# Machine Learning Foundations Homework 2

**Problem 1** (60 points)**.**

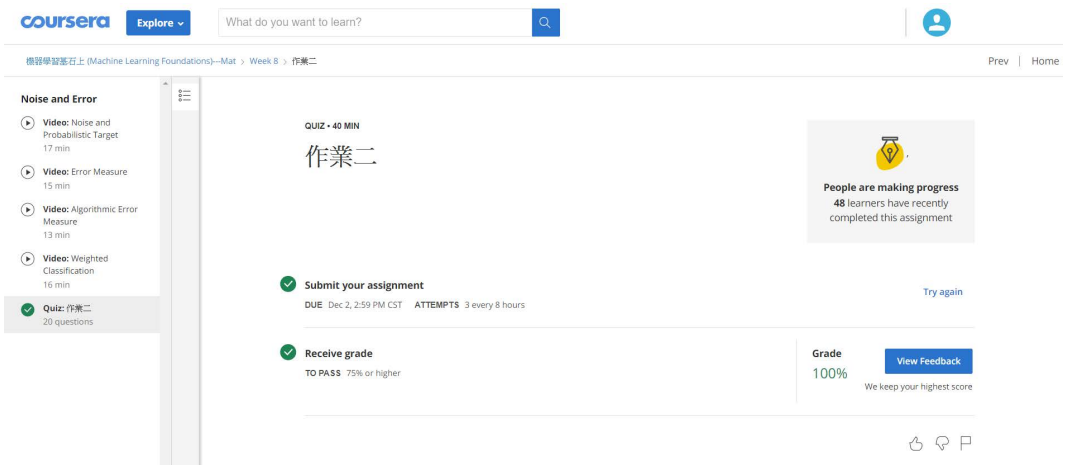



Figure 1: Q1 Screenshot (100 points). Please zoom in to see the details.

**Problem 2** (20 points)**.**

***Sol.*** We can show that there exists a set of 4 points that allows all $2^4 = 16$ dichotomies to be implemented.

Consider a set of 4 points which form a rectangle with diagonal points at $(0, 0)$ and $(2, 1)$.

Rotating these four points by 30 degrees, we obtain the following points: $(0, 0)$, $(\sqrt{3}, 1)$, $(\sqrt{3} - \frac{1}{2}, 1 + \frac{\sqrt{3}}{2})$, and $(-\frac{1}{2}, \frac{\sqrt{3}}{2})$.

Figure 2: The resultant points.

Using these points, it is possible to shatter $\mathcal{H}$, as there exists the following dichotomies:

- 1 dichotomy that excludes all 4 points;
- 4 dichotomies that includes just 1 point and excludes the other three;
- 6 dichotomies that includes 2 points and excludes the other two;
- 4 dichotomies that includes 3 points and excludes the remaining one;
- and 1 dichotomy that includes all 4 points.

Combined, this yields $16 = 2^4$ dichotomies for 4 points.

Specifically, note that with these set of points, it is possible to include only the north-south diagonal pair (that is, $(\sqrt{3} - \frac{1}{2}, 1 + \frac{\sqrt{3}}{2})$ and $(0, 0)$) and exclude the other two points. In addition, it is also possible to include only the east-west diagonal pair (that is, $(-\frac{1}{2}, \frac{\sqrt{3}}{2})$ and $(\sqrt{3}, 1)$) and exclude the other two points. It is not possible to do this with some other configurations of 4 points.

Since shattering is possible for $N = 4$, the VC dimension is *at least* 4.

□

**Problem 3** (20 points)**.**

***Sol.*** For this $\mathcal{H}$, the VC dimension $d_{\text{vc}} = \infty$; that is, for every possible input set size $N$, it is possible to find a set of $N$ points that can be shattered, i.e. we need to find $2^N$ values for $\alpha$ (that is, $\alpha_1, \alpha_2, \ldots, \alpha_{2^N}$) such that all possible dichotomies can be created.

To make this problem easier, we can think of $h_\alpha$ as taking the result of $\alpha x$ modulo 4 and checking if it falls inside the range of $[1, 3]$. If it does, then $h_\alpha(x) = -1$; else if it falls outside (i.e. $[0, 4) - [1, 3]$) then $h_\alpha(x) = +1$.

To this end, let the $i$th point's value be at $4^i$, and the first $\alpha$ be 0 (i.e. $\alpha_1 = 0$). At this point all values are strictly in $+1$ territory. For the remaining $\alpha_k$, we can see that each $4^i\alpha_k$ moves the point at different "speeds"; that is, it is possible to move the $i$th point a different distance to the other points. In this way, it is possible to shatter $N$ points.

