Assignment 1
Advanced Machine Learning

Armand Nicolicioiu (AI - 407)

April 2021

Exercise 1 Give an example of a finite hypothesis class $\mathcal{H}$ with $VC\dim(\mathcal{H}) = 2021$.

Solution:
Having $A = \{e1, e2, ..., e_n\}$ the orthonormal basis of $\mathbb{R}^n$, we define a finite hypothesis class $\mathcal{H}$ as follows:

$$\mathcal{H} = \{h_{w,0} \colon \mathbb{R}^n \to \{-1, 1\}, h_{w,0} = sign(\sum_{i=1}^{n} w_i x_i) \mid w_i = \begin{cases} 1, & \text{if } e_i \in B \\ -1, & \text{if } e_i \notin B \end{cases} \text{ for all subsets } B \text{ of } A\}.$$

We will show that $VC\dim(\mathcal{H}) = n$ and pick $n = 2021$ for our particular case.

Proof.

Part 1. We first show that $VC\dim(\mathcal{H}) \geq n$ by finding a set A of n points in $\mathbb{R}^n$ that is shattered by $\mathcal{H}$. As in the halfspaces example from *Lecture 7*, we take $A = \{e1, e2, ..., e_n\}$ to be the orthonormal basis of $\mathbb{R}^n$. Using the alternative definition of shattering from *Lecture 6*, we need to show that for every subset B of A there is a function h_B that labels $+1$ all elements in B and -1 all elements in $A \setminus B$.

By construction, our $\mathcal{H}$ contains exactly one corresponding hypothesis for each subset B in A:

$$h_B = h_{w,0} = sign(\sum_{i=1}^{n} w_i x_i) \text{ where } w_i = \begin{cases} 1, & \text{if } e_i \in B \\ -1, & \text{if } e_i \notin B \end{cases}.$$

We have $h_B = sign(\langle w, e_i \rangle) = w_i$, so it will assign $+1$ for all elements in B and -1 for all elements in $B \setminus A$, therefore proving that $VC\dim(\mathcal{H}) \geq n$.

Part 2. Now we will show that $VC\dim(\mathcal{H}) < n + 1$. We use the property presented in *Lecture 6* that for a finite hypothesis class $\mathcal{H}$ we have the upper bound:

$$VC\dim(\mathcal{H}) \leq \lfloor \log_2(|\mathcal{H}|) \rfloor$$

For our choice of $\mathcal{H}$ we have exactly one hypothesis for each subset of A, resulting in $|\mathcal{H}| = 2^{|A|} = 2^n$. Therefore:

$$VC\dim(\mathcal{H}) \leq \lfloor \log_2(2^n) \rfloor = n$$

Using both parts results in $VC\dim(\mathcal{H}) = n$. □

Exercise 2 Give an example of a finite set A in $\mathbb{R}^n$ of size 4 that is shattered by $\mathcal{H}_{balls}$ or justify why you cannot find such set.

Solution:
We will show that there is not possible to find an example of a set A in $\mathbb{R}^2$ of size 4 that is shattered by $\mathcal{H}_{balls}$.

Let's group the positioning of the 4 points into two generic situations:

Case 1) The 4 points form a concave polygon. In this case, there is no way to assign label 1(+) to the points that form the convex hull without assigning 1(+) to the point in the middle as well. Therefore, no set in this configuration can be shattered by $\mathcal{H}_{balls}$.

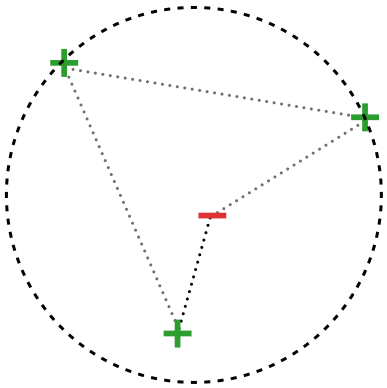


Figure 1: Concave polygon with negative label only for the point inside the convex hull.

Case 2) The 4 points form a convex polygon. It is easy to show that we can group together any number of adjacent points (label them with $1(+)$ by covering with a circle) and exclude the others (leave them outside that circle, thus labeling with $0(-)$).

The problem appears when we want to label as $1(+)$ two opposite points and as $0(-)$ the other two. If it is possible to label as $1(+)$ the first opposing points (let's say, without restraining generality, A and C) without covering the other two, then, if we want the label B and D as $1(+)$, we will not be able to do this without covering at least one of the remaining points, A or C as well. Therefore, the cases in which (A, B, CD) are labeled $(1, 0, 1, 0)$ or, equivalently $(0, 1, 0, 1)$ cannot be both realized by hypotheses in $\mathcal{H}_{balls}$.

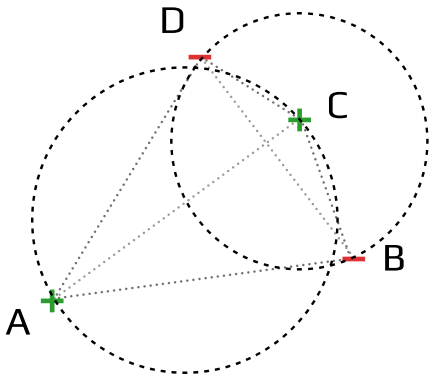


Figure 2: Convex polygon with labels $(1, 0, 1, 0)$ or $(0, 1, 0, 1)$.

