Homework 2

Benson Huang

Perceptrons

1.

First we prove the following lemma:

Lemma 1.1
Suppose that $X = \{\boldsymbol{x}_1, ..., \boldsymbol{x}_N\}$ is a linearly dependent set and $\mathcal{H}$ is the perceptron hypothesis of any dimension. Then X can't be shattered by $\mathcal{H}$.

Proof.
Suppose that $X = \{\boldsymbol{x}_1, ..., \boldsymbol{x}_N\}$ is a linearly dependent set and $\mathcal{H}$ is the perceptron hypothesis of any dimension. Then X can't be shattered by $\mathcal{H}$

Since $\mathcal{D}$ is linearly dependent, there exist a non-empty set $C = \{c_i | c_i \neq 0, 1 \leq i < N\}$ such that

$$\boldsymbol{x}_N = \sum_{c_i \in C} c_i \boldsymbol{x}_i.$$

Without loss of generality, we let

$$\boldsymbol{x}_N = \sum_{i=1}^{k} c_i \boldsymbol{x}_i, \text{ where } \ c_i \neq 0, 1 \leq k < N. \tag{1}$$

If $\mathcal{H}$ shatters X, then it also shatters $X \setminus \{\boldsymbol{x}_N\}$. Therefore, the dichotomy $(\text{sign}(c_1), ..., \text{sign}(c_k), 1, ..., 1) \in \{-1, 1\}^{N-1}$ can be implemented by $\mathcal{H}$ on $X \setminus \boldsymbol{x}_N$. For any hypothesis $h(\boldsymbol{x}) = \text{sign}(\boldsymbol{w}^T \boldsymbol{x}) \in \mathcal{H}$ that can implements this dichotomy, thus we have

$$\begin{aligned} & h(\boldsymbol{x}_1, ..., \boldsymbol{x}_{N-`1}) = (h(\boldsymbol{x}_1), ..., h(\boldsymbol{x}_{N-1})) = (\text{sign}(c_1), ..., \text{sign}(c_k), 1, 1, ..., 1) \\ \Rightarrow\ & c_i \cdot h(\boldsymbol{x}_i) = \|c_i\| \geq 0, \ \forall i \in 1, ..., k. \end{aligned} \tag{2}$$

Combine (1) and (2), we obtain

$$\begin{aligned} h(\boldsymbol{x}_N) &\overset{(1)}{=} \text{sign}\left(\boldsymbol{w}^T \sum_{c_i \in C} c_i \boldsymbol{x}_i\right) \\ &= \text{sign}\left(\sum_{c_i \in C} c_i \boldsymbol{w}^T \boldsymbol{x}_i\right) \\ &= \text{sign}\left(\sum_{c_i \in C} c_i \cdot h(\boldsymbol{x}_i)\right) \\ &\overset{(2)}{=} 1. \end{aligned} \tag{3}$$

Therefore, h can't implement the dichotomy $(\text{sign}(c_1), ..., \text{sign}(c_k), -1) \in \{-1, 1\}^{k+1}$ given the input $(\boldsymbol{x}_1, \boldsymbol{x}_k, \boldsymbol{x}_N)$, which leads to a contradiction. □

By the lemma above, if the set X (with $x_0 = 1$) is linearly dependent, then it can't be shattered by $\mathcal{H}$. We can thus use this property to eliminate impossible candidates.

[a]

$$\mathrm{rk}\left(\begin{bmatrix}1 & 7 & 8 & 9\\ 1 & 17 & 18 & 19\\ 1 & 27 & 28 & 29\end{bmatrix}\right) = 2 < 3.$$

Therefore, the set $\{(7,8,9),(17,18,19),(27,28,29)\}$ can't not be shattered by $\mathcal{H}$.

[b]

$$\mathrm{rk}\left(\begin{bmatrix}1 & 1 & 1 & 1\\ 1 & 7 & 8 & 9\\ 1 & 15 & 16 & 17\\ 1 & 21 & 23 & 25\end{bmatrix}\right) = 3 < 4.$$

Therefore, the set $\{(1,1,1),(7,8,9),(15,16,17),(21,23,25)\}$ can't not be shattered by $\mathcal{H}$.

[c]

$$\mathrm{rk}\left(\begin{bmatrix}1 & 1 & 1 & 3\\ 1 & 7 & 8 & 9\\ 1 & 15 & 16 & 17\\ 1 & 21 & 23 & 25\end{bmatrix}\right) = 4.$$

Notice that the inverse matrix of the matrix above exists. Now we can show that given inputs $(1,1,3),(7,8,9),(15,16,17),(21,23$ there exists a hypothesis $h \in \mathcal{H}$ that can implement any dichotomy in $\{-1,1\}^4$. Suppose the given dichotomy is $(y_1,y_2,y_3,y_4) \in \{-1,1\}^4$. Let

$$\begin{bmatrix}1 & 1 & 1 & 3\\ 1 & 7 & 8 & 9\\ 1 & 15 & 16 & 17\\ 1 & 21 & 23 & 25\end{bmatrix}\boldsymbol{w} = \begin{bmatrix}y_1\\ y_2\\ y_3\\ y_4\end{bmatrix},$$

then

$$\boldsymbol{w} = \begin{bmatrix}1 & 1 & 1 & 3\\ 1 & 7 & 8 & 9\\ 1 & 15 & 16 & 17\\ 1 & 21 & 23 & 25\end{bmatrix}^{-1}\begin{bmatrix}y_1\\ y_2\\ y_3\\ y_4\end{bmatrix}.$$

Choose $h(x) = \boldsymbol{w}^T\boldsymbol{x} \in \mathcal{H}$, we can implements any given dichotomy in $\{-1,1\}^4$.