For example, consider a case with $N = 3$, where point 3 moves twice as fast as point 2, which in turn moves twice as fast as point 1. From $\alpha_1$ to $\alpha_4$, point 1 is in the $[0, 1)$ region, but for the remaining $\alpha$ it is safely in $[1, 3)$ region. However, point 2, which travels twice as fast as point 1, is in the $h(x) = +1$ region during $\alpha_1$, $\alpha_2$, $\alpha_5$, $\alpha_6$. Meanwhile point 3 alternates between $+1$ and $-1$ every $\alpha$. This enumerates all possible 8 dichotomies.

□

**Problem 4** (20 points)**.**

***Sol.*** For any two hypothesis sets $A$ and $B$ with nonempty intersections, let their VC dimensions be $d_{\text{vc}}(A)$ and $d_{\text{vc}}(B)$ respectively.

To prove $d_{\text{vc}}(\mathcal{H}_1 \cap \mathcal{H}_2) \leq d_{\text{vc}}(\mathcal{H}_1)$, first observe the following:

If $A \subseteq B$, then $d_{\text{vc}}(A) \leq d_{\text{vc}}(B)$. This is because if $d_{\text{vc}}(A)$ is the largest amount of points that can be shattered by $A$, then $B$ can shatter **at least** $d_{\text{vc}}(A)$ points, since $B$ contains all of the hypotheses in $A$. If $B$ has other hypotheses not in $A$, then $B$ can possibly shatter more points than $A$.

Also notice that for any sets $S$ and $T$, $S \cap T \subseteq S$ by definition of intersection.

Combining these two observations, we can conclude that since $\mathcal{H}_1 \cap \mathcal{H}_2 \subseteq \mathcal{H}_1$, then $d_{\text{vc}}(\mathcal{H}_1 \cap \mathcal{H}_2) \leq d_{\text{vc}}(\mathcal{H}_1)$. □

**Problem 5** (20 points)**.**

***Sol.*** We can think of $\mathcal{H}_1 \cup \mathcal{H}_2$ as: for every $h \in \mathcal{H}_1$ that determines $h(x) = +1$ when $x < \theta$ and $-1$ when $x > \theta$, there is a corresponding $h' \in \mathcal{H}_2$ that determines $h'(x) = +1$ when $x < \theta$ and $+1$ when $x > \theta$.

Thus, for a set of $N$ points $x_1, x_2, \ldots, x_N$, $\mathcal{H}_1 \cup \mathcal{H}_2$ can create 2 dichotomies in between each possible point, as well as two extra dichotomies that determines all points to be $+1$ or all $-1$.

That is, if we sort the points so that they are in ascending order, then between points $x_i$ and $x_{i+1}$, two dichotomies are formed: one which determines that all points to the right of $x_1$ are $+1$ (and everything else $-1$), and one which determines that all points to the right of $x_1$ are $-1$. Combined with t he dichotomy that determines all points to be $+1$ and the dichotomy that determines all points to be $-1$, we arrive at $m_{\mathcal{H}_1 \cup \mathcal{H}_2}(N) = 2(N-1) + 2 = 2N$ dichotomies.

Note that this function holds even if there is is only one point ($N = 1$), where $m_{\mathcal{H}_1 \cup \mathcal{H}_2}(1) = 2$ as the only two possible dichotomies are $h(x_1) = +1$ and $h(x_1) = -1$.

Since $m_{\mathcal{H}_1 \cup \mathcal{H}_2}(N) = 2N$, the VC dimension can easily be determined as $d_{\text{vc}} = 2$ as it is the largest value that enables $m_{\mathcal{H}_1 \cup \mathcal{H}_2}(N) = 2^N$.

□

**Problem 6** (20 points)**.**

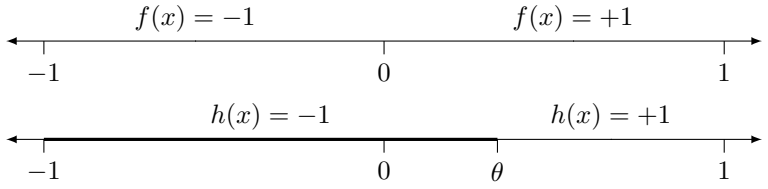
***Sol.*** To calculate $E_{\text{out}}(h_{s,\theta})$, we need to find the probability that the hypothesis $h_{s,\theta} = h$, given the input vector $x$, incorrectly predicts the correct output $y$ where $h(x)$ is attempting to approximate a noisy function $f(x)$ and $y$ is iid on the distribution $P(y|x)$,