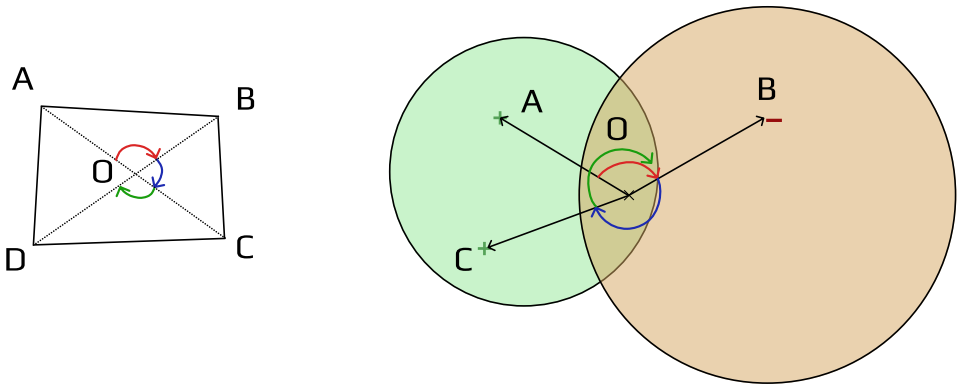


Figure 3: Assignment of points in clock-wise order

We prove this formally in the following. We are in the case of convex polygons ABCD. Thus, there exists O the point of intersection of AB and CD (see Figure 3).

Let $l_1 = (1, 0, 1, 0)$ and $l_2 = (0, 1, 0, 1)$ be two labelings for some points $\{A, B, C, D\}$. We assume that these labelings can be realised, meaning we assume that there is a ball B_1 that contains only A and C (realising l_1) and a ball B_2 that contains only B and D (realising l_2). As the intersection of AC and BD, the point O must be contained by both balls.

Let's select the points $\{A, B, C, D\}$ by rotating clock-wise a ray centered in O. We start from A that must be in the first ball B_1. We rotate clock-wise and select B from the second ball B_2. We continue and select C from the ball B_1. As D has different label than C, it should be in the second ball B_2, and the last rotation will overpass the full circle as it passes over 360°. This is a contradiction.

Thus, the labelings $l1 = (1, 0, 1, 0)$ and $l2 = (0, 1, 0, 1)$ cannot be both realized by any hypotheses h_1 (ball B_1) and h_2 (ball B_2) in $\mathcal{H}_{balls}$.

In conclusion, no set A of size 4 exists such that it is shattered by $\mathcal{H}_{balls}$.

Exercise 3 Show that the class $\mathcal{H}_\alpha$ can be (ϵ, δ) - PAC learned by giving an algorithm A and determining an upper bound on the sample complexity $m_{\mathcal{H}}(\epsilon, \delta)$ such that the definition of PAC-learnability is satisfied.

Solution:

The hypothesis $\mathcal{H} = \mathcal{H}_\alpha$ is (ϵ, δ) - PAC learnable if for every $\epsilon, \delta > 0$, for every labeling function $f \in \mathcal{H}^2_{rec}$ (realizability case) and for every distribution $\mathcal{D}$ on $\mathbb{R}^2$ we can find a function $m_{\mathcal{H}} \colon (0, 1)^2 \to \mathbb{N}$ and a learning algorithm A such that, when ran on a training set S consisting of $m \geq m_{\mathcal{H}}(\epsilon, \delta)$ examples sampled i.i.d. from $\mathcal{D}$ and labeled by f, the algorithm A returns a hypothesis $h_S \in \mathcal{H}$ that has a real risk lower than ϵ with probability at least $1 - \delta$.

$$\begin{aligned} &\underset{S \sim \mathcal{D}^m}{P}(L_{f,\mathcal{D}}(h_S) \leq \epsilon) \geq 1 - \delta \\ \implies &\underset{S \sim \mathcal{D}^m}{P}(L_{f,\mathcal{D}}(h_S) > \epsilon) < \delta \end{aligned}$$

We will first define the algorithm A.

We denote the hypotheses represented by a right triangle $\triangle ABC$ with catheti AB and AC parallel to the axes (Ox and Oy) and $AB/AC = \alpha$ as $h_{a1,b1,a2,b2}$, where $a1, b1, a2, b2$ are the coordinates of the triangle's vertices as follows:

$$A = (a1, b1)$$
$$B = (a2, b1)$$
$$C = (a1, b2)$$

We are under the realizability assumption, so there exists a labeling function $f \in \mathcal{H}_\alpha, f = h^*_{(a_1^*, b_1^*, a_2^*, b_2^*)}$ that labels the training data.

