ε with respect to R and $\mathcal{D}$

$$\mathbb{P}_{S \sim \mathcal{D}^m}\left[L_{\mathcal{D}}\left(h_S\right) \leq \varepsilon\right] \geq 1-\delta$$

We begin with the following observation: the tightest-fit rectangle R' is always contained in the true rectangle R: $R' \subseteq R$. As such, the error of this algorithm can only come from positiev samples that fall in the area of $R \setminus R'$. We can express this area as the union of four rectangular strips (top, bottom, left, right) as follows: T' is the axis-aligned rectangle with the upper boundary of R' and lower boundary R (left and right boundaries are left and right boundaries of R. Similarly we can define the three other rectangular strips (with overlaps between these strips at the corners).



Notice that if we are able to guarantee that the weight under $\mathcal{D}$ of each strip (that is, that the probability with respect to $\mathcal{D}$ of falling in each strip) is at most $\varepsilon/4$, then the error of R' is at most ε. Thus, let us analyse the weight of T'.

Define T to be the rectangular strip along the inside top of R, weighing exactly $\varepsilon/4$ under $\mathcal{D}$. The rectangular strip T' has weight exceeding $\varepsilon/4$ if and only if $T' \subseteq T$. In addition, $T \subseteq T'$ if and only if there are no points of S in T. If there was such a point $p \in T$ that is contained in S then this point must be labeled positive as it is contained in R. Then, by the definition of the algorithm R' must extend upwards into T to include p.

So, by the definition of T having weight of $\varepsilon/4$, the probability of drawing a sample from $\mathcal{D}$ that misses T is exactly $1-\varepsilon/4$. The probability of drawing m independent samples from $\mathcal{D}$ all missing T is therefore $(1-\varepsilon/4)^m$. By the union bound, the probability of these points missing all four rectangular strips is at most $4(1-\varepsilon/4)^m$.

This means that if we choose m is such a way that it satisfies $4(1-\varepsilon/4)^m \leq \delta$, then with probability $1-\delta$ over the m random samples the weight of the error is at most ε. Using the inequality

$$1-x \leq e^{-x} \quad \forall x \in \mathbb{R}$$

we see that

$$4(1-\varepsilon/4)^m \leq 4e^{-\varepsilon m/4}$$

Therefore, if we choose m to satisfy $4e^{-\varepsilon m/4} \leq \delta$ it will also satisfy the previous condition. By solving for m:

$$\ln\left(e^{-\varepsilon m/4}\right) \leq \ln(\delta/4) \quad \iff \quad \frac{\varepsilon m}{4} \geq \ln(4/\delta) \quad \iff \quad m \geq \left\lceil \frac{4\ln(4/\delta)}{\varepsilon} \right\rceil$$

we find that if we sample at least $\left\lceil \frac{4\ln(4/\delta)}{\varepsilon} \right\rceil$ *i.i.d* samples from $\mathcal{D}$ then the tightest-fit algorithm will have an error of at most ε with probability at least $1-\delta$.

2 VC-Dimension

Many interesting hypothesis classes are not finite (e.g. halfspaces), but it turns out that in some cases we can still learn them efficiently. In fact, in the case of binary classification, the property of PAC learnability of a class is completely determined by its VC-dimension.

Definition 2.1 — Restriction. Let $\mathcal{H}$ be a hypothesis class for binary classification over domain $\mathcal{X}$. Let $C = \{\mathbf{x}_1, \dots, \mathbf{x}_m\} \subset \mathcal{X}$ be a set of some size $m \in \mathbb{N}$. The *restriction* of $\mathcal{H}$ to C is

$$\mathcal{H}_C := \{h_C = (h(\mathbf{x}_1), \dots, h(\mathbf{x}_m)) \,|\, h \in \mathcal{H})\}$$

We can think of $\mathcal{H}_C$ as a set of binary vectors: $\mathcal{H}_C \subseteq \{0,1\}^{|C|}$ where $\mathcal{H}_C$ defined as above. As such we notice that $|\mathcal{H}_C| \leq 2^{|C|}$.

Definition 2.2 — Shattering. A hypothesis class $\mathcal{H}$ is said to *shatter* a finite set $C \subset \mathcal{X}$ if and only if $|\mathcal{H}_C| = 2^{|C|}$. Namely, if the restriction of $\mathcal{H}$ to C enables the points in C to have any possible labeling.