[d]

$$\mathrm{rk}\left(\begin{bmatrix}1 & 1 & 3 & 5\\ 1 & 7 & 8 & 9\\ 1 & 15 & 16 & 17\\ 1 & 21 & 23 & 25\end{bmatrix}\right) = 3 < 4.$$

Therefore, the set $\{(1,3,5),(7,8,9),(15,16,17),(21,23,25)\}$ can't not be shattered by $\mathcal{H}$.

[e]

Since the set $\{(1,2,3),(4,5,6),(7,8,9),(15,16,17),(21,23,25)\}$ has more than 4 vectors, $\{(1,1,2,3),(1,4,5,6),(1,7,8,9),(1,15,$ must be linearly dependent. Therefore, it can't be shattered by $\mathcal{H}$.

2.

First, let's try to upper bound the growth function of the axis-aligned perceptrons in 2D (denoted by $\mathcal{H}$) for $N \geq 4$. We can first divide $\mathcal{H}$ into two hypothesis set:

$$\mathcal{H} = \underbrace{\left\{h(\boldsymbol{x})\middle|h(\boldsymbol{x}) = \boldsymbol{w}^T\boldsymbol{x}, \langle \boldsymbol{w}, (0,0,1)^T\rangle = 0\right\}}_{\mathcal{H}_1} \cup \underbrace{\left\{h(\boldsymbol{x})\middle|h(\boldsymbol{x}) = \boldsymbol{w}^T\boldsymbol{x}, \langle \boldsymbol{w}, (0,1,0)^T\rangle = 0\right\}}_{\mathcal{H}_2}. \tag{4}$$

Therefore, the growth function of $\mathcal{H}$ can be bounded below by the sum of the growth functions of $\mathcal{H}_1$ and $\mathcal{H}_2$;

$$m_{\mathcal{H}}(N) \leq m_{\mathcal{H}_1}(N) + m_{\mathcal{H}_2}(N).$$

Note that

$$m_{\mathcal{H}_1} = 2 \underbrace{(N-1)}_{w_1 \in (x_1, x_1)} + 2 = 2N. \tag{5}$$

Moreover, there must be two dichotomies that can be implemented by both $\mathcal{H}_1$ and $\mathcal{H}_2$: $(-1, ..., -1), (1, ..., 1) \in \{-1, 1\}^N$. Hence, the upper bound of $m_{\mathcal{H}}(N)$ can be tighter:

$$m_{\mathcal{H}}(N) \leq m_{\mathcal{H}_1}(N) + m_{\mathcal{H}_2}(N) - 2 = 2N + 2N - 2 = 4N - 2.$$

Now, we need to show that $m_{\mathcal{H}}(N) = 4N - 2$. This means that the dichotomies implemented by $\mathcal{H}_1$ and $\mathcal{H}_2$ must not overlap (except the two dichotomies $(-1, ..., -1), (1, ..., 1) \in \{-1, 1\}^N$). To achieve this condition, our inputs should satisfy some properties.

Lemma 2.1
Suppose the ordered inputs of 2-D is $X = \{\boldsymbol{x}^{(1)}, ..., \boldsymbol{x}^{(N)}\} = \{(1, x_1^{(1)}, x_2^{(1)})^T, ..., (1, x_1^{(N)}, x_2^{(N)})^T\}$ with $x_1^{(1)} < \cdots < x_N^{(N)}$ and all $x_2^{(i)}$ are distinct. First, we define

$$S_{i,j}^X = \{\boldsymbol{x}^{(i)}, ..., \boldsymbol{x}^{(j)}\},\ i \leq j.$$

Then, the dichotomies (with $(-1, ..., -1)$ and $(1, ..., 1)$ excluded) implemented by $\mathcal{H}_1$ and $\mathcal{H}_2$ will overlap if and only if there exists a integer $k \in \{1, ..., N-1\}$ such that

$$\max_{\boldsymbol{x}^{(i)} \in S_{1,k}^X} x_2^{(i)} < \min_{\boldsymbol{x}^{(j)} \in S_{k+1,N}^X} x_2^{(j)} \tag{6}$$

or

$$\min_{\boldsymbol{x}^{(i)} \in S_{1,k}^X} x_2^{(i)} > \max_{\boldsymbol{x}^{(j)} \in S_{k+1,N}^X} x_2^{(j)}. \tag{7}$$

Proof.
($\Rightarrow$)
Suppose there is a dichotomy that can be implemented by both $\mathcal{H}_1$ and $\mathcal{H}_2$. That is, there exists two hypotheses $\exists h_1(x) = \text{sign}(\boldsymbol{w}_1^T \boldsymbol{x}) \in \mathcal{H}_1, h_2 = \text{sign}(\boldsymbol{w}_2^T \boldsymbol{x}) \in \mathcal{H}_2$ such that

$$h_1(\boldsymbol{x}^{(1)}, ..., \boldsymbol{x}^{(N)}) = h_2(\boldsymbol{x}^{(1)}, ..., \boldsymbol{x}^{(N)}). \tag{8}$$

Since we have assumed that $x_1^{(1)} < \cdots < x_1^{(N)}$, from the viewpoint of $\mathcal{H}_1$ there must be a integer k such that the previous k terms of $h_1(\boldsymbol{x}^{(1)}, ..., \boldsymbol{x}^{(N)})$ have the same values that are different from the values of the last $N-k$ terms of $h_1(\boldsymbol{x}^{(1)}, ..., \boldsymbol{x}^{(N)})$ (e.g. the previous k terms are all -1 but the last $N-k$ terms are all 1). Without loss of generality, we assume that

$$h_1(\boldsymbol{x}^{(1)}, ..., \boldsymbol{x}^{(N)}) = (\underbrace{-1, ..., -1}_{k \text{ terms}}, \underbrace{1, ..., 1}_{N-k \text{ terms}}) = h_2(\boldsymbol{x}^{(1)}, ..., \boldsymbol{x}^{(N)}) \tag{9}$$

