Bharadwaj Chukkala 118341705 bchukkal@umd.edu

ENPM808A: Introduction to Machine Learning

Homework 2

Problem 1 (2.1 LFD)

Check the code for the problem

```
The Examples we need to make Er <= 0.05 for M = 1:     N = 839.9410155759854
The Examples we need to make Er <= 0.05 for M = 100:   N = 1760.9750527736035
The Examples we need to make Er <= 0.05 for M = 10000: N = 2682.009089971222
```

Problem 2 (2.2 LFD)

For $N=4$, if we consider 4 non-aligned points. this $\mathcal{H}$ shatters these points (we can effectively enumerate them to see that all dichotomies are generated, so in this case we have $m_{\mathcal{H}}(4) = 2^4$

However for $N=5$, no matter how you place your five points, some points will be inside a rectangle defined by others, In this case we are not able to generate all dichotomies, consequently $m_{\mathcal{H}}(5) < 2^5$.

$\Rightarrow$ From these two observations, we may effectively conclude that for positive rectangles, we have $d_{vc} = 4$

thus $\underline{m_{\mathcal{H}}(N) \leq N^4 + 1}$

Problem 3 (2.11 LFD)

Check the code for the problem

```
For N = 100   the VC Bound = 0.8481596247015304
For N = 10000 the VC Bound = 0.10427815497178729
```

Problem 4 (2.12 LFD)

Check the code for the problem

$$N \geq \frac{8}{\epsilon^2} \ln\left(\frac{4((2N)^{d_{vc}}+1)}{\delta}\right)$$

Trying an initial guess of N = 1, 000 in the RHS and then use that as a new guess until N converges to a value.

```
The Sample Size Converges at N = 452956
```

Problem 5 (2.23 LFD)

Check the code for outputs and graphs

1 $\Rightarrow$ find best hypothesis that approximates mean squared error :-
best hypothesis will minimize least sq. error.

i) $h(x) = ax + b$.

$$E_x[(f(x) - h(x))^2] = E_x[(\sin \pi x - (ax+b))^2]$$

Now to minimize this, we shall equate the first derivative to zero.

writing $E_x[(\sin \pi x - (ax+b))^2]$ in a form that can be minimized.

$$= \int_{-1}^{1} [\sin \pi x - (ax+b)]^2 p(x)\, dx.$$

$$= \int_{-1}^{1} \frac{1}{2} [\sin \pi x - (ax+b)]^2 dx$$

take derivative w.r.t a and b and equate to 0.

$$\frac{\partial E}{\partial a} = \int_{-1}^{1} x[\sin \pi x - (ax+b)]\, dx = 0$$

$$\Rightarrow \frac{2}{\pi} - \frac{2}{3}a = 0$$

$$\frac{\partial E}{\partial b} = \int_{-1}^{1} [\sin \pi x - (ax+b)]\, dx = 0$$

$$\Rightarrow b = 0$$

$\therefore$ we have $a = \frac{3}{\pi}$ //

ii) when $h(x) = ax$

take derivative w.r.t a

$$\frac{\partial E}{\partial a} = \int_{-1}^{1} x\,[\sin\pi x - (ax)]\,dx = 0.$$

$$\Rightarrow \frac{2}{\pi} - \frac{2}{3}a = 0$$

$\therefore$ we have $a = \frac{3}{\pi}$ //

(iii) when $h(x) = b$

$$\frac{\partial E}{\partial b} = \int_{-1}^{1} [\sin\pi x - b]\,dx = 0$$

$$\Rightarrow b = 0$$

$\therefore$ we have $b = 0.$

(a) Expected values w.r.t D

$$E_{in}(g) = \sum_{i=1}^{N} (f(x_i) - h(x_i))^2$$

$$= \sum_{i=1}^{N} (\sin(\pi x_i) - (ax_i + b))^2$$

$$\frac{\partial E_{in}(g)}{\partial a} = -2\sum_{i=1}^{N} x_i\,(\sin(\pi x_i) - (ax_i + b)) = 0$$

$$\frac{\partial E_{in}(g)}{\partial b} = -2\sum_{i=1}^{N} (\sin(\pi x_i) - (ax_i + b)) = 0$$