First observe the following:

- The target function $f(x)$ will output the correct output $y$ with probability 0.8, and will output an incorrect value with probability 0.2.
- Since $f$ can only ever output $-1$ or $+1$, we only need to consider the values of $s = -1, 1$ in $h_{s,\theta}(x) = s \cdot \text{sgn}(x - \theta)$ so that $h$ also only outputs $\pm 1$.
- If we let $\mu$ be the probability that $h$ **incorrectly** predicts the value of $f$ (i.e. the chance that $h(x) \neq f(x)$), then the probability that $h$ fails to output the correct output $y$ is $0.8\mu + 0.2(1 - \mu) = 0.6\mu + 0.2$ (that is, either $f$ is correct but $h(x) \neq f(x)$, or $f$ is wrong and $h(x) = f(x)$).
- Note that if $f(x) \neq y$ and $h(x) \neq f(x)$, then $h(x) = y$, since both $h$ and $f$ can only output $+1$ or $-1$.

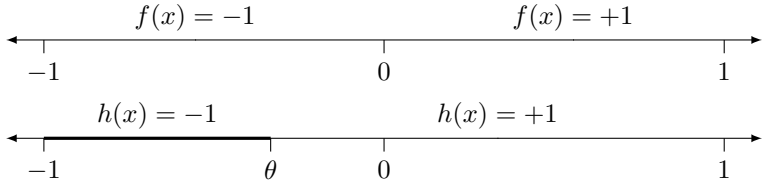
To find $\mu$, we can split this problem into four cases, which is when $s = +1$ or $s = -1$, and $\theta > 0$ or $\theta \leq 0$. We can assume for these cases that $f(x)$ is correct, since the $\mu$ in $0.8\mu$ refers to when $f$ is correct; in addition, the $(1 - \mu)$ in $0.2(1 - \mu)$ refers to when $f(x)$ is wrong and $h(x) = f(x)$, which is related to to $\mu$ and thus trivial to obtain.

**Case 1**: $s = +1$ and $\theta > 0$.



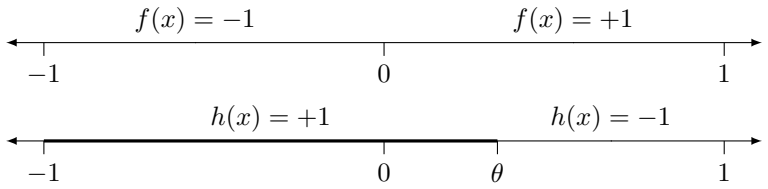

We can see that from $0 < x < \theta$, $h(x) \neq f(x)$. Since $[0, 1]$ is half of the possible range, and in that range $\theta$ of it is wrong, $\mu = \frac{\theta}{2}$.

**Case 2**: $s = +1$ and $\theta \leq 0$.



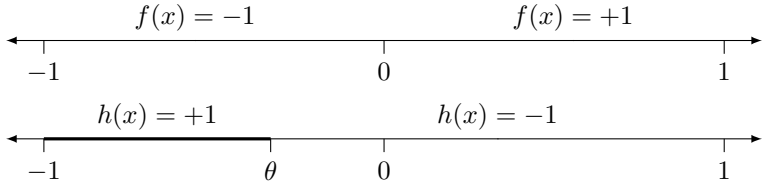

Similarly, we can see that when $\theta < x \leq 0$, $h(x) \neq f(x)$. By symmetry, we obtain $\mu = \frac{|\theta|}{2}$ when $s = 1$.

**Case 3**: $s = -1$ and $\theta > 0$.




We can see that $h(x) \neq f(x)$ everywhere *except* $0 < x < \theta$. Using the result from case 1, the $\mu$ for this case is $\mu = 1 - \frac{\theta}{2}$.

**Case 4**: $s = -1$ and $\theta \leq 0$.




We can see that $h(x) \neq f(x)$ everywhere *except* $0 < x < \theta$. Using the result from case 2 and by symmetry, $\mu = 1 - \frac{|\theta|}{2}$ when $s = -1$.

