Machine Learning, 2024 Spring
Assignment 5

Notice

Plagiarizer will get 0 points.
LaTeXis highly recommended. Otherwise you should write as legibly as possible.

Problem 1

(15 points)Which of the following are possible growth functions $m_{\mathcal{H}}(N)$ for some hypothesis set: $1+N; 1+N+\frac{N(N-1)}{2}; 2^N; 2^{\lfloor\sqrt{N}\rfloor}; 2^{\lfloor N/2\rfloor}; 1+N+\frac{N(N-1)(N-2)}{6}$.

Solution:

It is known that $m_{\mathcal{H}}(N)$ is either equal to 2^N or has a polynomial upper bound. Therefore, all the functions except $1+N+\frac{N(N-1)(N-2)}{6}, 2^{\lfloor\sqrt{N}\rfloor}, 2^{\lfloor N/2\rfloor}$ are possible growth functions $m_{\mathcal{H}}(N)$.

Problem 2

(15 points)For an $\mathcal{H}$ with $d_{\mathrm{vc}}=10$, what sample size do you need (as prescribed by the generalization bound) to have a 95% confidence that your generalization error is at most 0.05?

Solution:

From $E_{\text{out}}(g) \leq E_{\text{in}}(g)+\sqrt{\frac{8}{N}\ln\left(\frac{4\left((2N)^{d_{\mathrm{vc}}}+1\right)}{\delta}\right)}$, we can get that

$$\sqrt{\frac{8}{N}\ln\left(\frac{4\left((2N)^{d_{\mathrm{vc}}}+1\right)}{\delta}\right)} \leq 0.05$$

By solving the inequality, we can get the smallest $N=452957$.

Problem 3

(15 points)Let $\mathcal{H}=\{h_1,h_2,\ldots,h_M\}$ with some finite M. Prove that $d_{\mathrm{vc}}(\mathcal{H}) \leq \log_2 M$.

Solution:

Since there are M hypotheses, a total of M scenarios can be distinguished. $d_{\mathrm{vc}}(\mathcal{H})$ means that for $n=d_{\mathrm{vc}}(\mathcal{H})$ sets of data, these M hypotheses can distinguish all $2^{d_{\mathrm{vc}}(\mathcal{H})}$ cases, and at most M cases in total, thus

$$2^{d_{\mathrm{vc}}(\mathcal{H})} \leq M$$
$$d_{\mathrm{vc}}(\mathcal{H}) \leq \log_2 M$$

Problem 4

(15 points)Let $\mathcal{H}_1,\mathcal{H}_2,\ldots,\mathcal{H}_K$ be K hypothesis sets with finite VC dimension d_{vc}. Let $\mathcal{H}=\mathcal{H}_1\cup\mathcal{H}_2\cup\cdots\cup\mathcal{H}_K$ be the union of these models. Show that $d_{\mathrm{vc}}(\mathcal{H})<K(d_{\mathrm{vc}}+1)$.

Solution:

First, we prove a conclusion

$$d_{\mathrm{vc}}\left(\bigcup_{k=1}^{K}\mathcal{H}_k\right) \leq K-1+\sum_{k=1}^{K} d_{\mathrm{vc}}\left(\mathcal{H}_k\right)$$

We prove this conclusion by mathematical induction.

Proof: when $K=2$, we can get that

$$d_{\mathrm{vc}}\left(\bigcup_{k=1}^{2}\mathcal{H}_k\right) \leq 1+\sum_{k=1}^{2} d_{\mathrm{vc}}\left(\mathcal{H}_k\right)$$

Suppose that

$$d_{\mathrm{vc}}\left(\bigcup_{k=1}^{2}\mathcal{H}_k\right) \geq 2+\sum_{k=1}^{2} d_{\mathrm{vc}}\left(\mathcal{H}_k\right)$$

then

$$m_{\mathcal{H}_1 \bigcup \mathcal{H}_2}\left(d_1+d_2+2\right) \geq 2^{d_1+d_2+2}$$

However,

$$\begin{aligned} m_{\mathcal{H}_1 \bigcup \mathcal{H}_2}\left(d_1+d_2+2\right) &\leq m_{\mathcal{H}_1}\left(d_1+d_2+2\right)+m_{\mathcal{H}_2}\left(d_1+d_2+2\right) \\ &\leq \sum_{i=0}^{d_1}\tbinom{d_1+d_2+2}{i}+\sum_{i=0}^{d_2}\tbinom{d_1+d_2+2}{i} \\ &= 2^{d_1+d_2+2}-\tbinom{d_1+d_2+2}{d_1+1} < 2^{d_1+d_2+2} \end{aligned}$$