From the viewpoint of $\mathcal{H}_2$, on the other hand, (9) means that $\{x_2^{(1)}, ..., x_2^{(k)}\}$ and $\{x_2^{(k+1)}, ..., x_2^{(N)}\}$ can be separable by the plane $x_2 = x^{**}, x^{**} \in (x_2^{(k)}, x_2^{(k+1)})$. This means that $\max_{\boldsymbol{x}^{(i)} \in S_{1,k}^X} x_2^{(i)} < \min_{\boldsymbol{x}^{(j)} \in S_{k+1,N}^X} x_2^{(j)}$ or $\min_{\boldsymbol{x}^{(i)} \in S_{1,k}^X} x_2^{(i)} > \max_{\boldsymbol{x}^{(j)} \in S_{k+1,N}^X} x_2^{(j)}$.

($\Leftarrow$)
Suppose that (6) holds. Now we need to construct $h_1 \in \mathcal{H}_1$ and $h_2 \in \mathcal{H}_2$ that implement the same dichotomy. Let $k \in \{1, ..., N-1\}$ be the integer that satisfy (6). Pick two number $x^* \in (x_2^{(1)}, x_1^{(k+1)})$ and $x^{**} \in (x_2^{(k)}, x_2^{(k+1)})$. Then, we construct two hypothesis from $\mathcal{H}_1$ and $\mathcal{H}_2$ respectively:

$$h_1(\boldsymbol{x}) = \text{sign}\left((-x^*, 1, 0)^T \boldsymbol{x}\right) \in \mathcal{H}_1 \tag{10}$$

and

$$h_2(\boldsymbol{x}) = \text{sign}\left((-x^{**}, 0, 1)^T \boldsymbol{x}\right) \in \mathcal{H}_2. \tag{11}$$

From (10) we and the assumption we made about the input have

$$\begin{aligned}
& x_1^{(1)} < \cdots < x_1^{(k)} < x_1^{(k+1)} < \cdots < x_1^{(N)} \\
\Rightarrow\ & x_1^{(1)} - x^* < \cdots < x_1^{(k)} - x^* < 0 < x_1^{(k+1)} - x^* < \cdots < x_1^{(N)} - x^* \\
\Rightarrow\ & \boldsymbol{w}_1^T \boldsymbol{x}^{(1)} < \cdots < \boldsymbol{w}_1^T \boldsymbol{x}^{(k)} < \boldsymbol{w}_1^T x^{(k+1)} < \cdots < \boldsymbol{w}_1^T x^{(N)} \\
\Rightarrow\ & h_1(\boldsymbol{x}^{(1)}) = \cdots = h_1(\boldsymbol{x}^{(k)}) = -1,\ h_1(\boldsymbol{x}^{(k+1)}) = \cdots = h_1(\boldsymbol{x}^{(N)}) = 1 \\
\Rightarrow\ & h_1(\boldsymbol{x}_1, ..., \boldsymbol{x}_{N-`1}) = (h_1(\boldsymbol{x}_1), ..., h_1(\boldsymbol{x}_{N-1})) = (\underbrace{-1, ..., -1}_{k \text{ terms}},\ \underbrace{1, ..., 1}_{N-k \text{ terms}}\ ).
\end{aligned} \tag{12}$$

On the other hand, from (6) and (11), we have

$$\begin{aligned}
& \max_{\boldsymbol{x}^{(i)} \in S_{1,k}^X} x_2^{(i)} < x^{**} < \min_{\boldsymbol{x}^{(j)} \in S_{k+1,N}^X} x_2^{(j)} \\
\Rightarrow\ & \max_{\boldsymbol{x}^{(i)} \in S_{1,k}^X} x_2^{(i)} - x^{**} < 0 < \min_{\boldsymbol{x}^{(j)} \in S_{k+1,N}^X} x_2^{(j)} - x^{**} \\
\Rightarrow\ & \boldsymbol{w}_2^T \boldsymbol{x}^{(i)} < 0, \forall \boldsymbol{x}^{(i)} \in S_{1,k}^X \quad \text{and} \quad \boldsymbol{w}_2^T \boldsymbol{x}^{(j)} > 0, \forall \boldsymbol{x}^{(j)} \in S_{k+1,N}^X \\
\Rightarrow\ & h_2(\boldsymbol{x}^{(1)}) = \cdots = h_2(\boldsymbol{x}^{(k)}) = -1, h_2(\boldsymbol{x}^{(k+1)}) = \cdots = h_2(\boldsymbol{x}^{(N)}) = 1 \\
\Rightarrow\ & h_2(\boldsymbol{x}_1, ..., \boldsymbol{x}_{N-`1}) = (h_2(\boldsymbol{x}_1), ..., h_2(\boldsymbol{x}_{N-1})) = (\underbrace{-1, ..., -1}_{k \text{ terms}},\ \underbrace{1, ..., 1}_{N-k \text{ terms}}\ ).
\end{aligned} \tag{13}$$

Combine the result of (12) and (13), we obtain

$$h_1(\boldsymbol{x}^{(1)}, ..., \boldsymbol{x}^{(N)}) = h_2(\boldsymbol{x}^{(1)}, ..., \boldsymbol{x}^{(N)}) = (\underbrace{-1, ..., -1}_{k \text{ terms}},\ \underbrace{1, ..., 1}_{N-k \text{ terms}}\ ).$$