Putting these cases together, we have

$$\mu = \begin{cases} \frac{|\theta|}{2} & \text{if } s = +1 \\ 1 - \frac{|\theta|}{2} & \text{if } s = -1 \end{cases}$$

Thus

$$E_{\text{out}}(h) = \begin{cases} 0.8\mu + 0.2(1-\mu) = 0.6(\frac{|\theta|}{2}) + 0.2 & \text{if } s = +1 \\ 0.8\mu + 0.2(1-\mu) = 0.6(1 - \frac{|\theta|}{2}) + 0.2 & \text{if } s = -1 \end{cases}$$

We can simplify this by making moving this into a single equation, by combining the two cases together:

$$E_{\text{out}} = (\frac{s+1}{2})[0.6(\frac{|\theta|}{2}) + 0.2] + (\frac{1-s}{2})[0.6(1 - \frac{|\theta|}{2}) + 0.2]$$

After some simplification, this results in

$$E_{\text{out}} = 0.5 + 0.3s(|\theta| - 1)$$

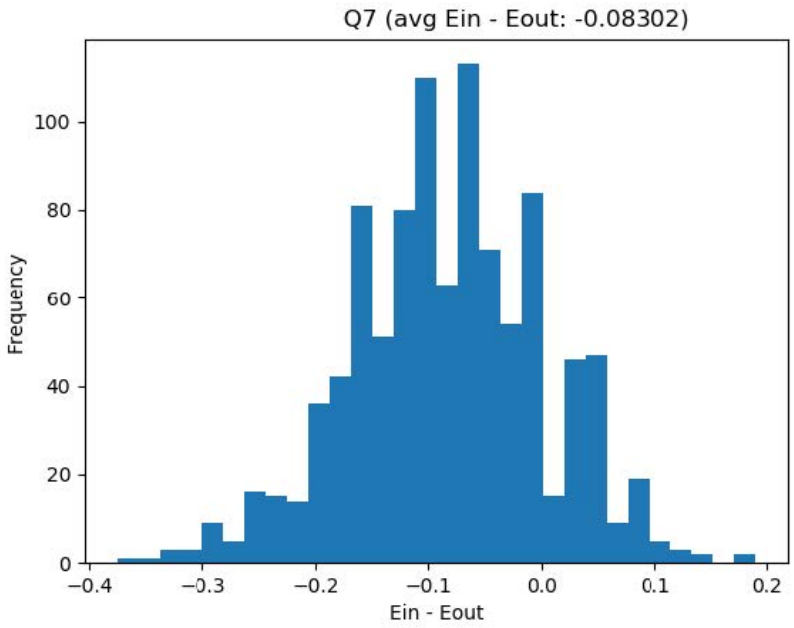
which completes the proof.

□

**Problem 7** (20 points)**.**

- The experiment was run 1000 times using seeds from 1234 to 2235.
- The histogram was plotted using 30 bins.
- The program prints the execution time and saves the file as an image named "20.png".

The histogram is as follows:



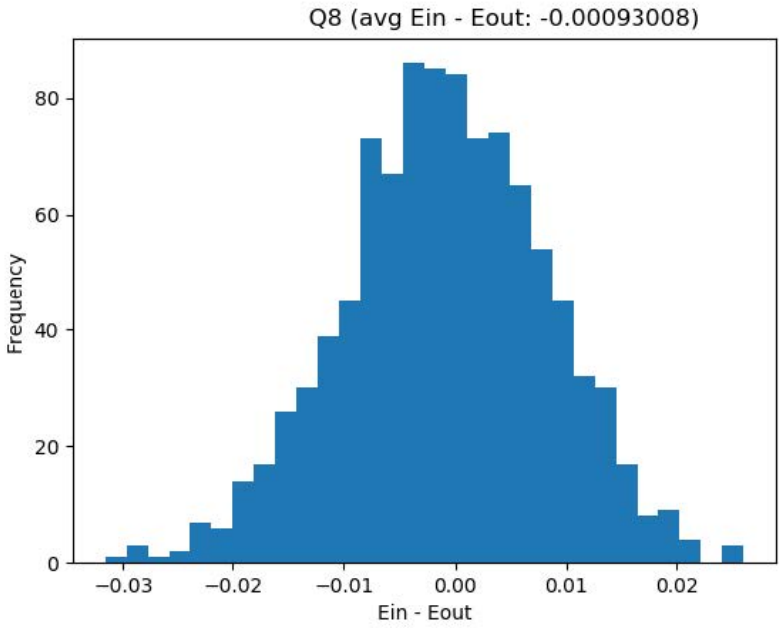

As we can see, a data size of 20 produces results mostly centered around $-0.1$ with small outliers beyond $-0.3$ and 0.1. The average error is $-0.083$. Since the data size is smaller, the results are more volatile than the graph in the next question.

**Problem 8** (20 points)**.**

***Sol.*** The following environment was used:

- The experiment was run 1000 times using seeds from 1234 to 2235.
- The histogram was plotted using 30 bins.