Therefore, $d_{\mathrm{vc}}\left(\bigcup_{k=1}^{2}\mathcal{H}_k\right) \leq 1+\sum_{k=1}^{2} d_{\mathrm{vc}}\left(\mathcal{H}_k\right)$. Suppose that when $K=n$, the conclusion is valid, then for $K=n+1$

$$\begin{aligned} d_{\mathrm{vc}}\left(\bigcup_{k=1}^{n+1}\mathcal{H}_k\right) &= d_{\mathrm{vc}}\left(\left(\bigcup_{k=1}^{n}\mathcal{H}_k\right)\bigcup\mathcal{H}_{n+1}\right) \\ &\leq 1+d_{\mathrm{vc}}\left(\bigcup_{k=1}^{n}\mathcal{H}_k\right)+d_{\mathrm{vc}}\left(\mathcal{H}_{n+1}\right) \\ &\leq 1+n-1+\sum_{k=1}^{n} d_{\mathrm{vc}}\left(\mathcal{H}_k\right)+d_{\mathrm{vc}}\left(\mathcal{H}_{n+1}\right) \\ &= n+\sum_{k=1}^{n+1} d_{\mathrm{vc}}\left(\mathcal{H}_k\right) \end{aligned}$$

Therefore, for $K=n+1$, the conclusion is still valid.

As $\mathcal{H}=\mathcal{H}_1 \cup \mathcal{H}_2 \cup \cdots \cup \mathcal{H}_K$, $d_{\mathrm{vc}}\left(\mathcal{H}_k\right)=d_{\mathrm{vc}}\ \forall\, k=1,2,...,K$, we can get that

$$d_{\mathrm{vc}}(\mathcal{H}) \leq K-1+K d_{\mathrm{vc}} < K\left(d_{\mathrm{vc}}+1\right)$$

Problem 5

(40 points)In this part, you need to complete some mathematical proofs about VC dimension. Suppose the hypothesis set

$$\mathcal{H}=\{f(x,\alpha)=\operatorname{sign}(\sin(\alpha x))\ |, \alpha \in \mathbb{R}\}$$

where x and f are feature and label, respectively.

- Show that $\mathcal{H}$ cannot shatter the points $x_1=1, x_2=2, x_3=3, x_4=4$.

(Key: Mathematically, you need to show that there exists y_1, y_2, y_3, y_4, for any $\alpha \in \mathbb{R}, f(x_i) \neq y_i, i = 1, 2, 3, 4$, for example , +1, +1, −1, +1)

- Show that the VC dimension of $\mathcal{H}$ is ∞. (Note the difference between it and the first question)

(Key: Mathematically, you have to prove that for any label sets $y_1, \cdots, y_m, m \in \mathbb{N}$, there exists $\alpha \in \mathbb{R}$ and $x_i, i = 1, 2, \cdots, m$ such that $f(x; \alpha)$ can generate this set of labels. Consider the points $x_i = 10^{-i} \ldots$)

Solution:

1) Shattering means that for any possible labeling of the points, there exists a parameter α such that $f(x, \alpha) = \text{sign}(\sin(\alpha x))$ can produce those labels. Consider the labeling $y_1 = +1, y_2 = +1, y_3 = -1, y_4 = +1$ for $x_1 = 1, x_2 = 2, x_3 = 3, x_4 = 4$. The sine function, $\sin(\alpha x)$ is is periodic with a period of 2π. This means that it repeats its values every 2π units. For the chosen labeling, the sign of $\sin(\alpha x)$ must change between x_2 and x_3, and then again between x_3 and x_4. Given the periodic nature of the sine function and the distances between the points, it's impossible to choose an α that will result in the sine function having the required sign changes between these specific points. The sine function would need to complete more than half a period between x_2 and x_3 and less than half a period between x_3 and x_4, which is contradictory given the uniform spacing of the points. Therefore, it is impossible to find an α that allows $\mathcal{H}$ to shatter points $x_1 = 1, x_2 = 2, x_3 = 3, x_4 = 4$ with the labeling $+1, +1, -1, +1$.

2) Consider the labeled data set $\left(2\pi 10^{-i}, y_i\right)_{i=1}^{n}$ and choose, for any such data set, the parameter $\alpha = \frac{1}{2}\left(1 + \sum_{i=1}^{n} \frac{1-y_i}{2} 10^i\right)$. 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

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

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

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

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

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

Regarding the remaining sum, we have

$$
\sum_{i: y_i = -1, i < j} 10^{i-j} < \sum_{i=1}^{+\infty} 10^{-i} = \sum_{i=0}^{+\infty} 10^{-i} - 1 = \frac{1}{9}
$$

Let define $\epsilon = 10^{-j} + \sum_{i: y_i = -1, i < j} 10^{i-j}$ and rewrite αx_j as $\alpha x_j = \pi(1 + \epsilon) + 2k\pi$. Since $0 < \epsilon < 1$, thus $\pi < \pi(1 + \epsilon) < 2\pi$ and $\sin(\alpha x_j) < 0$. Hence, the classifier correctly predicts all negative labels $y_j = -1 = \text{sign}(\sin(\alpha x_j)) = f(x_j)$.

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 defining α. This leads to

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

with $0 < \pi\epsilon < \pi$ and $\sin(\alpha x_j) > 0$.

Thus, all positively labeled points are also correctly classified by f using the particular choice of α. 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.