Advanced Machine Learning

Assignment 1 – April 2023

Tomeh Lara - 507

Question – 1

Give an example for

a) *A finite hypothesis class $\mathcal{H}$ with VCdim ($\mathcal{H}$) = 2023. Justify your choice.*
b) *An infinite hypothesis class $\mathcal{H}$ with VCdim ($\mathcal{H}$) = 2023. Justify your choice.*
c) *An infinite hypothesis class $\mathcal{H}$ with VCdim ($\mathcal{H}$) = ∞. Justify your choice.*

Answer – 1 – a)

Let $I \subseteq \{1,2,\dots.2023\}, B^{2023} = \{(x_1, x_2, \dots\ x_{2023}) \mid x_i \in \{0,1\}, \forall 1 \ \leq i \ \leq 2023\}\ and\ \mathcal{H} = \{h_I : B^{2023} \longrightarrow \{0,1\} \mid h_I(x) = \prod_{i \in I} x_i\}$. We consider that $h_\phi = 1$ for any x.

Claim: $\mathcal{H}$ is a finite hypothesis class with VCdim(H) = 2023.

Proof. $|\mathcal{H}|$ is given by the number of possible values of I. Any value of I can be thought of as a 2023-dimensional binary vector, where $x_i = 1$ if $i \in I$ and 0 otherwise. There are 2^{2023} such vectors, so $|\mathcal{H}| = 2^{2023}$. (finite)

For any finite hypothesis class $\mathcal{H}', VCdim(\mathcal{H}') \leq \ log_2|\mathcal{H}|$. For our $\mathcal{H}$, VCdim $(\mathcal{H}) \leq log_2\, 2^{2023} = 2023$. (1)

Let $C = \{\, \overline{e_i} \in \ B^{2023}|$ all bits of $\overline{e_i}$ are 1, except for the one at the i^{th} position, $1 \leq i \leq 2023\}$.

We want to show that $\mathcal{H}$ shatters C.

Let $C_I \subseteq C, C_I = \{\overline{e_i} \in C \mid i \in I\}$, and $\bar{I} = [2023] \setminus I$. So $C_{\bar{I}} = C \setminus C_I$.

Proving that for any subset of C, namely $C_{\bar{I}}$, there is a hypothesis in $\mathcal{H}$ that labels all elements of $C_{\bar{I}}$ with 1 is the same as proving that there is a hypothesis in H that labels all elements of C_I
with 0.

1. If I = ∅ then C_I = ∅ and $h_\emptyset$ labels all points in C with 1.
2. If I = [2023], then C_I = C and h_I = 0 for any $\overline{e_i}$ in C.

3. If I ⊂ C and I ≠ ∅ then $h_I(\overline{e_\iota}) = \begin{cases} 0, for\ any\ \overline{e_\iota}\ in\ C_I \\ 1, for\ any\ \overline{e_\iota}\ in\ C_{\bar{I}} \end{cases}$

This happens because h_I computes the product of bits on positions in I. Any $\overline{e_\iota}$ with i in I by definition has a 0 on the i^{th} position so the whole product is 0. In contrast, any $\overline{e_\iota}$ with i in $\bar{I}$ will have 1 on any position in I. So, H shatters C with |C| = 2023. Then VCdim(H) ≥ 2023. But VCdim(H) ≤ 2023 from (1). Which means that VCdim(H) = 2023.

In this example, the hypothesis class H is defined as the set of all functions that take a 2023-dimensional binary vector and output the product of its elements indexed by a fixed subset I of {1,2, ... ,2023}. The claim is that H is a finite hypothesis class with VCdim(H) = 2023.

To prove this claim, we first note that the size of $\mathcal{H}$ is given by the number of possible subsets of {1,2, ... ,2023}, which is 2^{2023}. Since $\mathcal{H}$ is a finite set, this satisfies the first condition for being a finite hypothesis class.

Next, we use the fact that for any finite hypothesis class $\mathcal{H}'$, its VC dimension is bounded by $log_2|\mathcal{H}'|$. Using this bound, we can show that VCdim(H) is at most 2023, which satisfies the second condition for being a finite hypothesis class.

Finally, we need to show that $\mathcal{H}$ shatters a set of 2023 points in the Boolean hypercube of dimension 2023, i.e., that for any subset C of size 2023, there exists a function in H that correctly classifies all points in C. To do this, we consider the set C consisting of all 2023 points where all bits are 1, except for the i^{th} bit which is 0 for each i =1, 2, ..., 2023. We show that for any subset I of {1, 2, ..., 2023}, there exists a function in H that correctly classifies all points in C_I = {c in C | $C_i = 1$ for $i\ in\ I$ } as 1 and all points in $C_{\bar{I}}$ = {c in C | C_i = 1 for i not in I } as 0. This implies that $\mathcal{H}$ shatters C, and hence VCdim(H) is at least 2023. Therefore, we have shown that $\mathcal{H}$ is a finite hypothesis class with VCdim(H) = 2023, as claimed.