The proof is similar for the case in which (7) holds. □

From the lemma above, showing that $m_{\mathcal{H}}(N) = 4N - 2$ is equivalent to showing that there exist a inputs set X of size N that dose not follow the both of the conditions (6) and (7). That is, it must satisfy the following two conditions:

$$\max_{\boldsymbol{x}^{(i)} \in S_{1,k}^X} x_2^{(i)} > \min_{\boldsymbol{x}^{(j)} \in S_{k+1,N}^X} x_2^{(j)},\ \forall k \in \{1, ..., N-1\} \tag{14}$$

and

$$\min_{\boldsymbol{x}^{(i)} \in S_{1,k}^X} x_2^{(i)} < \max_{\boldsymbol{x}^{(j)} \in S_{k+1,N}^X} x_2^{(j)},\ \forall k \in \{1, ..., N-1\}. \tag{15}$$

Now we construct a inputs set of size N by the following steps:

1. First, we add $(1, x_1^{(1)}, x_2^{(1)})^T, (1, x_1^{(2)}, x_2^{(2)})^T$ and $(1, x_1^{(3)}, x_2^{(3)})^T$ into the inputs set, where $x_1^{(1)} < x_1^{(2)} < x_1^{(3)}$ and $x_2^{(3)} < x_2^{(1)} < x_2^{(2)}$

2. When the size of the inputs set is $k-1$ add a new input $(1, x_1^{(k)}, x_2^{(k)})$ into the inputs set, with the following 2 constrains:

$$x_1^{(k)} > x_1^{(k-1)} \tag{16}$$

and

$$\min\{x_1^{(k-2)}, x_1^{(k-1)}\} < x_1^{(k)} < \max\{x_1^{(k-2)}, x_1^{(k-1)}\}. \tag{17}$$

3. Repeat step 2. until the size of the inputs set reaches N.

Lemma 2.2
The procedure described above will generate a inputs set with size $N \geq 4$ that satisfies both of (14) and (15).

Proof.
We prove this lemma by induction. For $n = 4$ the two conditions trivially hold. Now suppose that the statement holds for $n = N - 1$. We have to discuss 3 possible cases for the value of k denoted by the two conditions.

case 1: $k \leq N - 3$

From (17), the value of $\min_{\boldsymbol{x}^{(j)} \in S^X_{k+1,N}} x_2^{(j)}$ or $\max_{\boldsymbol{x}^{(j)} \in S^X_{k+1,N}} x_2^{(j)}$ would not be $x_2^{(N)}$. Therefore, the two conditions holds by the inductions hypothesis.

case 2: $k = N - 2$
Without loss of generality, we assume that $x_2^{(N-1)} < x_2^{(N)} < x_2^{(N-2)}$. Hence, we have

$$\max_{\boldsymbol{x}^{(i)} \in S^X_{1,k}} x_2^{(i)} \geq x_2^{(N-2)} > x_2^{(N-1)} = \min_{\boldsymbol{x}^{(j)} \in S^X_{k+1,N}} x_2^{(j)}$$

and

$$\min_{\boldsymbol{x}^{(i)} \in S^X_{1,k}} x_2^{(i)} \overset{\text{I.H.}}{<} x_2^{(N-1)} \leq \max_{\boldsymbol{x}^{(j)} \in S^X_{k+1,N}} x_2^{(j)}.$$

case 3: $k = N - 1$
From (17), we have

$$\max_{\boldsymbol{x}^{(i)} \in S^X_{1,k}} x_2^{(i)} \geq x_2^{(N-2)} > x_2^{(N)} = \min_{\boldsymbol{x}^{(j)} \in S^X_{k+1,N}} x_2^{(j)}$$

and

$$\min_{\boldsymbol{x}^{(i)} \in S^X_{1,k}} x_2^{(i)} \leq x_2^{(N-1)} < x_2^{(N)} = \max_{\boldsymbol{x}^{(j)} \in S^X_{k+1,N}} x_2^{(j)}.$$

By the discussions above, the statement holds for $n = N$. Thus, we have proved Lemma 2.2. □

Based on Lemma 2.2 and Lemma 2.1, we have successfully constructed a inputs set with size N through which $\mathcal{H}_1$ and $\mathcal{H}_2$ can implement $2(N-1)$ dichotomies respectively. Plus the two dichotomies $(-1, ..., -1), (1, ..., 1) \in \{-1, 1\}^N$, we have shown that

$$m_{\mathcal{H}}(N) = 4N - 2.$$

3.

Since the hypothesis set of positively-biased 2-D perceptron (denoted by $\mathcal{H}$) is a subset of the hypothesis set of all 2-D perceptron (denoted by $\mathcal{H}'$), the VC dimension of the former must not be greater than the VC dimension of the latter. Thus, from the context of lecture 7, we know that the VC dimension of $\mathcal{H}$ must $\leq d + 1 = 2 + 1$. Now, we assume that

$$\mathcal{D} = \{\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, \boldsymbol{x}^{(3)}\} = \{(1, x_1^{(1)}, x_2^{(1)}), (1, x_1^{(2)}, x_2^{(2)}), (1, x_1^{(3)}, x_2^{(3)})\}$$

is an inputs set of size 3 that can be shattered by $\mathcal{H}'$. If the output of $\boldsymbol{w}^T \boldsymbol{x}^{(i)} = a_i$, we have

$$\begin{bmatrix} 1 & x_1^{(1)} & x_2^{(1)} \\ 1 & x_1^{(2)} & x_2^{(2)} \\ 1 & x_1^{(3)} & x_2^{(3)} \end{bmatrix} \cdot \begin{bmatrix} w_0 \\ w_1 \\ w_2 \end{bmatrix} \overset{\text{let}}{=} X \cdot \begin{bmatrix} w_0 \\ w_1 \\ w_2 \end{bmatrix} = \begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix}.$$