To solve for a and b

$$(x_2 - x_1)(\sin\pi x_2 - ax_2 - b) = 0$$
$$(x_1 - x_2)(\sin\pi x_1 - ax_1 - b) = 0$$

if we assume $x_1 \neq x_2$ we can solve

$$a = \frac{\sin\pi x_2 - \sin\pi x_1}{x_2 - x_1}$$

$$b = \frac{x_2\sin\pi x_1 - x_1\sin\pi x_2}{x_2 - x_1}$$

when $h(x) = ax+b$

$$g^D(x) = \frac{\sin\pi x_2 - \sin\pi x_1}{x_2 - x_1} x + \frac{x_2 \sin\pi x_1 - x_1 \sin\pi x_2}{x_2 - x_1}$$

similiarly we can get a, b value for $h(x) = ax$ and $h(x) = b$.

when $h(x) = ax$

$$g^D(x) = \frac{x_2 \sin\pi x_1 + x_1 \sin\pi x_2}{x_1^2 + x_2^2} x$$

when $h(x) = b$

$$g^D(x) = \frac{\sin\pi x_2 + \sin\pi x_1}{2}$$

here we can see that x is uniformly distributed between -1 and 1 only for 1 hypothesis, $h(x) = b$.

$$E_D[g^D(x)] = E_D\left[\frac{\sin\pi x_1 + \sin\pi x_2}{2}\right] = 0$$

Calculation of Bias component

$$\begin{aligned} \underline{\text{bias}} &= E_x((\bar{g}(x) - f(x))^2 \\ &= E_x[(0 - \sin(\pi x))^2] \\ &= E_x(\sin(\pi x))^2 \\ &= 1/2 \end{aligned}$$

Calculation of Variance component

$$\begin{aligned} \underline{\text{variance}} &= E_x[E_D[(g^D(x) - \bar{g}(x))^2]] \\ &= E_x\left[E_D\left(\frac{1}{2}(\sin\pi x_1 + \sin\pi x_2) - 0\right)^2\right] \\ &= E_x\left[\frac{1}{4} E_D(\sin\pi x_1 + \sin\pi x_2)^2\right] \\ &= E_x[E_D(\sin^2\pi x_1 + \sin^2\pi x_2 + 2\sin\pi x_1 \sin\pi x_2)] \\ &= E_x\left[\frac{1}{4}\left(\frac{1}{2} + \frac{1}{2} + 0\right)\right] \\ &= \underline{0.25} \end{aligned}$$

Compute the out-of sample error :-

$$\begin{aligned} E_D[E_{out}(g^D)] &= \text{bias} + \text{variance} \\ &= \frac{1}{2} + \frac{1}{4} \\ &= \underline{0.75} \end{aligned}$$

Problem 6

To show that k is a break point for H

- ☐ Show a set of k points $x_1, \ldots x_k$ which H can shatter
- ☐ Show H can shatter any set of k points.
- ☐ Show a set of k points $x_1, \ldots, x_k$ which H cannot shatter
- ☐ Show H cannot shatter any set of k points $x_1, \ldots x_k$.

To understand these statements, we need to understand what a break point k and what is meant by shattering.

→ Shattering is the ability of a model to classify a set of p ts perfectly.

A set of S examples is shattered by a set of functions H, if for every dichotomy of S into positive & negative points, there is a function h(x) in H that gives the labels perfectly to the dichotomy.

→ Break point :- If no data set of size k can be shattered by $H = \{h(x_1), \ldots h(x_k)\}$, then k is the break point for H. Basically at a break point, the hypothesis set fails to achieve all dichotomies.