Answer – 1 – b)

Let's consider the class of half-spaces in n-dimensional Euclidean space, which is defined as the set of all functions $h_{(x)} = sign(w.x + b)$, where w is a vector of weights and b is a scalar bias term, and $sign$ denotes the sign function which returns +1 if its argument is positive, -1 if it is negative, and 0 if it is zero. The hypothesis class $\mathcal{H}$ is then the set of all half-spaces that can be defined in this way, for any choice of w and b.

To show that the VC dimension of $\mathcal{H}$ is 2023, we need to show that there exists a set of $n+1$ points that can be shattered by $\mathcal{H}$, and that no set of $n+2$ points can be shattered by $\mathcal{H}$. Where $n = 2023$ in our example.

First, let's show that $\mathcal{H}$ can shatter any set of $2023+1$ points. Suppose we have a set of $2023+1$ points $x_1, x_2, \ldots, x_{(2023+1)}$ in 2023-dimensional space. Without loss of generality, we can assume that the first 2023 points $x_1, x_2, \ldots, x_{(2023)}$ are linearly independent, since otherwise we can remove some redundant points without affecting the shattering property.

To shatter this set of points, we need to show that for any labeling of the points, there exists a half-space $h_{(x)} = sign(w.x+b)$ in $\mathcal{H}$ that correctly classifies all points.

Let $y_1, y_2, \ldots, y_{(2023+1)}$ be the labels of the points, where $y_i = +1\ if\ x_i$ is positive, and $y_i = -1\ if\ x_i$ is negative. We want to find a choice of w and b such that $h_{x_i} = y_i$ for all i.

We can construct such a half-space as follows. First, let's define a matrix x whose columns are the 2023-dimensional vectors $x_1, x_2, \ldots, x_{2023}$. We know that x has full rank, since its columns are linearly independent. Let's also define a vector y whose components are $y_1, y_2, \ldots, y_{2023}$. We can then solve for the weights w and bias b by minimizing the following quadratic function: $\|X_w - y\|^2$. Subject to the constraint that w is a unit vector, i.e., $\|w\| = 1$. This is a standard optimization problem in linear algebra, known as the least squares problem, and it has a unique solution given by: $w = X^{(-1)y} / \left\|X^{(-1)y}\right\|, b = -sign(w.x_{(n+1)})$.

Where $X^{(-1)}$ denotes the inverse of X. This solution corresponds to the hyperplane that passes through the first 2023 points $x_1, x_2, \ldots, x_{2023}$, and whose normal vector w points towards the point $x_{(2023+1)}$ if it is labeled positive, and away from it if it is labeled negative. The sign function ensures that the hyperplane correctly classifies $x_{(2023+1)}$.

Since we can choose any set of 2023+1 points in this way, we have shown that $\mathcal{H}$ can shatter any set of 2023+1 points.

Next, let's show that $\mathcal{H}$ cannot shatter any set of 2023+2 points. Suppose we have a set of 2023+2 points $x_1, x_2, \dots, x_{(2023+2)}$ in 2023-dimensional space. Without loss of generality, we can assume that the first 2023 points $x_1, x_2, \dots, x_{2023}$ are linearly independent.

To show that $\mathcal{H}$ cannot shatter this set of points, we need to show that there exists a labeling of the points that cannot be realized by any half-space in $\mathcal{H}$. Let's choose the labeling $y_1 = y_2 = \cdots = y_{2023} = +1$, and $y_{(2023+1)} = y_{(2023+2)} = -1$. We want to show that there is no half-space $h_{(x)} = sign(w.x + b)$ in $\mathcal{H}$ that can correctly classify these points.

Suppose such a half-space exists. Let w and b be the weights and bias of this half-space. Since $h_{x_1} = h_{x_2} = \dots = h_{x_{2023}} = +1$, we have $w.x_i + b > 0$ for all $i = 1,2,\dots,2023$. Similarly, since $h_{(x_{\{2023+1\}})} = h_{(x_{\{2023+2\}})} = -1$, we have $w.x_{\{2023+1\}} + b < 0$ and $w.x_{\{2023+2\}} + b < 0$.

We can then take the average of the first 2023 inequalities and subtract the last two inequalities to obtain: $w.(x_1 + x_2 + \cdots + x_{2023} - x_{\{2023+1\}} - x_{\{2023+2\}}) > 0$

But since $x_1 + x_2 + \cdots + x_{2023}$ are linearly independent, the vector $x_1 + x_2 + \cdots + x_{2023} - x_{\{2023+1\}} - x_{\{2023+2\}}$ is nonzero and lies in the 2023-dimensional hyperplane defined by the first 2023 points. This means that we have found a vector v such that $w.v > 0$, which contradicts the fact that w is a solution to the least squares problem, since v is orthogonal to the columns of X.

Therefore, there is no half-space in $\mathcal{H}$ that can correctly classify the points $x_1, x_2, \dots x_{\{2023+2\}}$ with the chosen labeling, and we have shown that $\mathcal{H}$ cannot shatter any set of 2023+2 points.

Thus, we have shown that the VC dimension of the class of half-spaces in 2023-dimensional Euclidean space is 2023. This implies that any classification algorithm that uses this hypothesis class can achieve arbitrarily good performance on any binary classification problem with n-dimensional inputs, as long as the true underlying distribution is well-behaved and the number of training examples is sufficiently large.