From Lemma 1.1 we know that if the inputs can be shattered by the perceptron, then it must be a linearly independent set. This means that the inverse of X does exist. Therefore, we have

$$\begin{bmatrix} w_0 \\ w_1 \\ w_2 \end{bmatrix} = X^{-1} \begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix}.$$

Apply the formula of the inverse matrix for 3×3 matrix, we can thus compute w_0's value:

$$w_0 = \frac{a_1}{\det(X)} \underbrace{\begin{vmatrix} x_1^{(2)} & x_2^{(2)} \\ x_1^{(3)} & x_2^{(3)} \end{vmatrix}}_{\Delta_1} + \frac{a_2}{\det(X)} \underbrace{\begin{vmatrix} x_2^{(2)} & 1 \\ x_2^{(3)} & 1 \end{vmatrix}}_{\Delta_2} + \frac{a_3}{\det(X)} \underbrace{\begin{vmatrix} 1 & x_2^{(2)} \\ 1 & x_2^{(3)} \end{vmatrix}}_{\Delta_3}. \tag{18}$$

If we want to implement the dichotomy $(y_1, y_2, y_3) \in \{-1, 1\}^3$, where

$$(y_1, y_2, y_3) = \left(\text{sign}\left(\frac{-\Delta_1}{\det(X)}\right), \text{sign}\left(\frac{-\Delta_2}{\det(X)}\right), \text{sign}\left(\frac{-\Delta_3}{\det(X)}\right)\right) \tag{19}$$

$$= (\text{sign}(a_1), \text{sign}(a_2), \text{sign}(a_3)). \tag{20}$$

In this way, w must not be greater than zero. This means that (y_1, y_2, y_3) can't be implemented by $\mathcal{H}$ for any inputs of size 3. Thus, the VC dimension of $\mathcal{H}$ must less than 3 (i.e., ≤ 2). Now, we show that the VC dimension

of $\mathcal{H}$ is 2:
Let $\mathcal{D}' = \{\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}\} = \{(1,-1,2),(1,-2,1)\}$. Now we have to implement the four dichotomies of $\{-1,1\}^2$ using the hypothesis in $\mathcal{H}$:

case 1: $(y_1, y_2) = (1,1)$
Let $\boldsymbol{w} = (3,0,-1)^T$and $h(\boldsymbol{x}) = \text{sign}(\boldsymbol{w}^T x) \in \mathcal{H}$,

$$\begin{bmatrix} 1 & -2 & 1 \\ 1 & -1 & 2 \end{bmatrix} \cdot \boldsymbol{w} = \begin{bmatrix} 1 & -2 & 1 \\ 1 & -1 & 2 \end{bmatrix} \cdot \begin{bmatrix} 3 \\ 0 \\ -1 \end{bmatrix} = \begin{bmatrix} 2 \\ 1 \end{bmatrix}$$

$$\Rightarrow h(\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}) = (1,1).$$

case 2: $(y_1, y_2) = (1,-1)$
Let $\boldsymbol{w} = (3,0,-2)^T$and $h(\boldsymbol{x}) = \text{sign}(\boldsymbol{w}^T x) \in \mathcal{H}$,

$$\begin{bmatrix} 1 & -2 & 1 \\ 1 & -1 & 2 \end{bmatrix} \cdot \boldsymbol{w} = \begin{bmatrix} 1 & -2 & 1 \\ 1 & -1 & 2 \end{bmatrix} \cdot \begin{bmatrix} 3 \\ 0 \\ -2 \end{bmatrix} = \begin{bmatrix} 1 \\ -1 \end{bmatrix}$$

$$\Rightarrow h(\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}) = (1,-1).$$

case 3: $(y_1, y_2) = (-1,1)$
Let $\boldsymbol{w} = (3,2,0)^T$and $h(\boldsymbol{x}) = \text{sign}(\boldsymbol{w}^T x) \in \mathcal{H}$,

$$\begin{bmatrix} 1 & -2 & 1 \\ 1 & -1 & 2 \end{bmatrix} \cdot \boldsymbol{w} = \begin{bmatrix} 1 & -2 & 1 \\ 1 & -1 & 2 \end{bmatrix} \cdot \begin{bmatrix} 3 \\ 2 \\ 0 \end{bmatrix} = \begin{bmatrix} -1 \\ 1 \end{bmatrix}$$

$$\Rightarrow h(\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}) = (-1,1).$$

case 4: $(y_1, y_2) = (-1,-1)$
Let $\boldsymbol{w} = (1,0,-2)^T$and $h(\boldsymbol{x}) = \text{sign}(\boldsymbol{w}^T x) \in \mathcal{H}$,

$$\begin{bmatrix} 1 & -2 & 1 \\ 1 & -1 & 2 \end{bmatrix} \cdot \boldsymbol{w} = \begin{bmatrix} 1 & -2 & 1 \\ 1 & -1 & 2 \end{bmatrix} \cdot \begin{bmatrix} 1 \\ 0 \\ -2 \end{bmatrix} = \begin{bmatrix} -1 \\ -3 \end{bmatrix}$$

$$\Rightarrow h(\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}) = (-1,-1).$$

From above we have shown that $\mathcal{D}'$ can be shattered by $\mathcal{H}$. Which means that the VC dimension of $\mathcal{H}$ is at least 2. Combine the results we derived above, we have proved that the VC dimension of $\mathcal{H}$ is 2.