The histogram is as follows:




We can see that by increasing the data size to 2000, $|E_{\text{in}} - E_{\text{out}}|$ is much more uniform than the previous question, decreasing the average distance to $-0.0009$, with most of the values being centered around $-0.01$ and $0.01$.

□

**Problem 9** (20 points)**.**

***Sol.*** We can split the problem into two parts: showing that $2^d$ points can be shattered, and then showing that $2^d + 1$ points cannot be.

**$2^d$ points can be shattered**

First, observe that $v$ represents a series of choices determined by the input $x$ as it goes through the binary decision tree based on the threshold vector $t$. Specifically, at the $i$th layer of the tree, the input $x$ decides to enter the left child (if $x_i \leq t_i$) or the right child (if $x_i > t_i$). The choice made at the $i$th layer is recorded as $v_i = 0$ or 1 respectively. This continues until after the $d$th layer has been processed.

From this perspective, we can see that an input $x$ can result in up to $2^d$ possible paths (i.e. at each layer, you can either go left or right). Specifically, for any possible input $x$, it is possible to construct $2^d$ threshold vectors $t$: for each dimension $x_i$, the value of $t_i$ is either larger or no larger than $x_i$, done for all $d$ dimensions.

This means that, for $N = 2^d$ inputs, it is possible to create all possible $v$ vectors by assigning each $x_i$ to its own path.

For example, if $d = 2$, then $N = 2^d = 4$ can enumerate the following outcomes: $x_1$ to $v_1 = [0, 0]$, $x_2$ to $v_2 = [0, 1]$, $x_3$ to $v_3 = [1, 0]$, and $x_4$ to $v_4 = [1, 1]$.

We can collectively say that there exists a $t$ such that for $2^d$ inputs, $x_i$ corresponds to $v_i$, which is a unique path. If there are **more** than $2^d$ inputs, some of them will correspond to the same $v_i$ by pigeonhole principle.

Next, since $S \subseteq \{0, 1\}^d$, we can effectively create $2^{2^d}$ possible $S$ sets. Using the $d = 2$ example, some possible sets $S$ may be $\{[0, 1], [1, 0]\}$, $\emptyset$, $\{[0, 0], [1, 1]\}$ and so on. This implies that there are $2^{2^d}$ dichotomies, and if $v_i \in$ some set $S$, then $h(x) = +1$ and $h(x) = -1$ otherwise.

Using $N = 2^d$ inputs, there are thus exactly $2^{2^d}$ dichotomies that can be created, as each of the possible $2^d$ paths $v$ are possible for some $t$, and $\mathcal{H}$ can effectively create a powerset of all possible $v$s (of which members are called $S$).

For example, let $S_1$, $S_2$, ..., $S_{16}$ be all possible subsets of $\{0, 1\}^2$. Then

$h_{S_1=\emptyset,t}(x_1, x_2, x_3, x_4) = (-1, -1, -1, -1)$;

$h_{S_2=\{[0,0]\},t}(x_1, x_2, x_3, x_4) = (+1, -1, -1, -1)$;

...

$h_{S_{16}=\{[0,0],[0,1].[1,0],[1,1]\},t}(x_1, x_2, x_3, x_4) = (+1, +1, +1, +1)$.

Which effectively shatters the $N = 2^d$ points.

Note that with less than $2^d$ points, they can all be shattered as some set of $N < 2^d$ points can correspond to $N$ (which are some, but not all) possible paths $v$; the powerset that can be derived from all possible $v$s has a cardinality of $2^N$.

**$2^d + 1$ points cannot be shattered**

Next, we show that since $2^d + 1$ points cannot be shattered,

Assume that the first $2^d$ points each has a corresponding path $v$. By the pigeonhole principle, the next point $x'$ must also correspond to an existing path (since there are only $2^d$ paths). That is, some point $x$ from the first $2^d$ points

will correspond to the same $v$. We can then see that $h(x) = h(x')$, meaning it is impossible to shatter the set with more than $2^d$ points.

For example, if $d = 2$, then $N = 4$ points can be shattered as shown above. However, if $N = 5$, then $x_5$ and some other point (say, $x_1$) will correspond to the same path. Then, this will happen:

$h_{S_2=\{[0,0]\},t}(\mathbf{x_1}, x_2, x_3, x_4, \mathbf{x_5}) = (+\mathbf{1}, -1, -1, -1, +\mathbf{1})$;

It is impossible that $(+\mathbf{1}, -1, -1, -1, -\mathbf{1})$ can occur; thus this set of points cannot be shattered.

With these two parts in mind, we can conclude that the VC dimension of $\mathcal{H}$ is $2^d$.

□