The VC-Dimension

Problem A

a The VC dimension is the measure of the capacity of a hypothesis class. It describes the largest set of points $\mathcal{F}$ can shatter meaning it can correctly classify all the possible labellings of the set.

$\mathcal{F}$ has VC dimension at least 3 if there exist three points in the input space such that for every possible combination of labels $((y_1, y_2, y_3))$ there exist a function $f \in \mathcal{F}$ that correctly labels all of them.

input space $= 3 \Rightarrow 2^3 = 8$ possible labellings [binary classification]
VC dimension $= 3 \Rightarrow f$ is strong enough to correctly label any of the possible ~~tas~~ input

b $VC(\mathcal{F}_1) \leq VC(\mathcal{F}_2)$ if $\mathcal{F}_1 \subset \mathcal{F}_2$

$VC(\mathcal{F}_1)$ is the ~~number~~ largest set shattered by $\mathcal{F}_1$
$VC(\mathcal{F}_2)$ is the largest set shattered by $\mathcal{F}_2$
if $\mathcal{F}_1 \subset \mathcal{F}_2$ any set shattered by $\mathcal{F}_2$ is already shattered by $\mathcal{F}_1$ since $\mathcal{F}_1$ is a subset of $\mathcal{F}_2$

hence $VC(\mathcal{F}_1) \leq VC(\mathcal{F}_2)$

Problem B

a $\mathcal{X} = \{1, 2, 3\}$

$\mathcal{F} = \{f: \mathcal{X} \to \{-1, 1\} ; f(1) = 1\}$

$f(1) = 1$ $f(2)$ and $f(3)$ can be 1 or -1

$f_1 : f(1) = 1 \quad f(2) = 1 \quad f(3) = 1$

$f_2 : f(1) = 1 \quad f(2) = 1 \quad f(3) = -1$

$f_3 : f(1) = 1 \quad f(2) = -1 \quad f(3) = 1$

$f(4) : f(1) = 1 \quad f(2) = -1 \quad f(3) = -1$

So there are 4 functions in $\mathcal{F}$.

b VC dim is 2

c input: $\{1, 2\}$ Label: $(1, 1) \to$ Use f_1 or f_2
$(1, -1) \to$ Use f_3 or f_4

$\{1, 3\}$ Label: $(1, 1) \to$ Use f_1 or f_3
$(1, -1) \to$ Use f_2 or f_4

$\{2, 3\}$ Labels: $(1, 1) \to$ Use f_1
$(1, -1) \to$ Use f_2
$(-1, 1) \to$ Use f_3
$(-1, -1) \to$ Use f_4

input 1 cannot have label -1 since $f(1) = 1$ ~~we can't have 3 in~~

Largest subset shattered by $\mathcal{F}$ is $\{2, 3\}$, $2^2 = 4$

If we take 3 $(-1, \ , \ )$ cannot be shattered

$\therefore$ VC dimension is 2

Problem C

$\mathcal{F} = \{ \cancel{f} x \to \text{sign}(\max\{0, wx+b\}) : w, b \in \mathbb{R} \}$

So in simple words we can define the function as

1 if $wx + b > 0$ lets say f_1
-1 otherwise lets say f_2

because ~~of~~ This ~~is~~ is equivalent to a single linear threshold like a perceptron with an additional constrain The key difference from standard linear classifiers is that the decision boundary is strict

a On 2 points the possible labellings are

$(1, 1)$ use f_1 for both
$(1, -1)$ f_1 for first and f_2 for second
$(-1, 1)$ f_2 for first and f_1 for second
$(-1, -1)$ f_2 for both

On 3 points the $\mathcal{F}$ cannot label all combinations correctly mainly when the label gets flipped after the threshold

Eg: $(1, -1, 1)$ $(-1, 1, -1)$

b Since it cannot correctly label all the combinations for 3 points but can correctly do for 2 points the VC Dimension is 2

Neural Network Approximation

Problem A

1 The Weirstrass Approximation theorem states that any continuous ~~v~~ real valued function $f(x)$ defined on a closed interval $[a, b]$ can be uniformly approximated by a polynomial to any desired accuracy.

ie For any $\varepsilon > 0$ there exist polynomial $P(x)$ such that

$$\sup_{x \in [a,b]} |f(x) - P(x)| < \varepsilon$$

2 $f(x) = \sin(x)$ $\quad I = [-\pi, \pi]$

To approximate $\sin(x)$ within an error 10^{-9} we can use its taylor series expansion

$$\sin x = \sum_{n=0}^{\infty} \frac{(-1)^n}{(2n+1)!} x^{2n+1}$$

We need to find smallest n value such that

$$\left| \sin x - \sum_{n=0}^{\infty} \frac{(-1)^n}{(2n+1)!} x^{2n+1} \right| < 10^{-9}$$

The maximum error occurs at $x = \pi$

$$\frac{\pi^{2n+1}}{2n+1} < 10^{-9}$$

Trying different values for N, at $N=9$ the error is $\approx 10^{-8}$ and for $N=10$ the error is $\approx 10^{-10}$. Therefore we choose $N=9$ such that

$$\sup_{x\in[a,b]} |\sin(x) - P(x)| < 10^{-9}$$

Problem B

Given a 2π periodic function $f(x)$, the Fourier Series representation is

$$f(x) \sim a_0 + \sum_{n=1}^{\infty} a_n \cos(nx) + b_n \sin(nx)$$

The Fourier coefficients are given by

Constant term : $a_0 = \frac{1}{2\pi}\int_{-\pi}^{\pi} f(x)\,dx$

cosine ~~term~~ coeff : $a_n = \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)\cos(nx)\,dx$

sine coeff : $b_n = \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)\sin(nx)\,dx$