Ring Hypothesis Sets

4.

We define $r(\boldsymbol{x})$ to be the distance between the input $\boldsymbol{x}$ and the origin. By the definition of $h(\boldsymbol{x})$,

$$h(\boldsymbol{x}) = \begin{cases} +1 & \text{if } a \leq [r(\boldsymbol{x})]^2 \leq b \\ -1 & \text{otherwise} \end{cases}.$$

Thus, the growth function of this hypothesis set is the same as the growth function for positive intervals mentioned in page 19 of lecture 5. So the growth function is $\binom{N+1}{2} + 1$.

5.

From 4., we know that the VC dimension of this hypothesis set should be equal to the VC dimension of the hypothesis of positive intervals, which is 2.

Deviation from Optimal Hypothesis

6.

$\forall h \in \mathcal{H}$, we have

$$E_{\text{out}}(g) - E_{\text{out}}(g_*) = E_{\text{out}}(g) - E_{\text{in}}(g) + E_{\text{in}}(g) - E_{\text{in}}(g*) + E_{\text{in}}(g_*) - E_{\text{out}}(g_*). \tag{21}$$

Since $g = \arg\min_{h\in\mathcal{H}} E_{\text{in}}(h)$, we have

$$E_{\text{in}}(g) - E_{\text{in}}(g*) \leq 0. \tag{22}$$

Combine (21) and (22), we obtain

$$\begin{aligned} E_{\text{out}}(g) - E_{\text{out}}(g_*) &= E_{\text{out}}(g) - E_{\text{in}}(g) + E_{\text{in}}(g) - E_{\text{in}}(g*) + E_{\text{in}}(g_*) - E_{\text{out}}(g_*) \\ &\leq \sup_{h\in\mathcal{H}} |E_{\text{out}}(h) - E_{\text{in}}(g)| + E_{\text{in}}(g) - E_{\text{in}}(g*) + E_{\text{in}}(g_*) - E_{\text{out}}(g_*) \\ &\leq \sup_{h\in\mathcal{H}} |E_{\text{out}}(h) - E_{\text{in}}(h)| + E_{\text{in}}(g) - E_{\text{in}}(g*) + \sup_{h\in\mathcal{H}} |E_{\text{out}}(h) - E_{\text{in}}(h)| \\ &\overset{(22)}{\leq} 2\sup_{h\in\mathcal{H}} |E_{\text{out}}(h) - E_{\text{in}}(h)| \,. \end{aligned} \tag{23}$$

$$\tag{24}$$

VC bound implies that

$$\mathbb{P}\left[sup_{h\in\mathcal{H}}|E_{\text{in}}(h) - E_{\text{out}}(h)| \leq \epsilon\right] \geq 1 - 4m_{\mathcal{H}}(2N)e^{-\frac{1}{8}\epsilon^2 N}. \tag{25}$$

Hence, if we want $\mathbb{P}\left[sup_{h\in\mathcal{H}}|E_{\text{in}}(h) - E_{\text{out}}(h)| \leq \epsilon\right]$ be less than $1-\delta$, then

$$\begin{aligned} 1 - \delta &\geq 1 - 4m_{\mathcal{H}}(2N)e^{-\frac{1}{8}\epsilon^2 N} \\ \Rightarrow \epsilon &\leq \sqrt{\frac{8}{N}\ln\frac{4m_{\mathcal{H}}(2N)}{\delta}}. \end{aligned} \tag{26}$$

Combine (23) and (26), we get an upper bount of $E_{\text{out}}(g) - E_{\text{out}}(g_*)$:

$$E_{\text{out}}(g) - E_{\text{out}}(g_*) \leq 2\sup_{h\in\mathcal{H}} |E_{\text{out}}(h) - E_{\text{in}}(h)| \leq 2\sqrt{\frac{8}{N}\ln\frac{4m_{\mathcal{H}}(2N)}{\delta}}.$$

The VC Dimension

7.

If the VC dimension of $\mathcal{H}$ is greater than d, it implies there exists a inputs set of size d through which $\mathcal{H}$ can implement 2^d dichotomies. Since the number of hypothesis in $\mathcal{H}$ is most M, $\mathcal{H}$ can implement at most M distinct dichotomies. Therefore, to ensure that $\mathcal{H}$ can shatter the inputs, we have

$$2^d \leq M$$

as a constrain on d. So the largest possible value of d is $\lfloor \log_2 M \rfloor$.

8.

Let $\mathcal{H}$ be the hypothesis set of all symertric boolean function. First we show that $d_{\text{vc}}(\mathcal{H}) \geq k+1$. To make $\mathcal{H}$ shatter $k+1$ inputs, we just need to make all the inputs have different number of 1. This allows $\mathcal{H}$ to implement any possible dichotomy because the output of the hypothesis in $\mathcal{H}$ only depends on the number of 1 in the input. If the number of the inputs is greater than $k+1$, then pigeonhole principle tells us that there must be two inputs with same number of 1. This make it impossible for $\mathcal{H}$ to implement all dichotomies because the output of these two inputs must be the same. Thus, we can ensure that $d_{\text{vc}}(\mathcal{H}) \leq k+1$.

9.

The following conditions are necessary condition for $d_{\text{vc}}(\mathcal{H}) = d$:

- Some set of d distinct inputs is shattered by $\mathcal{H}$.

  Explanation: This is the the definition of the VC dimension.