Consider the training set $S = \left\{ (x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m) \;\middle|\; \begin{array}{l} y_i = h^*_{(a_1^*, b_1^*, a_2^*, b_2^*)}(x_i), \\ x_i \in \mathbb{R}^2, x_i = (x_{i1}, x_{i2}) \end{array} \right\}$

Let define the algorithm A that learns from the training set S as the tightest right triangle in $\mathcal{H}_\alpha$ that covers all the points in S, assigning $1(+)$ to all points inside it and $0(-)$ to all points outside. As in Figure 4, we pick $a1$ as the leftmost Ox coordinate of a point in S and $b1$ as the lowermost Oy coordinate of a point in S. Now that we found point A, we need to determine the other ends of the segments AB and AC — the points B and C. To do that, we draw a line of slant $-ctg^-1(\alpha)$ through each positive example x_i in S until it intersects the catethi in points B_i and C_i, respectively. To assure that all examples in the training set are labeled correctly by our algorithm, we pick the point $P(x_{P1}, x_{P2})$ that has the longest distance to the imaginary line drawn through A and parallel with BC. As a result, we have:

$$a_1 = \min(x_{i1})$$
$$b_1 = \min(x_{i2})$$
$$b_2 = b_{2i} = \arg\max_{x_i \in S} \{x_{i2} + \alpha(x_{i1} - a_1)\}$$
$$a_2 = a_{2i} = \arg\max_{x_i \in S} \{\frac{b_1 - x_{i2}}{\alpha} - x_{i1}\}$$

If there are no positive examples in the training set, we choose a new point $z = (z_1, z_2)$ outside the training set and pick $A = B = C = z$.

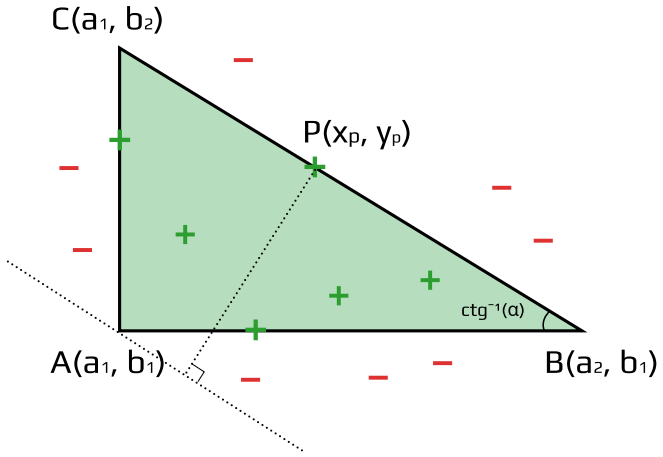


Figure 4: Algorithm A represented by the tightest right triangle in $\mathcal{H}_\alpha$ that covers all the points in S labeled with $1(+)$.

We defined an algorithm that generates $h_S = h_{a1,b1,a2,b2}$ as the tightest right triangle in $\mathcal{H}_\alpha$ enclosing all the positives examples in S. By construction, A is an ERM algorithm with $L_{h^*,\mathcal{D}}(h_S) = 0$ because h_S doesn't mislable any example in the training set.

We denote by T^* the triangle region determined by hypothesis h^* given by the realizability assumption. All the points inside T^* have $1(+)$ as the correct label, and all the points outside it have $0(-)$ as the correct label.

We denote by T_S the triangle region determined by hypothesis h_S given by our learning algorithm A trained on the set S.

Now we want to find the sample complexity $m_{\mathcal{H}}(\epsilon, \delta)$ such that

$$\underset{S\sim\mathcal{D}^m}{P}(L_{h^*,\mathcal{D}}(h_S) \leq \epsilon) \geq 1-\delta \text{ where } S \text{ contains } m \geq m_{\mathcal{H}}(\epsilon,\delta) \text{ examples.}$$

Because of the realizability assumption and our of choice of A, it is assured that h_S will label correctly any point inside T_S. Also, we make the observation that $T_S \subset T^*$, meaning that h_S also labels correctly all the examples outside T^* because they are outside T_S too. Thus, the only region where h_S can make mistakes is $T^* \setminus T_S$.

Let's fix $\epsilon > 0, \delta > 0$ and consider a distribution $\mathcal{D}$ over $\mathbb{R}^2$.

Case 1)