Answer – 1 – c)

Let us define an axis-aligned rectangle in the two-dimensional plane as follows: An axis-aligned rectangle R is determined by four real numbers a, b, c, and d, where a < b and c < d. The rectangle R is defined by the set of points {(x, y): a ≤ x ≤ b and c ≤ y ≤ d}.

Now let H be the class of all possible axis-aligned rectangles in the two-dimensional plane, i.e.,

H = {R(a,b,c,d) : a < b, c < d, a,b,c,d ∈ ℝ}

Note that H is an infinite hypothesis class since there are an infinite number of possible axis-aligned rectangles in the plane.

Next, we will show that VC dimension (H) = ∞.

To do this, we will show that for any positive integer N, there exists a set of N points in the plane that can be shattered by H.

Consider any set of N points in the plane, denoted by P = {(x1, y1), (x2, y2), ..., (xN, yN)}. We will construct a labeling function $f: P \longrightarrow \{0,1\}$ such that for any labeling function $g: P \longrightarrow \{0,1\}$, there exists an axis-aligned rectangle R in H such that R contains all points labeled 1 by g and does not contain any point labeled 0 by g.

Let us define f as follows:

If there exists a point in P with label 1, such that x < min(x_i), then f(x, y) = 0.

Otherwise, if there exists a point in P with label 1, such that y < min(y_i), then f(x, y) = 0.

Otherwise, if there exists a point in P with label 1, such that x > max(x_i), then f(x, y) = 0.

Otherwise, if there exists a point in P with label 1, such that y > max(y_i), then f(x, y) = 0.

Otherwise, f(x, y) = 1.

In other words, f labels a point (x, y) as 1 if it is not dominated by any point in P (i.e., no point in P has a smaller x-coordinate and no point in P has a smaller y-coordinate). Otherwise, f labels (x, y) as 0.

Now, let $g: P \longrightarrow \{0,1\}$ be any labeling function. We will construct an axis-aligned rectangle R in H such that R contains all points labeled 1 by g and does not contain any point labeled 0 by g.

Let us define a, b, c, and d as follows:

a = min(x_i) for all i such that $g(x_i)$ = 1

b = max(x_i) for all i such that $g(x_i)$ = 1

c = min(y_i) for all i such that $g(y_i)$ = 1

d = max(y_i) for all i such that $g(y_i)$ = 1

Note that a, b, c, and d are well-defined since g is a valid labeling function.

Now consider any point (x, y) labeled 1 by g. By definition of f, we have that (x, y) is not dominated by any point in P. Thus, we have x ∈ [a, b] and y ∈ [c, d].

Therefore, any point labeled 1 by g is contained in the rectangle R = R(a,b,c,d), which is an axis-aligned rectangle in H.

Next, consider any point (x, y) labeled 0 by g. We will show that (x, y) is not contained in R. Suppose, for contradiction, that (x, y) is contained in R. Then we have a ≤ x ≤ b and c ≤ y ≤ d. Since (x, y) is labeled 0 by g, we have either x < a or x > b or y < c or y > d. However, this contradicts the fact that (x, y) is contained in R. Therefore, (x, y) is not contained in R.

Thus, we have shown that for any labeling function $g: P \longrightarrow \{0,1\}$, there exists an axis-aligned rectangle R in H such that R contains all points labeled 1 by g and does not contain any point labeled 0 by g. Therefore, H can shatter any set of N points in the plane, for any positive integer N.

Since H can shatter sets of arbitrarily large size, it follows that VC dimension (H) = ∞.

Question – 2

Consider H to be the following hypothesis class:

$$\mathcal{H} = \{h_a: \mathbb{R}^3 \longrightarrow \{0,1\} \mid h_a(x) = 1_{[\|x\|_2 \le a]}(x), x = (x_1, x_2, x_3) \in \mathbb{R}^3, \|x\|_2 = \sqrt{x_1^2 + x_2^2 + x_3^2}\ \}$$

Consider the realizability assumption.

- **a)** Show that H can be (ϵ, δ)-PAC learned by giving an algorithm A and determining the sample complexity $m_H(\epsilon, \delta)$ such that the definition of PAC learnability is satisfied.
- **b)** Compute VCdim(H).

Answer – 2 – a)

To show that H can be (ϵ, δ)-PAC learnable under the realizability assumption, we need to provide an algorithm A that outputs a hypothesis h from $\mathcal{H}$, such that with probability at least $1 - \delta$, the error of h is at most ϵ. Additionally, we need to determine the sample complexity $m_{\mathcal{H}}(\epsilon, \delta)$, which is the minimum number of examples required to achieve this level of accuracy.

Algorithm A:

1. Let $S = \{(x_1, y_1), \dots (x_m, y_m)\}$be a sample of m examples drawn independently and identically from an unknown distribution D over R^3, where $x_i = (x_{i1}, x_{i2}, x_{i3})$ and $y_i \in \{0,1\}$.
2. Compute the empirical mean squared error (EMSE) of the hypotheses in $\mathcal{H}$ on S. That is, for each hypothesis h_a in $\mathcal{H}$, compute the average squared error on the sample: $EMSE(h_a) = \left(\frac{1}{m}\right) * \sum_{(i=1)^m} (h_a(x_i) - y_i)^2$,where x_i is the i^{th} example in S, y_i is the corresponding label (0 or 1), and $h_a(x_i) = 1_{\{\|x_i\|_2 \le a\}}$ is the predicted label of h on x_i.
3. Choose the hypothesis h_{a^*} in $\mathcal{H}$ that minimizes the $EMSE: h_{a^*} = argmin_{(h_a \in H)} EMSE(h_a)$
4. Output the hypothesis h = h_{a^*}.