Definition 2.3 — VC-Dimension. The VC-Dimension of a hypotheses class $\mathcal{H}$ is the size of the largest shattered set of points:

$$VC-Dim(\mathcal{H}) := sup\{m \in \mathbb{N} | \exists C \subset \mathcal{X} \ |C| = m\ s.t.\ \mathcal{H}\ shatters\ C\}$$

To show that $d \in \mathcal{N}$ is the VC-Dimension of a hypothesis class we need to show two things:

1. There *exists* a set $C \subseteq \mathcal{X}$ of size d which is shattered by $\mathcal{H}$. This implies that $VC-Dim(\mathcal{H}) \geq d$.
2. For *every* set $C \subseteq \mathcal{X}$ of size $d+1$ $\mathcal{H}$ does not shatter C. This also means that all sets of size $m > d+1$ are not shattered. Thus $VC-Dim(\mathcal{H}) \leq d$.

2.1 VC-Dimension Of Finite Classes

Exercise 2.1 Let $\mathcal{H}$ be some finite hypothesis class. Show that $VC-Dim(\mathcal{H}) \leq \log_2 |\mathcal{H}|$.
Solution: Let $VC-Dim(\mathcal{H}) = d$. By definition, it means that there exists a set of size d that is shattered by $\mathcal{H}$. As such, there are at least 2^d different hypothesis in $\mathcal{H}$, i.e. $2^d \leq |\mathcal{H}| \iff d \leq \log_2 |\mathcal{H}|$. ■

Exercise 2.2 Let $\mathcal{H}$ be some finite hypothesis class. Show that there can be an arbitrary gap between $VC-Dim(\mathcal{H})$ and $\log_2 |\mathcal{H}|$.

Solution: Consider the class of singletons, $\mathcal{H}_{singleton} := \{h_z : h_z(\mathbf{x}) = \mathbb{1}[\mathbf{x} = z], z \in \mathcal{X}\} \cup \{h^0\}$ where h^0 is defined by $h^0(x) \equiv 0\ \forall x \in \mathcal{X}$. Then $VC-Dim(\mathcal{H}) = 1$ but $|\mathcal{H}|$ can be arbitrarily large. ■

Exercise 2.3 Show there exists a finite hypothesis class $\mathcal{H}$ with $VC-Dim(\mathcal{H}) = \log_2 |\mathcal{H}|$.
Solution: Let $\mathcal{X} = \{\pm 1\}^d$ be the d-dimensional cube, and $\mathcal{H} = \mathcal{Y}^{\mathcal{X}}$, that is all the possible functions from $\mathcal{X}$ to $\mathcal{Y}$.

In this case, $|\mathcal{H}| = |\mathcal{Y}|^{|\mathcal{X}|} = 2^{2^d}$. And indeed, let C be all the possible different samples, then $|C| = 2^d$. From the definition of $\mathcal{H}$, for every labeling of C we can find an hypothesis $h \in \mathcal{H}$ that fits this labeling, and thus shatters this set. ■

2.2 VC-Dimension And Number Of Parameters

In many cases, such as seen in the singletons- and axis-aligned rectangle hypothesis classes, the VC-Dimension turns out to be the numbers of parameters needed in order to specify the hypotheses. Though this is common it is not always the case. Consider the following hypothesis class as is demonstrated by the following hypothesis class:

$$\mathcal{H} := \{x \mapsto sgn(sin(\theta \cdot x)) : \theta \in \mathbb{R}\}$$

Hypotheses in this class are parameterized via a single parameter θ. Even so, this hypothesis class can shatter an arbitrarily large set with arbitrary labeling. Therefore it has an infinite VC-Dimension and is not PAC learnable.

R We will need to know θ with ever growing precision if we are to shatter data sets with very dense points and alternating labels.

For examples already seen in the course, recall that the class of homogeneous linear separators over $\mathcal{X} = \mathbb{R}^d$ defined as

$$\mathcal{H} := \left\{ h_{\mathbf{w}} : h_{\mathbf{w}}(\mathbf{x}) = sign(\langle \mathbf{w}, \mathbf{x} \rangle), \mathbf{w} \in \mathbb{R}^d \right\}$$

Exercise 2.4 Let $\mathcal{H}$ be the class of homogeneous halfspaces in $\mathbb{R}^d$, then $VC - Dim(\mathcal{H}) = d$.

Solution: Let us show a set of size d that is shattered by $\mathcal{H}$ and that for any set of size $d+1$, $\mathcal{H}$ does not shatter it.