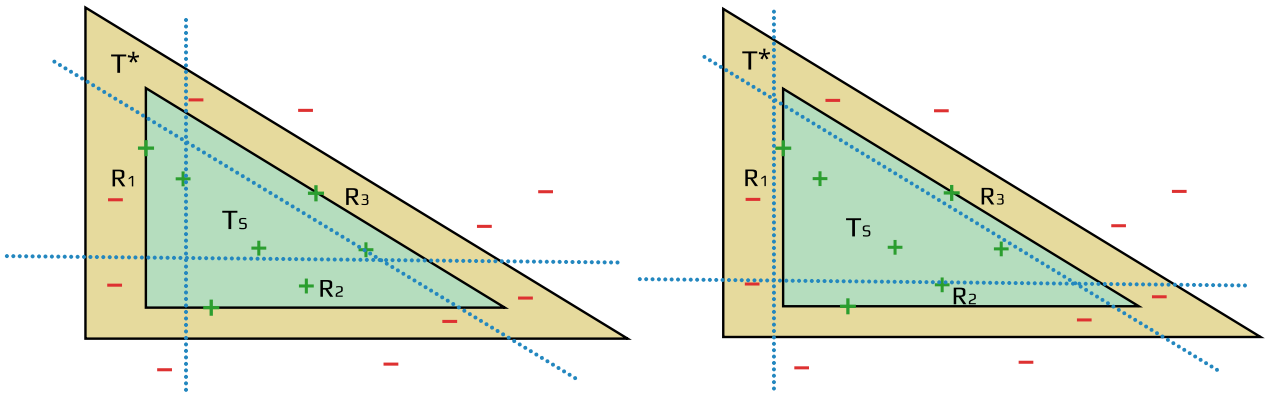
$$\begin{aligned} &\mathcal{D}(T^*) = \underset{x\sim\mathcal{D}}{P}(x \in T^*) \leq \epsilon \\ \implies &L_{h^*,\mathcal{D}}(h_S) = \underset{x\sim\mathcal{D}}{P}(h_S(x) \neq h^*(x)) = \underset{x\sim\mathcal{D}}{P}(x \in T^* \setminus T_S) \leq \underset{x\sim\mathcal{D}}{P}(x \in T^*) \leq \epsilon \\ \implies &\underset{S\sim\mathcal{D}^m}{P}(L_{h^*,\mathcal{D}}(h_S) \leq \epsilon) = 1 \text{ (this happens all the time)} \end{aligned}$$

Case 2) $\mathcal{D}(T^*) = \underset{x\sim\mathcal{D}}{P}(x \in T^*) > \epsilon$

We construct as in Figure 5 the regions $R1, R2, R3$ by drawing lines parallel to the triangle's sides, having $\mathcal{D}(R_i) = \underset{x\sim\mathcal{D}}{P}(x \in R_i) = \frac{\epsilon}{3}$.

Case 2.1) T_S intersects all regions $R1, R2, R3$ (See Figure 5a). This results in:

$$\begin{aligned} L_{h^*,\mathcal{D}}(h_S) = \underset{x\sim\mathcal{D}}{P}(h^*(x) \neq h_S(x)) = \underset{x\sim\mathcal{D}}{P}(x \in T^* \setminus T_S) \leq \underset{x\sim\mathcal{D}}{P}(x \in R_1 \cup R_2 \cup R_3) \leq \\ \leq \sum_{i=1}^{3} \underset{x\sim\mathcal{D}}{P}(x \in R_i) = \sum_{i=1}^{3} \mathcal{D}(R_i) = 3 \cdot \frac{\epsilon}{4} = \epsilon \quad \text{(with probability 1)} \end{aligned}$$



(a) T^* overlaps with all regions $R1, R2, R3$. (b) T^* does not overlap with at least one region R_i

Figure 5: The regions defined by the hypotheses: The yellow triangle is T^*, the green triangle is T_S and $R1, R2, R3$ are the trapezoids separated by the dotted lines. The examples in $T^* \setminus T_S$ area will be misclassified by our h_S as negative when in fact they should be positive because they are located inside T^*.

Case 2.2) T_S will not intersect at least one region of $R1, R2, R3$ (See Figure 5b). This allows for the possibility to have $L_{h^*,\mathcal{D}}(h_S) > \epsilon$.

We denote with F_i this event of T_S not intersecting R_i, so we have $F_i = \{S \sim \mathcal{D}^m \mid T_S \cap R_i = \emptyset\}$.

$$\underset{S\sim\mathcal{D}^m}{P}(L_{h^*,\mathcal{D}}(h_S) > \epsilon) \leq \underset{S\sim\mathcal{D}^m}{P}(\overset{\text{at least one } F_i \text{ happens}}{\nearrow} F1 \cup F2 \cup F3) \leq \sum_{i=1}^{3} \underset{S\sim\mathcal{D}^m}{P}(F_i)$$

Now, $\underset{S\sim\mathcal{D}^m}{P}(F_i) =$ the probability that T_S will not intersect R_i

$$= \text{the probability that no point from } R_i \text{ is sampled in } S$$
$$= \left(1 - \frac{\epsilon}{3}\right)^m$$

So

$$\underset{S\sim\mathcal{D}^m}{P}(L_{h^*,\mathcal{D}}(h_S) > \epsilon) \leq \sum_{i=1}^{3} \underset{S\sim\mathcal{D}^m}{P}(F_i) = 3 \cdot \left(1 - \frac{\epsilon}{3}\right)^m$$

From *Lecture 2* we know that $1 - x \leq e^{-x}$, so $1 - \frac{\epsilon}{3} \leq e^{-\frac{\epsilon}{3}}$, meaning that

$$\underset{S\sim\mathcal{D}^m}{P}(L_{h^*,\mathcal{D}}(h_S) > \epsilon) \leq 3 \cdot \left(1 - \frac{\epsilon}{3}\right)^m \leq 3 \cdot e^{-\frac{\epsilon}{3}m}$$

We want $\underset{S\sim\mathcal{D}^m}{P}(L_{h^*,\mathcal{D}}(h_S) > \epsilon) < \delta$.