Now we need to determine the sample complexity $m_{\mathcal{H}}(\epsilon, \delta)$ required for algorithm A to be (ϵ, δ)-PAC learnable. To do so, we will use the following theorem:

Theorem: Let $\mathcal{H}$ be a hypothesis class with finite VC dimension d_{VC}, defined as $\mathcal{H} = \{h_a: R^3 \rightarrow \{0,1\} \mid h_a(x) = 1_{[\|x\|_2 \leq a]}(x),\ x = (x_1, x_2, x_3) \in R^3, \|x\|_2 = \sqrt{(x_1^2 + x_2^2 + x_3^2)}\}$.

For any δ > 0 and ϵ > 0, if $m \geq (8/d_{VC}) * \text{In}(4/\delta) * (1/\epsilon)^2$, then with probability at least 1-δ, algorithm A is (ϵ, δ)-PAC learnable.

To apply this theorem to our hypothesis class $\mathcal{H}$, we need to compute its VC dimension d_{VC}. Recall that the VC dimension of a hypothesis class is the maximum size of a set of points that can be shattered by the class. In other words, it is the largest integer d such that there exist d points that can be labeled in all 2^d possible ways by the hypotheses in $\mathcal{H}$.

In our case, $\mathcal{H}$ is a class of threshold functions that depend on the Euclidean norm of the input vector x Therefore, the VC dimension of $\mathcal{H}$ is equal to the number of linearly separable points in the unit sphere of R^3. It is a well-known result that the maximum number of linearly separable points in the unit sphere of d-dimensional space is 2d. Therefore, the VC dimension of $\mathcal{H}$ is 3.

Using this value of d_{VC} in the theorem above, we obtain the following sample complexity for (ϵ, δ)-PAC learnability of $\mathcal{H}$: $m_{\mathcal{H}(\epsilon,\delta)} \geq (8/3) * \text{In}(4/\delta) * (1/\epsilon)^2$.

Therefore, we have shown that $\mathcal{H}$ can be (ϵ, δ)-PAC learnable with sample complexity given by the above formula.

Answer – 2 – b)

Let's first define the hypothesis class H:

$$\mathcal{H} = \{h_a: \mathbb{R}^3 \longrightarrow \{0,1\} \mid h_a(x) = 1_{[\|x\|_2 \leq a]}(x), x = (x_1, x_2, x_3) \in \mathbb{R}^3, \|x\|_2 = \sqrt{(x_1^2 + x_2^2 + x_3^2)}\}$$

Where a is a positive real number, and $1_{[\|x\|_2 \leq a]}$ is the indicator function that takes the value 1 if the condition $\|x\|_2 \leq a$ is true, and 0 otherwise.

The VC dimension of $\mathcal{H}$ is defined as the largest number of points that can be shattered by $\mathcal{H}$. To find the VC dimension of $\mathcal{H}$, we need to show that there exists a set of k points that can be shattered by $\mathcal{H}$, but any set of k+1 points cannot be shattered by $\mathcal{H}$.

Let S be a set of k points on the surface of a sphere of radius r, centered at the origin. We will consider the case where k=3, and k=4 separately.

Case k=3: Let A, B, and C be three points on S such that they form an equilateral triangle. Let a be the distance between any two points on the triangle. We can assume without loss of generality that $\mathrm{a} = 2\mathrm{r}$. Now, let h_a be a hypothesis in $\mathcal{H}$ such that $h_a(x) = 1$ if $\|x\|_2 \leq a$ and $h_a(x) = 0$ otherwise. Since A, B, and C form an equilateral triangle with distance a between them, we can assign the labels 0 and 1 to each point such that the hypothesis h_a correctly classifies all possible dichotomies of the three points. For example, we can assign the label 1 to points that are inside the triangle and label 0 to points that are outside the triangle. Similarly, we can assign the label 0 to points that are inside the triangle and label 1 to points that are outside the triangle. Therefore, H can shatter a set of three points on S.

Case k=4: Let A, B, C, and D be four points on S such that they form a tetrahedron. Let a be the distance between any two points on the tetrahedron. We can assume without loss of generality that $\mathrm{a} = 2\mathrm{r}$. Let h_a be a hypothesis in $\mathcal{H}$ such that $h_a(x) = 1$ if $\|x\|_2 \leq a$ and $h_a(x) = 0$ otherwise. We can show that there exist dichotomies of the four points that cannot be correctly classified by any hypothesis in $\mathcal{H}$.

For example, let us consider the dichotomy where points A and B are in one class, and points C and D are in the other class. It is easy to see that for any value of a, there exists a point x on S such that $h_a(x) \neq 1$ for this dichotomy. Therefore, $\mathcal{H}$ cannot shatter a set of four points on S.