1) When we show that set of k points which H can shatter, we prove that k might not be a break point for H. I didn't choose this because this has is not guaranteed proof to show

2) When we show that H can shatter any set of k points, we prove with gurantee that k is definitely not a break point. I didn't choose this because this is converse of what is asked to show

3) When we show a set of k points which H cannot shatter, we prove that k might be a break point for H. I didn't chose this because this again is not guaranteed proof to show

4) When we show that H cannot shatter any set of k points, we can guarantee that k is a break point from the definition of break point.
Hence i chose this.

Problem 7

The VC dimension will be provided in 2 steps

1) There exist d+1 points that perceptron can shatter
2) No d+2 (or more) points can be shattered by H.

1) Suppose the Perceptron f(x)

$$f(x) = \begin{cases} 1 & \text{if } w^T x + b > 0 \\ -1 & \text{otherwise} \end{cases}$$

consider d+1 points

$$x^{(0)} = (0 \ldots 0)^T$$
$$x^{(1)} = (1, 0 \ldots 0)^T \qquad y = (y_0, y_1 \ldots y_d)^T \in \{-1, 1\}^{d+1}$$
$$x^{(2)} = (0, 1, \ldots 0)^T$$
$$\vdots$$
$$x^{(d)} = (0, \ldots d)^T$$

Let $b = 0.5$

y_0 and $w = (w_1, w_2 \ldots w_d)$ where $w_i = y_i$

$i \in \{1, 2 \ldots d\}$

thus f(x) can label all these points correctly.

so VC dimension of perceptron is atleast d+1.

2) Expand $x \in R^d$ to $x \in R^{d+1}$, by letting $(X)^T = (x^T, 1)$

and let $W^T = (w^T, b)$

Thus

$$f(X) = \begin{cases} 1 & \text{if } W^T X > 0 \\ -1 & \text{otherwise} \end{cases}$$

Assume that there are d+2 points that perceptron in R^d can shatter, namely $x^{(1)}, x^{(2)} \ldots x^{(d+2)} \in R^d$ corresponding to $X^{(1)}, X^{(2)}, \ldots X^{(d+2)} \in R^{d+1}$

since d+2 points in R^{d+1}, there exists certain i

such that $X^{(i)} = \sum_{j \neq i} a_j \cdot X(j)$

where at least one $a_j \neq 0$. Let $S = \{ j / j \neq i, a_j \neq 0 \}$

$\forall j \in S$, we give $x^{(j)}$ the label $\text{sign}(a_j)$ and give $x(i)$ a label -1.

By our assumption, there exists W that make $f(x)$ label those $d+2$ points correctly so $\forall j \in S$.

we have $a_j \cdot W^T x^{(j)} > 0$

and $W^T x^{(i)} \leq 0$

Also :

$$\begin{aligned} W^T x^{(i)} &= W^T \Big(\sum_{j \neq i} a_j \cdot x^{(j)} \Big) \\ &= W^T \Big(\sum_{j \in S} a_j \cdot x^{(j)} \Big) \\ &= \sum_{j \in S} a_j \cdot W^T x^{(j)} \\ &> 0 \end{aligned}$$

So our assumption is false. The VC dimension of Perceptron in $\mathbb{R}^d$ is at most $d+1$.

from 1 and 2, we conclude that the VC dimension of Perceptron in $\mathbb{R}^d$ is $d+1$.

Problem 8 (Exercise 2.6 LFD)

Check code for output

a) Apply the error bar in $E_{out}(g) \leq E_{in}(g) + \sqrt{\frac{1}{2N} \ln \frac{2M}{\delta}}$.

```
The Ein for the Training set is: 0.11509037065006825
The Ein for the Test set is:     0.09603227913199208
```

So, the error bar on the in-sample error is higher than the error bar from the test error.

b) If we reserve more examples for testing, we'll have fewer training samples. We may end up with a hypothesis that is not as good as we could have arrived at if using more training samples. So Etest(g) might be way too off, even when the error bar on it is small.