1. Consider the set of standard basis vectors, $C = \{e_1, \dots, e_d\}$. For any labeling $(y_1, \dots, y_d) \in \{\pm 1\}^d$ define $\mathbf{w} = (y_1, \dots, y_d)^\top$. Clearly, for all i it holds that $\langle \mathbf{w}, e_i \rangle = y_i$ and thus C is shattered. This means that $VC - Dim(\mathcal{H}) \geq d$.
2. Let $C = \{\mathbf{x}_1, \dots, \mathbf{x}_{d+1}\}$ be any set of $d+1$ points in $\mathbb{R}^d$. Since they must be linearly dependent, there is a non-trivial linear combination such that $\sum_{i=1}^{d+1} a_i \mathbf{x}_i = 0$. Assume w.l.o.g. that $a_{d+1} \neq 0$. In this case

$$\mathbf{x}_{d+1} = \sum_{i=1}^{d} b_i \mathbf{x}_i \quad b_i := -\frac{a_i}{a_{d+1}}$$

.

Let us show that it is impossible to achieve the labeling

$$y_i = \begin{cases} sign(b_i) & 1 \leq i \leq d \\ -1 & i = d+1 \end{cases}$$

Assume towards contradiction that there exists $\mathbf{w}$ that achieves such labeling, i.e.

$$y_i = sign(\langle \mathbf{w}, \mathbf{x}_i \rangle) = \begin{cases} sign(b_i) & 1 \leq i \leq d \\ -1 & i = d+1 \end{cases}$$

In such case we get that $\langle \mathbf{w}, \mathbf{x}_{d+1} \rangle = \left\langle \mathbf{w}, \sum_{i=1}^{d} b_i \mathbf{x}_i \right\rangle = \sum_{i=1}^{d} b_i \langle \mathbf{w}, \mathbf{x}_i \rangle > 0$ and so $y_{d+1} = sign(\langle \mathbf{w}, \mathbf{x}_{d+1} \rangle) = 1$ in contradiction to the assumption that $y_{d+1} = -1$.

■

Definition 2.4 Let $\mathbf{X} := (\mathbf{x}_1, \dots, \mathbf{x}_d) \subseteq \mathbb{R}^d$. A **monomial** is a function

$$\mathbf{x} \mapsto \prod_{i=1}^{d} x_i^{n_i}$$

for some set of $n_1, \dots, n_d \in \mathbb{N}$. The **degree of the monomial** is the sum $\sum_{i=1}^{d} n_i$.

We say that $p(\mathbf{x})$ is a **polynomial of degree** r if it is a linear sum of monomials of degree at most r.

■ **Example 2.1** Monomials of degree 2 are of the form $\left(x_1^2, x_1x_2, x_1x_3, \ldots, x_dx_{d-1}, x_d^2\right)$. So $p(\mathbf{x}) = x_1^2 - 7x_2x_5$ is a polynomial of degree 2. ■

Definition 2.5 Let $\mathcal{H}_{poly}^k$ be the set of all hypotheses that classify based on a polynomial of degree k, that is

$$\mathcal{H}_{poly}^k := \left\{\mathbf{x} \mapsto sign\left(p\left(\mathbf{x}\right)\right) | p\left(\mathbf{x}\right) : \mathbb{R}^d \to \mathbb{R} \text{ is a polynomial of degree } k\right\}$$

Exercise 2.5 Show that $VC - Dim\left(\mathcal{H}_{poly}^k\right) \leq \binom{d+k-1}{k}$.

Solution: Let $\mathbf{x} \in \mathbb{R}^d$ and define $\Psi_k(\mathbf{x})$ as the vector of all its monomials of degree k. For example

$$\Psi_2(x) = \begin{pmatrix} x_1^2 \\ x_1x_2 \\ \vdots \\ x_d^2 \end{pmatrix}$$

Notice that Ψ_k is a function whose image is in $\mathbb{R}^{\binom{d+k-1}{k}}$ as the number of monomials of degree k in a d dimensional space (with replacement and without relevance to order) is $\binom{d+k-1}{k}$.

Now, a polynomial of degree k over $\mathbf{x}$ is in fact some linear function over $\Psi_k(\mathbf{x})$. That is, there exists some $\mathbf{w} \in \mathbb{R}^{\binom{d+k-1}{k}}$ such that $p(\mathbf{x}) = \mathbf{w}^\top \Psi_k(\mathbf{x})$. If we look at the function $sign(p(\mathbf{x}))$ we are looking at homogeneous halfspaces over $\mathbb{R}^{\binom{d+k-1}{k}}$. Therefore, we can upper bound the VC-Dimension by $\binom{d+k-1}{k}$. ■