Since H can shatter a set of three points but cannot shatter a set of four points, we can conclude that the VC dimension of $\mathcal{H}$ is 3.

Question – 3

Compute VCdim(H) where $\mathcal{H}$ is the following hypothesis class:

$\mathcal{H} = \{h_{\theta_1,\theta_2}: \mathbb{R}^2 \longrightarrow \{0,1\} \mid h_{\theta_1,\theta_2}(\mathrm{x}) = h_{\theta_1,\theta_2}(x_1, x_2) = 1_{[\theta_1 + x_1\ .\ \sin(\theta_2) + x_2\ .\ \cos(\theta_2) > 0]},\ \theta_1, \theta_2 \in \mathbb{R},\ \mathrm{x} \in \mathbb{R}^2\}.$

Answer – 3

To compute the VC dimension of the hypothesis class $\mathcal{H}$, we need to find the maximum number of points that can be shattered by $\mathcal{H}$. In other words, we need to find the largest value of m such that there exists a set of m points that can be labeled arbitrarily by $\mathcal{H}$.

We will start by considering the case of m = 1. Let x_1 be a point in R^2. Then, there are two possible labels for x_1: $h_{(\theta_1,\theta_2)}(x_1) = 0$, or $h_{(\theta_1,\theta_2)}(x_1) = 1$.To see this, we can rewrite the condition for $h_{(\theta_1,\theta_2)}(x_1) = 1$ as:

$\theta_1 + x_1\ .\ \sin(\theta_2) + x_2\ .\ \cos(\theta_2)\ >\ 0$

We can now fix the values of θ_1 and θ_2 to obtain the two possible labels for x_1. For example, if we set θ_1 = 0 and θ_2 = 0, we get:

$x_1 + x_2 > 0\ (h_{(0,0)}(x_1, x_2) = 1)$

If we set $\theta_1 = \pi$ and $\theta_2 = 0$, we get:

$-x_1 + x_2 > 0\ (h_{(\pi,0)}(x_1, x_2) = 1)$

Therefore, $\mathcal{H}$ can shatter a single point.

Next, we consider the case of m = 2. Let x_1 and x_2 be two points in R^2. We need to show that there exists a labeling of x_1 and x_2 that cannot be realized by any $h_{(\theta_1,\theta_2)}$ in $\mathcal{H}$. Without loss of generality, we can assume that x_1 is to the left of x_2 (i.e., $x_1 < x_2$in the $x - axis$). We can then define the labeling as follows: $h(x_1) = 0$, $h(x_2) = 1$

To see that this labeling cannot be realized by any $h_{(\theta_1,\theta_2)}$ in $\mathcal{H}$, we note that for any $h_{(\theta_1,\theta_2)}$, the value of $h_{(\theta_1,\theta_2)}(x_1)$ is always equal to $h_{(\theta_1,\theta_2)}(x_2)$, since the value of $(\theta_1$ does not affect the ordering of x_1 and x_2. Therefore, it is impossible to realize the above labeling using $\mathcal{H}$, and hence the VC dimension of $\mathcal{H}$ is 1 for any value of m. Therefore, VCdim(H) = 1.

Question – 4

Consider $\mathcal{H}_a$ to be the class of axis aligned rectangles with fixed aspect-ratio a, where $a \in \mathbb{R}$ and $a > 0$:

$$\mathcal{H}_a = \{h_{\mathrm{a,b,c,d}}: \mathbb{R}^2 \longrightarrow \{0,1\} |\, h_{\mathrm{a,b,c,d}}(\mathrm{x}) = 1_{[\mathrm{a,b}]\times[\mathrm{c,d}]}(\mathrm{x}), \mathrm{a,b,c,d} \in \mathbb{R}, \mathrm{a} < \mathrm{b}, \mathrm{c} < \mathrm{d}, \frac{d-c}{b-a} = \mathrm{a}\, \}.$$

Consider the realizability assumption.

a) Give a learning algorithm A that returns a hypothesis h_s from $\mathcal{H}_a$, $h_s = A(S)$ consistent with the training set S (h_S has empirical risk 0 on S).

b) Find the sample complexity $m_{\mathcal{H}_a}(\epsilon, \delta)$ in order to show that $\mathcal{H}_a$ is PAC-learnable.

Answer – 4 – a)

To find a learning algorithm A that returns a hypothesis h_s from $\mathcal{H}_a$, we need to first understand the hypothesis class $\mathcal{H}_a$ and the constraints imposed by it.

Hypothesis class $\mathcal{H}_a$ consists of all axis-aligned rectangles in the two-dimensional space of real numbers with a fixed aspect ratio a. The rectangle is specified by four parameters, (a, b, c, d), where a, b, c, and d are real numbers, such that a ‹ b and c ‹ d, and the ratio of the height and width of the rectangle is equal to a. The output of the hypothesis function h is 1 if the input falls within the rectangle and 0 otherwise.