$$\implies 3 \cdot e^{-\frac{\epsilon}{3}m} < \delta$$
$$\implies e^{-\frac{\epsilon}{3}m} < \frac{\delta}{3}$$
$$\implies -\frac{\epsilon}{3} \cdot m < \log\frac{\delta}{3}$$
$$m > -\frac{3}{\epsilon}\log\frac{\delta}{3} = \frac{3}{\epsilon}\log\frac{3}{\delta}$$

Therefore, for a training set S of size $m \geq m_{\mathcal{H}}(\epsilon, \delta) = \frac{3}{\epsilon} \cdot \log\frac{3}{\delta}$ i.i.d. samples from $\mathcal{D}$, our learning algorithm A obtains a hypothesis h_S with $\underset{S\sim\mathcal{D}^m}{P}(L_{h^*,\mathcal{D}}(h_S) \leq \epsilon) \geq 1 - \delta$, showing that $\mathcal{H}_\alpha$ is (ϵ, δ) - PAC learnable. □

Exercise 4 Show that the class $\mathcal{H}$ can be (ϵ, δ) - PAC learned by giving an algorithm A and determining an upper bound on the sample complexity $m_{\mathcal{H}}(\epsilon, \delta)$ such that the definition of PAC-learnability is satisfied. Compute VCdim($\mathcal{H}$).

Solution:
Similar to the previous problem, we start by defining an algorithm A.
We consider the training set $S = \{(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)\}$.
We find the point x_S that is furthest away from the origin, from the set S_+ of points with label 1.

$$r_S = \max_{x_i \in S_+} ||x_i||_2 \tag{1}$$

We define the hyphotheis $h_S = B(r_S) = B_S$ the ball of radius equal to the distance from the origin to the point x_S. By construction, A is an ERM algorithm with $L_{h^*,\mathcal{D}}(h_S) = 0$ because h_S doesn't mislable any

example in the training set.

From the realizability assumption, there exists a hypothesis $h^* = B(r^*) = B^*$ such that all the points inside the ball $B(r^*)$ have $1(+)$ as the correct label, and all the points outside it have $0(-)$ as the correct label.

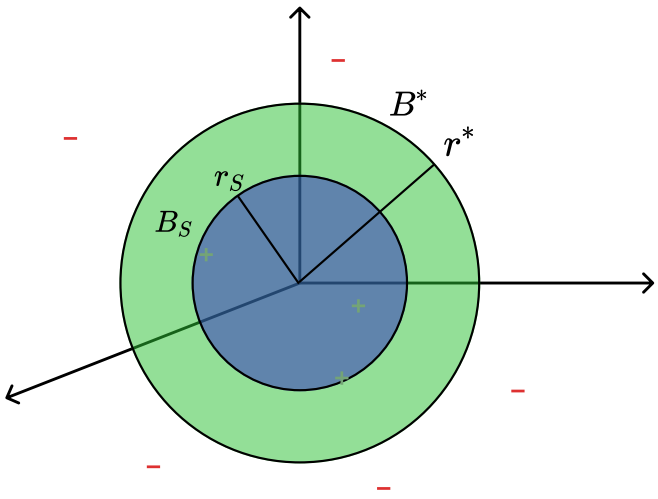


Figure 6

Now we want to find the sample complexity $m_{\mathcal{H}}(\epsilon, \delta)$ such that

$$\underset{S\sim\mathcal{D}^m}{P}(L_{h^*,\mathcal{D}}(h_S) \leq \epsilon) \geq 1 - \delta \text{ where } S \text{ contains } m \geq m_{\mathcal{H}}(\epsilon, \delta) \text{ examples.}$$

Because of the realizability assumption and our of choice of A, it is assured that h_S will label correctly any point inside B_S. Also, we make the observation that $B_S \subset B^*$, meaning that h_S also labels correctly all the examples outside B^* because they are outside B_S too. Thus, the only region of space where h_S can make mistakes is $B^* \setminus B_S$.

Case 1)

$$\begin{aligned} &\mathcal{D}(B^*) = \underset{x\sim\mathcal{D}}{P}(x \in B^*) \leq \epsilon \\ \implies &L_{h^*,\mathcal{D}}(h_S) = \underset{x\sim\mathcal{D}}{P}(h_S(x) \neq h^*(x)) = \underset{x\sim\mathcal{D}}{P}(x \in B^* \setminus B_S) \leq \underset{x\sim\mathcal{D}}{P}(x \in B^*) \leq \epsilon \\ \implies &\underset{S\sim\mathcal{D}^m}{P}(L_{h^*,\mathcal{D}}(h_S) \leq \epsilon) = 1 \text{ (this happens all the time)} \end{aligned}$$

Case 2)

$$\mathcal{D}(B^*) = \underset{x\sim\mathcal{D}}{P}(x \in B^*) > \epsilon \tag{2}$$

We define $T_\epsilon = B^* \setminus B_\epsilon = B(r^*) \setminus B(r_\epsilon)$ is a region of space close to the margins of the B^* such that $\underset{x\sim\mathcal{D}}{P}(x \in T_\epsilon) = \epsilon$ (see Figure 7)