- Some set of $d+1$ distinct inputs is not shattered by $\mathcal{H}$.

  Explanation: If there is no set of $d+1$ inputs can't be shattered by $\mathcal{H}$, it means that all the sets $d+1$ inputs can be shattered by $\mathcal{H}$. This fact implies that the $d_{\text{vc}}(\mathcal{H}) \geq d+1$, which leads to a contradiction.

- Any set of $d+1$ distinct inputs is not shattered by $\mathcal{H}$. Explanation: Same as previous explanation.

10.

Here we show that if $\mathcal{H} = \{f \ : \ f(x) = \text{sign}\left(\sin(\omega x)\right), \ \omega \geq 0\}$, then $d_{vc}(\mathcal{H}) = +\infty$. Consider the labeled data set $\{(2\pi 10^{-i}, y_i)\}_{i=1}^{n}$ and choose, for any such data set, the parameter

$$\omega = \frac{1}{2}\left(1 + \sum_{i=1}^{n} \frac{1-y_i}{2} 10^i\right). \tag{27}$$

case 1: $y_j = -1$

The parameter can be rewritten as $\omega = \frac{1}{2}\left(1 + \sum_{\{i \ : \ y_i = -1\}} 10^i\right)$. where we observe that, for any point $x_j = 2\pi 10^{-j}$ in the considered data set such that $y_j = -1$, the term 10^j appears in the sum. This leads to

$$\begin{aligned}
\omega x_j &= \pi 10^{-j}\left(1 + \sum_{\{i \ : \ y_i = -1\}} 10^i\right) \\
&= \pi 10^{-j}\left(1 + 10^j + \sum_{\{i \ : \ y_i = -1, \ i \neq j\}} 10^i\right) \\
&= \pi\left(10^{-j} + 1 + \sum_{\{i \ : \ y_i = -1, \ i \neq j\}} 10^{i-j}\right) \\
&= \pi\left(10^{-j} + 1 + \sum_{\{i \ : \ y_i = -1, \ i > j\}} 10^{i-j} + \sum_{\{i \ : \ y_i = -1, \ i < j\}} 10^{i-j}\right)
\end{aligned}$$

For all $i > j$, the terms 10^{i-j} are positive powers of 10 and thus are even numbers that can be written as $2k_i$ for some $k_i \in \mathbb{N}$. Therefore, we have

$$\sum_{\{i \ : \ y_i = -1, \ i > j\}} 10^{i-j} = \sum_{\{i \ : \ y_i = -1, \ i > j\}} 2k_i = 2k \tag{28}$$

for some $k \in \mathbb{N}$, which gives

$$\omega x_j = \pi\left(10^{-j} + 1 + \sum_{\{i \ : \ y_i = -1, \ i < j\}} 10^{i-j}\right) + 2k\pi.$$

Regarding the remaining sum, we have

$$\sum_{\{i \ : \ y_i = -1, \ i < j\}} 10^{i-j} < \sum_{i=1}^{+\infty} 10^{-i} = \sum_{i=0}^{+\infty} 10^{-i} - 1 \tag{29}$$

The geometric series $S_n = \sum_{i=0}^{n-1} 10^{-i}$, of first term 1 and common ratio 0.1, is known to converge towards

$$\sum_{i=0}^{+\infty} 10^{-i} = \frac{1}{1-0.1},$$

which gives

$$\sum_{\{i \ : \ y_i = -1, \ i < j\}} 10^{i-j} < \frac{1}{1-0.1} - 1 = \frac{1}{9}.$$

Define

$$\epsilon = 10^{-j} + \sum_{\{i \ : \ y_i = -1, \ i < j\}} 10^{i-j}$$

and rewrite ωx_j as

$$\omega x_j = \pi\left(1 + \epsilon\right) + 2k\pi.$$

Given that $10^{-j} \leq 0.1$ and the previous result, we have $0 < \epsilon < 1/9 + 1/10 = 1$, and thus

$$\pi < \pi(1+\epsilon) < 2\pi \quad \Rightarrow \quad \sin(\omega x_j) < 0.$$

Hence, the classifier correctly predicts all negative labels $y_j = -1 = \text{sign}(\sin(\omega x_j)) = f(x_j)$.

case 2: $y_j = 1$
The same steps can be reproduce with positive labels $y_j = +1$ with the difference that the term 10^j does not appear in the sum defning ω. This leads to

$$\begin{aligned}\omega x_j &= \pi 10^{-j}\left(1 + \sum_{\{i\,:\,y_i=-1,\ i\neq j\}} 10^i\right)\\ &= \pi\left(10^{-j} + \sum_{\{i\,:\,y_i=-1,\ i>j\}} 10^{i-j} + \sum_{\{i\,:\,y_i=-1,\ i<j\}} 10^{i-j}\right)\\ &= \pi\epsilon + 2k\pi\end{aligned}$$