Formally, $\mathcal{H}_a$ is defined as:

$$\mathcal{H}_a = \{h_{(\mathrm{a,b,c,d})}: \mathbb{R}^2 \longrightarrow \{0,1\} |\, h_{(\mathrm{a,b,c,d})}(\mathrm{x}) = 1_{([a,b]\times[c,d])}(x)\ if\ a \le x_1 \le b\ and\ c\ \le x_2 \le d, and\ 0\ otherwise, where\ a,b,c,d \in R, \mathrm{a} < \mathrm{b}, \mathrm{c} < \mathrm{d}, \frac{(d-c)}{(b-a)} = \mathrm{a}\, \}$$

Since we are given the realizability assumption, we know that there exists a hypothesis h that perfectly fits the training set S. Therefore, the goal of our learning algorithm is to find a hypothesis h_s from $\mathcal{H}_a$ that is consistent with S.

Algorithm A for finding a consistent hypothesis from $\mathcal{H}_a$ can be described as follows:

1. Initialize the hypothesis h_s as an empty rectangle, i.e., $h_s(x) = 0$ for all x in R^2.
2. For each example (x, y) in S:
    a. If $y = 1$, expand h_s to include x by updating its parameters as follows:
    - If $x_1 < a$, set $a = x_1$.

- If $x_1 > b$, set $b = x_1$.
- If $x_2 < c$, set $c = x_2$.
- If $x_2 > d$, set $d = x_2$.

b. If y = 0, shrink h_s to exclude x by updating its parameters as follows:

- If $x_1 = a$, set $a = a + \frac{1}{(2a)}$.
- If $x_1 = b$, set $b = b - \frac{1}{(2a)}$.
- If $x_2 = c$, set $c = c + \frac{1}{2}$.
- If $x_2 = d$, set $d = d - \frac{1}{2}$.

3. Return the hypothesis h_s.

Let's explain the algorithm in more detail. In step 1, we initialize h_s as an empty rectangle with 0 value for all inputs.

In step 2, we go through each example in the training set S. If the output y is 1, we expand h_s to include the input x. We update the parameters of the rectangle as follows: if the x-coordinate of x is less than the current left boundary a, we move the left boundary to x 's x -coordinate. If the x -coordinate of x is greater than the current right boundary b, we move the right boundary to x 's x -coordinate. Similarly, we update the top and bottom boundaries c and d based on the y-coordinate of x.

If the output y is 0, we shrink h_s to exclude the input x. We do this by moving the appropriate boundary of the rectangle inward by half a unit. If the x -coordinate of x is equal to the left boundary a, we move the left boundary inward by $\frac{1}{(2a)}$. If the x -coordinate of x is equal to the right boundary b, we move the right boundary inward by $\frac{1}{(2a)}$. Similarly, if the y-coordinate of x is equal to the bottom boundary c, we move the bottom boundary inward by $\frac{1}{2}$. If the y-coordinate of x is equal to the top boundary d, we move the top boundary inward by $\frac{1}{2}$.

In step 3, we return the hypothesis h_s, which is the final rectangle that includes all positive examples and excludes all negative examples from the training set S. Since we expand or shrink the rectangle only based on the positive or negative examples, the resulting hypothesis h_s is guaranteed to be consistent with the training set S.

The learning algorithm A for hypothesis class $\mathcal{H}$ uses a simple strategy of expanding or shrinking an initial empty rectangle to fit the training set S. The algorithm achieves 0 empirical risk on S by adjusting the boundaries of the rectangle based on the positive and negative examples in S. While the algorithm is simple and efficient, it requires the realizability assumption to hold and may not work well on noisy or non-linear data.

Answer – 4 – b)

To find the sample complexity $m_{\mathcal{H}}(\epsilon, \delta)$ required to PAC-learn $\mathcal{H}$, we need to find the minimum number of training examples m that guarantees that with probability of at least $1 - \delta$, any hypothesis h in $\mathcal{H}$ will have an error on the test set of at most ϵ.

First, let's examine the hypothesis class $\mathcal{H}$. Each hypothesis in $\mathcal{H}$ is a rectangle in the two-dimensional space R^2, defined by its four coordinates (a, b, c, d). The function $h_{\mathrm{a,b,c,d,a}}(\mathrm{x})$ returns 1 if x is inside the rectangle [a, b]×[c, d], and 0 otherwise.

We need to show that H is PAC-learnable, which means that there exists an algorithm A that, given a training set of m examples drawn independently from an unknown distribution D over R^2 and labeled by a target concept $c \in \mathcal{H}$, outputs a hypothesis $h \in \mathcal{H}$ such that, with probability of at least $1 - \delta$, the error of h on the test set is at most ϵ.

To find the sample complexity $m_{\mathcal{H}}(\epsilon, \delta)$, we can use the following theorem:

Theorem: Let $\mathcal{H}$ be a hypothesis class, and let ϵ, δ be positive real numbers. If $\mathcal{H}$ is PAC-learnable with sample complexity $m_{\mathcal{H}}(\epsilon/2, \delta/2)$, then $\mathcal{H}$ is PAC-learnable with sample complexity $m_{\mathcal{H}}(\epsilon, \delta)$.

Proof: Suppose that $\mathcal{H}$ is PAC-learnable with sample complexity $m_{\mathcal{H}}(\epsilon/2, \delta/2)$. Let A be an algorithm that achieves this, i.e., given a training set of $m \geq m_{\mathcal{H}}(\epsilon/2, \delta/2)$ examples drawn independently from D and labeled by a target concept c ∈ H, outputs a hypothesis $h \in \mathcal{H}$ such that, with probability of at least $1 - \delta/2$, the error of h on the test set is at most ϵ/2.