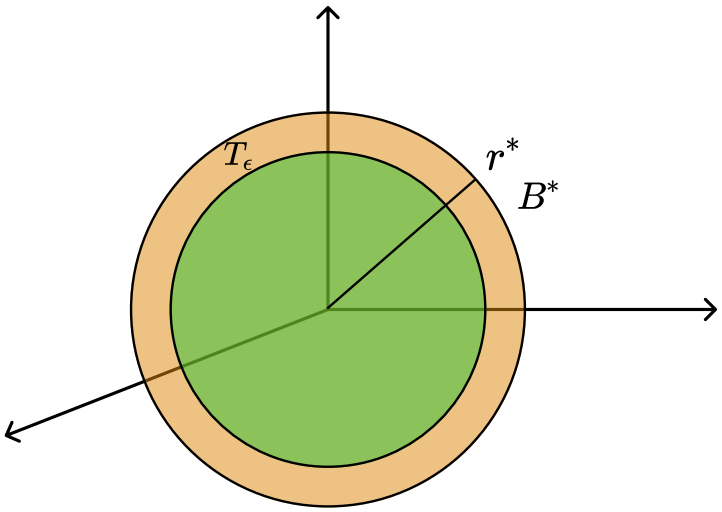


Figure 7: T_ϵ (orange color) represents the region at the margin of the B^* ball.

Case 2.1) B_S intersects T_ϵ, results:

$$L_{h^*,\mathcal{D}}(h_S) = \underset{x\sim\mathcal{D}}{P}(h^*(x) \neq h_S(x)) = \underset{x\sim\mathcal{D}}{P}(x \in B^* \setminus B_S) \leq \underset{x\sim\mathcal{D}}{P}(x \in T_\epsilon) = \epsilon \quad \text{(with probability 1)}$$

Case 2.2) $B_S \cap T_\epsilon = \emptyset$

$$\underset{S\sim\mathcal{D}^m}{P}(L_{h^*,\mathcal{D}}(h_S) > \epsilon) = \underset{S\sim\mathcal{D}^m}{P}(B_S \cap T_\epsilon = \emptyset)$$

$\underset{S\sim\mathcal{D}^m}{P}(B_S \cap T_\epsilon = \emptyset)$ is the probability that no point from T_ϵ is sampled in the training set S. This probability is $(1-\epsilon)^m$. We obtain:

$$\underset{S\sim\mathcal{D}^m}{P}(L_{h^*,\mathcal{D}}(h_S) > \epsilon) = \underset{S\sim\mathcal{D}^m}{P}(B_S \cap T_\epsilon = \emptyset) = (1-\epsilon)^m \leq e^{-\epsilon m}$$

$$\begin{aligned} & e^{-\epsilon m} < \delta \\ \implies & -\epsilon m < \log\delta \\ \implies & m > -\frac{1}{\epsilon}\log\delta \\ \implies & m > \frac{1}{\epsilon}\log\frac{1}{\delta} \end{aligned}$$

Therefore, for a training set S of size $m \geq m_{\mathcal{H}}(\epsilon,\delta) = \frac{1}{\epsilon}\log\frac{1}{\delta}$ i.i.d. samples from $\mathcal{D}$, our learning algorithm A obtains a hypothesis h_S with $\underset{S\sim\mathcal{D}^m}{P}(L_{h^*,\mathcal{D}}(h_S) \leq \epsilon) \geq 1-\delta$, showing that $\mathcal{H}_\alpha$ is (ϵ,δ) - PAC learnable. □

b. VCdim($\mathcal{H}$)

It is easy to see that any set of one point $C = \{x\}$ is shattered by $\mathcal{H}$. We denote $d = ||x||_2$ the distance from the point x to the origin. The two possible labelings can be realised by selecting a ball with radius $r < d$ for label $0(-)$ or selecting a ball with radius $r > d$ for label $1(+)$.

So:

$$\text{VCdim}(gH) \geq 1 \tag{3}$$

Let $C_2 = \{x_1, x_2\}$ a set of two points. We will show that $\mathcal{H}$ cannot shatter C_2.

Without loosing generality, we can sort any two points such that the first one is closer to the origin than the second point (or at equal distance) $||x_1||_2 \leq ||x_2||_2$.

Let's assume that the labeling (0,1) is realised by a hyphothesis $h = B(r)$. The second point x_2 has label 1(+), so:

$$\begin{aligned} x_2 \in B(r) &\implies ||x_2|| < r \\ ||x_1||_2 \leq ||x_2||_2 < r &\implies x_1 \in B(r) \end{aligned}$$

Thus, the first point x_1 also has label 1(+), contradicting the assumption. Thus, the labeling (0,1) cannot be realised for any set of two points, which means that any such set C_2 cannot be shattered.

This results in

$$\begin{aligned} &\text{VCdim}(gH) < 2 \\ Eq.3 \implies &\text{VCdim}(gH) = 1 \end{aligned}$$

Exercise 5 Let $\mathcal{H} = \{h_\theta : \mathbb{R} \to \{0,1\}, h_\theta = \mathbb{1}_{[\theta,\theta+1]\cup[\theta+2,+\infty)}(x), \theta \in \mathbb{R}\}$. Compute VCdim($\mathcal{H}$).

Solution:

We first show that there exists a set $C = \{a, b, c\}$ of size 3 that is shattered by $\mathcal{H}$, thus VCdim($\mathcal{H}$) ≥ 3.