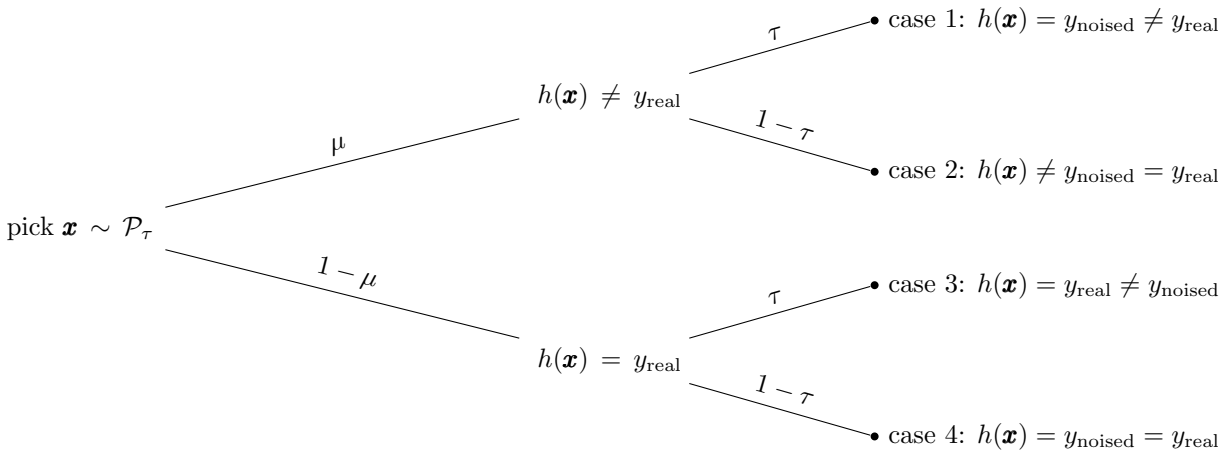
with $0 < \pi\epsilon < \pi$. So $\sin(\omega x_j) > 0$. Thus, all positively labeled points are also correctly classified by f using the particular choice of ω. Since the steps above are valid for any labeling of the points, we proved that $\mathcal{H}$ shatters the set of points. In addition, the proof is valid for any number of points n, which shows that $\mathcal{H}$ can shatter sets of points of any size and thus has infinite VC-dimension.

proof reference: https://mlweb.loria.fr/book/en/VCdiminfinite.html

Noise and Error

11.

We denote $E_{\text{out}}(h, 0)$ by μ. Also, we let $y_{\text{real}} = f(\boldsymbol{x})$, where f is the original target function, and we let y_{noised} represent the noised label. Let's draw the probability tree diagram:



From the figure above, we have

$$\begin{aligned}E_{\text{out}}(h, \tau) &= \mu(1-\tau) + (1-\mu)\tau\\ &= \mu(1-2\tau) + \tau.\end{aligned}$$

From the equation above, we obtain

$$\mu = E_{\text{out}}(h, 0) = \frac{E_{\text{out}}(h, \tau) - \tau}{1 - 2\tau}.$$

12.

Since $\boldsymbol{x}$ is drawn uniformly from $[0, 1]^3$, there are three cases with equal probability:

case 1: $f(x) = 1$
The expected square error is

$$0.1 \times [(1 \bmod 3 + 1) - 1]^2 + 0.2 \times [(2 \bmod 3 + 1) - 1]^2 = 0.9.$$

case 2: $f(x) = 2$
The expected square error is

$$0.1 \times [(2 \bmod 3 + 1) - 2]^2 + 0.2 \times [(3 \bmod 3 + 1) - 2]^2 = 0.3.$$

case 3: $f(x) = 3$
The expected square error is

$$0.1 \times [(3 \bmod 3 + 1) - 3]^2 + 0.2 \times [(4 \bmod 3 + 1) - 3]^2 = 0.6.$$

So the total expected value of the squared error is $\frac{1}{3} \times (0.9 + 0.3 + 0.6) = 0.6$.

13.

$$\begin{aligned}
\Delta(f, f_*) &= \mathbb{E}_{\boldsymbol{x} \sim P(\boldsymbol{x})}(f(\boldsymbol{x}) - f_*(\boldsymbol{x}))^2 \\
&= \sum_{i=1}^{3} P(f(\boldsymbol{x}) = i)(i - f_*(\boldsymbol{x}))^2 \\
&= P(f(\boldsymbol{x}) = 1)(1 - (0.7 \cdot 1 + 0.1 \cdot 2 + 0.2 \cdot 3))^2 + \\
&\quad P(f(\boldsymbol{x}) = 2)(2 - (0.7 \cdot 2 + 0.1 \cdot 3 + 0.2 \cdot 1))^2 + \\
&\quad P(f(\boldsymbol{x}) = 3)(3 - (0.7 \cdot 3 + 0.1 \cdot 1 + 0.2 \cdot 2))^2 \\
&= \frac{1}{3} \times (0.5^2 + 0.1^2 + 0.4^2) \\
&= 0.14.
\end{aligned}$$

Decision Stump

14.

VC bound says that

$$\mathbb{P}\left[\underset{h \in \mathcal{H}}{sup} |E_{\text{in}}(h) - E_{\text{out}}(h)| > \epsilon\right] \leq 4 m_{\mathcal{H}}(2N) e^{-\frac{1}{8}\epsilon^2 N}.$$

Hence, we have

$$\delta \leq 4 m_{\mathcal{H}}(2N) e^{-\frac{1}{8}\epsilon^2 N} \tag{30}$$

$$\Rightarrow \frac{8}{\epsilon^2} \ln \frac{4 m_{\mathcal{H}}(2N)}{\delta} \leq N. \tag{31}$$

Since $m_{\mathcal{H}}(\mathcal{H}) = 2N$, given $\epsilon = 0.1$ and $\delta = 0.1$, we get

$$800 \ln \frac{4 \cdot 4N}{0.1} \leq N.$$

Among the five choices, 12000 is the smallest number such that the inequality above holds.

15.

If $\theta > 0$, the hypothesis only make incorrect prediction when $x \in (0, \theta]$, so $E_{\text{out}} = \theta/2$. If $\theta \leq 0$, the hypothesis only make incorrect prediction when $x \in [\theta, 0)$, so $E_{\text{out}} = -\theta/2$. Therefore, $E_{\text{out}} = |\theta|/2$.