Now suppose that we want to PAC-learn $\mathcal{H}$ with sample complexity $m_{\mathcal{H}}(\epsilon, \delta)$. Let $m \geq m_{\mathcal{H}}(\epsilon/2, \delta/2)$, and let S be a training set of m examples drawn independently from D and labeled by a target concept $c \in \mathcal{H}$. We can split S into two equal-sized sets S1 and S2, and use A to learn a hypothesis $h1$ from S1 with error at most $\epsilon/2$, and a hypothesis $h2$ from S2 with error at most $\epsilon/2$. Then, by the union bound, with probability of at least $1-\frac{\delta}{2}+1-\frac{\delta}{2}=1-\delta$, the error of $h1$ on S2 and the error of $h2$ on S1 are both at most ϵ/2. We can then combine $h1$ and $h2$ into a single hypothesis h that agrees with both on their respective training sets, and whose error on the test set is at most ϵ.

Therefore, if we can find a PAC-learning algorithm A for $\mathcal{H}$ with sample complexity $m_{\mathcal{H}}(\epsilon/2, \delta/2)$, then we can use it to obtain a PAC-learning algorithm for $\mathcal{H}$ with sample complexity $m_{\mathcal{H}}(\epsilon, \delta)$

To find the sample complexity $m_{\mathcal{H}}(\epsilon/2, \delta/2)$, we can use the following theorem:

Theorem: Let $\mathcal{H}$ be a hypothesis class of functions from X to {0,1}, and let ϵ, δ be positive real numbers. If $\mathcal{H}$ is PAC-learnable with sample complexity $m_{\mathcal{H}}(\epsilon/2, \delta/2)$, then $\mathcal{H}$ is PAC-learnable with sample complexity $m_{\mathcal{H}}(\epsilon, \delta)$.

To apply the above theorem to our hypothesis class H, we need to find a PAC-learning algorithm A for $\mathcal{H}$ with sample complexity $m_{\mathcal{H}}(\epsilon/2, \delta/2)$,.

To learn a hypothesis h ∈ $\mathcal{H}$, we need to find the four coordinates (a, b, c, d) that define a rectangle containing the positive examples in the training set. We can do this by simply scanning the training set and finding the smallest rectangle that contains all positive examples.

Let R be the smallest rectangle that contains all positive examples in the training set. Then, for any point x outside of R, we can output $h(x) = 0$, since it is guaranteed to have zero error. For any point x inside R, we can output $h(x) = 1$, since it is guaranteed to have at most one error (i.e., the error on the positive example in x).

To analyze the error of h on the test set, let c be the target concept in $\mathcal{H}$, and let R_c be the smallest rectangle that contains all positive examples of c. Then, the error of h on the test set is at most the sum of the areas of R and R_c, divided by the area of the space R^2.

Let A_c be the area of R_c. Then, A_c is at most the area of any rectangle in $\mathcal{H}$, which is $(b-a)(d-c) = \frac{d-c}{(b-a)(b-a)} * (d-c) = (d-c)^2$ Therefore, A_c is at most $(d-c)^2$.

Let A be the area of R. Then, A is at most the area of any rectangle in $\mathcal{H}$ that contains all positive examples in the training set. This is a rectangle of width $(b-a)$ and height $(d-c)$, with a ratio of $\frac{(d-c)}{(b-a)} = a$. Therefore, we can write $A = a * (b-a) * (d-c)$. Since R contains all positive examples in the training set, we have $A \geq A_c$, which gives us a lower bound on a:

$$a \geq \frac{A_c}{((b-a)(d-c))} = \frac{(d-c)^2}{((b-a)(d-c))} = \frac{d-c}{(b-a)}$$

Therefore, we have:

$$A = a * (b-a)(d-c) \geq \frac{(d-c)^2}{(b-a)(b-a)} * (d-c) = \frac{(d-c)}{(b-a)} = a$$

Using the above inequalities, we can bound the error of h on the test set as follows:

$$error(h) \leq \frac{(A + A_c)}{area(R^2)} \leq \frac{(a + (d-c)^2)}{((b-a) * (d-c))}.$$

We want this error to be at most $\epsilon/2$. Therefore, we need to find values of a, b, c, d such that:

$$a < b, c < d, \frac{(d-c)}{(b-a)} = a, (a + (d-c)^2)/((b-a) * (d-c)) \leq \epsilon/2.$$

The first three conditions define a rectangle in $\mathcal{H}$. The last condition can be rewritten as:

$$d - c \leq \sqrt{(\frac{\epsilon}{2} * (b-a) * (d-c))} - a$$

Let $D = d - c$ and $A = b - a$. Then, we have:

$D \leq \sqrt{(\frac{\epsilon}{2AD})} - a$. Squaring both sides and rearranging, we get:

$$(a + \frac{\epsilon}{2} * \mathrm{A})^2 \leq (1 + \epsilon/2)AD$$

Since we want to minimize the sample complexity, we can assume that A and D are equal, and set them to their maximum value of 1. Then, we have:

$$(a + \frac{\epsilon}{2})^2 \leq (1 + \epsilon/2)$$

Taking the square root and subtracting $a + \epsilon/2$ from both sides, we get:

$$\sqrt{(1+\epsilon/2)} - \sqrt{\left(\frac{\epsilon}{2}\right)} \leq a \leq \sqrt{\left(1+\frac{\epsilon}{2}\right)} + \sqrt{\frac{\epsilon}{2}} - \frac{\epsilon}{2}$$

Therefore, we have found a rectangle in $\mathcal{H}$ that contains all positive examples in the training set and has error at most $\frac{\epsilon}{2}$ on the test set. The width of this rectangle is at most $\sqrt{(\epsilon/2)} + \sqrt{(1 + \epsilon/2)} - \epsilon/2$, and the height is at most 1. Therefore, the sample complexity required to find such a rectangle with probability at least 1 – δ/2 is:

$$m_{\mathcal{H}}\left(\frac{\epsilon}{2}, \frac{\delta}{2}\right) \leq \left(\frac{4}{\epsilon}\right) * (\log(4) + \log{(1/\delta)}).$$

Note that we used the fact that $\mathcal{H}$ is a finite hypothesis class with 4 parameters, and therefore its VC dimension is at most 4. This implies that the growth function of $\mathcal{H}$ is at most 2^{4m}, and we can apply the above theorem to find the sample complexity.

To summarize, the sample complexity required to PAC-learn $\mathcal{H}$ with error at most ϵ and confidence at least $1 - \delta$ is:

$$m_{\mathcal{H}}(\epsilon, \delta) \leq \left(\frac{4}{\epsilon}\right) * \left(\log(4) + \log\left(\frac{2}{\delta}\right)\right).$$

This means that $\mathcal{H}$ is PAC-learnable.

Question – 5

Compute VCdim(H) where $\mathcal{H}$ is the following hypothesis class:

$\mathcal{H} = \{h_\theta : \mathbb{R} \to \{0,1\} |\, h_\theta(\mathrm{x}) = 1_{[\theta,\theta+1]\cup[\theta+2,\theta+4]\cup[\theta+6,\theta+9]}(\mathrm{x}), \theta \in \mathbb{R}\}$.

Answer – 5

The VC dimension of a hypothesis class $\mathcal{H}$ is the size of the largest set of points that $\mathcal{H}$ can shatter, i.e., the maximum number of points such that $\mathcal{H}$ can realize all possible binary labelings of them.

In this case, we have a hypothesis class $\mathcal{H}$ of functions that map real numbers to the binary set {0,1}. Specifically, each function h_θ in $\mathcal{H}$ is defined by a parameter θ that determines the intervals $[\theta, \theta + 1]$, $[\theta + 2, \theta + 4]$, and $[\theta + 6, \theta + 9]$ where $h_\theta(x)$ outputs 1 for a given input x.

To compute the VC dimension of $\mathcal{H}$, we need to find the largest set of points that can be shattered by $\mathcal{H}$. In other words, we need to find the maximum number of points n such that there exist n points in R that can be labeled in all possible 2^n ways by the functions in $\mathcal{H}$.

First, let's consider the case where $n = 1$. There are only two possible labelings of a single point, namely 0 or 1. It is easy to see that $\mathcal{H}$ can realize both of these labelings by choosing appropriate values of θ. Specifically, if we set θ=0, then $h_0(x)$ outputs 1 for all x in the interval [0,1], which corresponds to the labeling 1 for the single point at $x = 0$. On the other hand, if we set$\theta = -1$, then $h_{\{-1\}}(x)$ outputs 0 for all x in the interval [0,1], which corresponds to the labeling 0 for the single point at $x = 0$. Therefore, $\mathcal{H}$ can shatter a set of 1 point, and we have VCdim(H) ≥1.

Next, let's consider the case where $n = 2$. There are $2^2 = 4$ possible labelings of two points, namely (0,0), (0,1), (1,0), and (1,1). It is not difficult to see that H cannot realize all four of these labelings. To see why, note that if $\mathcal{H}$ can realize all possible labelings of two points, then there must exist two points x_1 and x_2 such that $h_\theta(x_1) = 0$ and $h_\theta(x_2) = 1$ for some θ. However, the definition of the functions in $\mathcal{H}$ implies that if $h_\theta(x) = 1$ for some x, then $h_\theta(x') = 1$ for all x' in the same interval as x. In particular, if $h_\theta(x_1) = 0$ and $h_\theta(x_2) = 1$ for some θ, then there must exist an interval I_1 containing x_1 and an interval I_2 containing x_2 such that $h_\theta(x) = 1$ for all x in I_1 and $h_\theta(x) = 0$ for all x in I_2. But then, any point in the intersection of I_1 and I_2 would have to be labeled both 0 and 1, which is a contradiction. Therefore, $\mathcal{H}$ cannot shatter a set of 2 points, and we have VCdim(H) ≤1.

Combining these results, we have VCdim(H) = 1, i.e., the maximum number of points that $\mathcal{H}$ can shatter is 1. $\mathcal{H}$ can realize all possible labelings of a single point, but not all possible labelings of two points.