For a = 0, b = 0.3, c = 1.2, we show that we could choose different values of θ such that the resulting hypothesis realises all possible labelings.

For $\theta = 2$ we obtain the labeling $(0, 0, 0)$:

$$\begin{aligned} a = 0.0 < 2 = \theta &\implies h_\theta(a) = 0 \\ b = 0.6 < 2 = \theta &\implies h_\theta(b) = 0 \\ b = 1.2 < 2 = \theta &\implies h_\theta(c) = 0 \end{aligned}$$

For $\theta = -0.5$ we obtain the labeling $(1, 0, 0)$:

$$\begin{aligned} a = 0.0 \in [\theta, \theta+1] = [-0.5, 0.5] &\implies h_\theta(a) = 1 \\ b = 0.6 \in (\theta+1, \theta+2) = (0.5, 1.5) &\implies h_\theta(b) = 0 \\ b = 1.2 \in (\theta+1, \theta+2) = (0.5, 1.5) &\implies h_\theta(c) = 0 \end{aligned}$$

For $\theta = 0.1$ we obtain the labeling $(0, 1, 0)$:

$$\begin{aligned} a = 0.0 < \theta = 0.1 &\implies h_\theta(a) = 0 \\ b = 0.6 \in [\theta, \theta+1] = [0.1, 1.1] &\implies h_\theta(b) = 1 \\ b = 1.2 \in (\theta+1, \theta+2) = (1.1, 2.1) &\implies h_\theta(c) = 0 \end{aligned}$$

For $\theta = 1.0$ we obtain the labeling $(0, 0, 1)$:

$$\begin{aligned} a = 0.0 < \theta = 1.0 &\implies h_\theta(a) = 0 \\ b = 0.6 < \theta = 1.0 &\implies h_\theta(b) = 0 \\ c = 1.2 \in [\theta, \theta+1] = [1.0, 2.0] &\implies h_\theta(c) = 1 \end{aligned}$$

For $\theta = 0.5$ we obtain the labeling $(0, 1, 1)$:

$$\begin{aligned} a = 0.0 < \theta = 0.5 &\implies h_\theta(a) = 0 \\ b = 0.6 \in [\theta, \theta+1] = [0.5, 1.5] &\implies h_\theta(b) = 1 \\ c = 1.2 \in [\theta, \theta+1] = [0.5, 1.5] &\implies h_\theta(c) = 1 \end{aligned}$$

For $\theta = -0.1$ we obtain the labeling $(1, 1, 0)$:

$$\begin{aligned} a = 0.0 \in [\theta, \theta + 1] = [0.1, 0.9] &\implies h_\theta(a) = 1 \\ b = 0.6 \in [\theta, \theta + 1] = [0.1, 0.9] &\implies h_\theta(b) = 1 \\ c = 1.2 \in (\theta + 1, \theta + 2) = (0.9, 1.9) &\implies h_\theta(c) = 0 \end{aligned}$$

For $\theta = -0.9$ we obtain the labeling $(1, 0, 1)$:

$$\begin{aligned} a = 0.0 \in [\theta, \theta + 1] = [-0.9, 0.1] &\implies h_\theta(a) = 1 \\ b = 0.6 \in (\theta + 1, \theta + 2) = (0.1, 0.9) &\implies h_\theta(b) = 0 \\ c = 1.2 \in [\theta + 2, \infty) = [1.1, \infty) &\implies h_\theta(c) = 1 \end{aligned}$$

For $\theta = -2$ we obtain the labeling $(1, 1, 1)$:

$$\begin{aligned} a = 0.0 \in [\theta + 2, \infty] = [0.0, \infty) &\implies h_\theta(a) = 1 \\ b = 0.6 \in [\theta + 2, \infty] = [0.0, \infty) &\implies h_\theta(b) = 1 \\ c = 1.2 \in [\theta + 2, \infty] = [0.0, \infty) &\implies h_\theta(c) = 1 \end{aligned}$$

Thus, $C = \{0, 0.6, 1.2\}$ is shattered by $\mathcal{H}$, obtaining:

$$\text{VCdim}(\mathcal{H}) \geq 3. \tag{4}$$

Let's consider any set $C = \{a, b, c, d\}$ of size 4 with $a \leq b \leq c \leq d$.

If $\mathcal{H}$ shatters $C = \{a, b, c, d\}$ then there should exists θ such that h_θ realizes the labeling $(1, 0, 1, 0)$.

From the definition of h_θ if $h_\theta(a) = 1, h_\theta(b) = 0, h_\theta(c) = 1$ then $a \in [\theta, \theta+1], b \in (\theta+1, \theta+2), c \in [\theta+2, \infty]$. From $c \leq d \implies d \in [\theta + 2, \infty] \implies h_\theta(d) = 1$.

Thus, the labeling $(1, 0, 1, 0)$ cannot be realised and

$$\text{VCdim}(\mathcal{H}) < 4. \tag{5}$$

From Eq.4 and Eq.5 it result that

$$\text{VCdim}(\mathcal{H}) = 3. \tag{6}$$

Exercise 6 Let $\mathcal{X}$ be an instance space and consider $\mathcal{H} \subseteq \{0, 1\}^{\mathcal{X}}$ a hypothesis space with finite VC dimension. For each $x \in \mathcal{X}$ we consider the function $z_x : \mathcal{X} \to \{0, 1\}$ such that $z_x(h) = h(x)$ for each $h \in \mathcal{H}$. Let $Z = \{z_x : \mathcal{H} \to \{0, 1\}, x \in \mathcal{X}\}$. Prove that $\text{VCdim}(Z) < 2^{\text{VCdim}(\mathcal{H})+1}$.

Solution:

Let's denote $n = \text{VCdim}(\mathcal{H})$ and $M = 2^{n+1}$.

Let's assume $\text{VCdim}(Z) \geq 2^{\text{VCdim}(\mathcal{H})+1} = M$.

Thus, there exits a set $C = \{h_1, h_2, ..., h_M\}$ with elements $h_i \in \mathcal{H}$ from the domain of the functions $z \in Z$, $|C| = M$ that is shattered by Z.

Thus, for each labelings $l = [l(1), l(2), ..., l(M)] \in \{0, 1\}^M$ there exists a function z_x that realises it:

$$z_x(h_1) = l(1), \quad z_x(h_2) = l(2), \quad .... \quad z_x(h_M) = l(M)$$

or equivalently, for functions $C = \{h_1, h_2, ..., h_M\}$, $\forall$ labeling $l \in \{0, 1\}^M = l \in \{0, 1\}^{2^{n+1}}$ with 2^{n+1} elements, there exists a point x such that

$$h_1(x) = l(1), \quad h_2(x) = l(2), \quad .... \quad h_M(x) = l(M). \tag{7}$$

Let's consider the set of all labelings of $n + 1$ elements $Q = \{q_i | \forall i \leq M = 2^{n+1}\}$.

All labelings $q_i = [q_i(1), q_i(2), ...q_i(n+1)]$ could be organized into a matrix Q where each labeling reprezents a column

$$Q = \begin{bmatrix} | & | & & | \\ q_1^T & q_2^T & \dots & q_M^T \\ | & | & & | \end{bmatrix} \in \{0,1\}^{n \times M}$$

Each row i of the matrix Q could be seen as a labeling $l_i = [l_i(1), l_i(2), ..., l_i(M)] \in \{0,1\}^M$.

$$Q = \begin{bmatrix} | & | & & | \\ q_1^T & q_2^T & \dots & q_M^T \\ | & | & & | \end{bmatrix} = \begin{bmatrix} — & l_1 & — \\ — & l_2 & — \\ & \vdots & \\ — & l_{n+1} & — \end{bmatrix}$$

From Eq. 7, $\forall$ such labeling l_i, there $\exists x_i$ such that:

$$h_1(x_i) = l_i(1), \quad h_2(x_i) = l_i(2), \quad .... \quad h_M(x_i) = l_i(M) \tag{8}$$

Expanding the previous equation for all i, we obtain

$$\begin{gathered} h_1(x_1) = l_1(1), \quad h_2(x_1) = l_1(2), \quad .... \quad h_M(x_1) = l_1(M) \\ h_1(x_2) = l_2(1), \quad h_2(x_2) = l_2(2), \quad .... \quad h_M(x_2) = l_2(M) \\ \vdots \\ h_1(x_{n+1}) = l_{n+1}(1), \quad h_2(x_{n+1}) = l_{n+1}(2), \quad .... \quad h_M(x_{n+1}) = l_{n+1}(M) \end{gathered}$$

Rewriting the previous equations, we obtain

$$\begin{gathered} h_1(x_1) = l_1(1), \quad h_1(x_2) = l_2(1), \quad .... \quad h_1(x_{n+1}) = l_{n+1}(1) \\ h_2(x_1) = l_1(2), \quad h_2(x_2) = l_2(2), \quad .... \quad h_2(x_{n+1}) = l_{n+1}(2) \\ \vdots \\ h_M(x_1) = l_1(M), \quad h_M(x_2) = l_2(M), \quad .... \quad h_M(x_{n+1}) = l_{n+1}(M) \end{gathered}$$

As we defined it, $q_i = [l_1(i), l_2(i)], \dots l_M(i)$ is a column of Q, meaning one of the possible $M = 2^{n+1}$ labelings of size $n+1$. Thus, we have shown that there exists a set of points $C_2 = \{x_1, x_2, ..., x_{n+1}\}$ and for any labeling q_i there exists a hypothesis $h_i \in \{h_1, h_2, ..., h_M\}$ such that:

$$h_i(x_1) = q_i(1), \quad h_i(x_2) = q_i(2), \quad .... \quad h_i(x_{n+1}) = q_i(n+1) \tag{9}$$

Thus, $C_2 = \{x_1, x_2, ...x_{n+1}\}$ is shattered by $\mathcal{H}$ resulting in $\text{VCdim}(\mathcal{H}) \geq n+1 = \text{VCdim}(\mathcal{H}) + 1$.
This results in a contradictions, therefore $\text{VCdim}(Z) < 2^{\text{VCdim}(\mathcal{H})+1}$.