2 $f(x) = e^{-|x|}$ is an even function since

$$f(-x) = e^{-|-x|}$$
$$= e^{-|x|}$$
$$= f(x)$$

Thus all sine terms vanishes $b_n = 0$

$$\therefore \quad f(x) \sim a_0 + \sum_{n=1}^{\infty} a_n \cos(nx)$$

$$a_0 = \frac{1}{2\pi}\int_{-\pi}^{\pi} e^{-|x|}\,dx$$

$$= \frac{1}{2\pi} \times 2 \int_0^{\pi} e^{-x}\,dx$$

$$= \frac{1}{\pi}\int_0^{\pi} e^{-x}\,dx$$

$$= \frac{1-e^{-\pi}}{\pi}$$

$$a_n = \frac{1}{\pi}\int_{-\pi}^{\pi} e^{-x}\cos nx\,dx$$

$$\frac{\pi a_n}{2} = \int_0^{\pi} e^{-x}\cos nx\,dx$$

$$\frac{\pi a_n}{2} = \left[-e^{-x}\cos nx - \int n e^{-x}\sin nx\,dx\right]$$

$$= -e^{-x}\cos nx - \left[-e^{-x} n\sin(nx) - \int n^2 \cos nx\, e^{-x}\,dx\right]$$

$$= -e^{-x}\cos nx + e^{-x} n\sin(nx) - n^2\int e^{-x}\cos nx\,dx$$

$$= -e^{-x}\cos nx + e^{-x} n\sin(nx) - n^2\frac{\pi a_n}{2}$$

$$\frac{\pi a_n}{2} + \frac{n^2\pi a_n}{2} = -e^{-x}\cos(nx) + e^{-x} n\sin(nx)$$

$$a_n\frac{\pi}{2}(n^2+1) = -e^{-x}\cos(nx) + e^{-x} n\sin(nx)$$

$$\therefore \quad a_n = \left[\frac{\pi}{2(n^2+1)} e^{-x}\left(n\sin(nx) - \cos(nx)\right)\right]_0^{\pi}$$

$$\therefore a_n = \frac{2}{\pi}\left[\frac{e^{-\pi}(-\cos(n\pi)) + 1}{n^2+1}\right]$$

$$= \frac{2}{\pi}\left[\frac{1-(-1)^n e^{-\pi}}{n^2+1}\right]$$

$$\cos n\pi = \begin{cases} 1 & \text{if } n \text{ is even} \\ -1 & \text{if } n \text{ is odd} \end{cases}$$

$$\therefore e^{-|x|} \sim \frac{1-e^{-\pi}}{\pi} + \sum_{n=1}^{\infty} \frac{2\left[1-(-1)^n e^{-\pi}\right]}{\pi(n^2+1)} \cos(nx)$$

Problem C

A neural network with a single hidden layer and sufficient width can ~~be~~ approximate any continuous function on a domain arbitrarily well given an appropriate activation function is used. This is universal approximation theorem

Shallow networks can approximate any function but may require ~~to~~ extremely high number of neurons for approximating complex function

A deep network is more efficient than a shallow network but the process of training is complex due to optimization challenges.

Comparitive analysis and discussion

From the plots we can confidently say that deep neural network performs much better than the shallow network. From the plots of actual function against model approximation we can observe that deep networks approximate the function more smoothly. Though the shallow network is also able to approximate the function reasonably well it may require more number of neurons for smoothness.

In the light of training shallow network trains faster since gradient calculations and updations happen only for a single layer. But for deeper network the same happens for multiple layers and also there are optimization challenges. Hence deep networs require more time for training.

Deep neural networs are beneficial than shallow network in approximating complex functions or problems. For example when it comes to data with high dimensionality or complex/discontinuous/ ~~or~~ non-smooth functions deep networks perform better.

If you observe the plots for loss for deep networks, at some points you can see that the loss is spiking. This is probable because the model is overfitting to the data. In simple words the model has already reached the optimum parameters but now its overly fitting the data or any noise. This could be either because the network is deeper than it actually need (overqualified) or it is running ~~for~~ more iterations/epochs than it actually require.

Neural Networks Optimization

Problem A

b Sauared loss is a metric used to evaluate a model
In simple words it is the aggregate sum of ~~or~~ squares of difference between the actual and the predicted result

Hence
$$L(\theta) = \sum_{i=1}^{n} [f_\theta(x_i) - y_i]^2$$

$f_\theta(x_i) - y_i$ determines how different the predicted output is from the actual result.

but if $f_\theta(x_i) < y_i$ then $f_\theta(x_i) - y_i < 0$
So we square it to get a positive value since error is supposed to added up always.

c
$$f_\theta(x) = \alpha_1 \sigma(w_1 x + b_1) + \alpha_2 \sigma(w_2 x + b_2)$$

$$\theta = (\alpha_1, \alpha_2, w_1, w_2, b_1, b_2)$$

$$L(\theta) = \sum_{i=1}^{3} [f_\theta(x_i) - y_i]^2$$

$$\frac{\partial L}{\partial \alpha_1} = \sum_{i=1}^{3} \frac{\partial}{\partial \alpha_1} [f_\theta(x_i) - y_i]^2$$

$$= \sum_{i=1}^{3} 2(f_\theta(x_i) - y_i) \frac{\partial}{\partial \alpha_1} (f_\theta(x_i) - \cancel{y_i}^{\,0})$$

$$= \sum_{i=1}^{3} 2(f_\theta(x_i) - y_i) \frac{\partial}{\partial \alpha_1} (\alpha_1 \sigma(w_1 x + b_1) + \cancel{\alpha_2 \sigma(w_2 x + b_2)}^{\,0})$$

$$= \sum_{i=1}^{3} 2\left(f_\theta(x_i) - y_i\right) \cdot \sigma(w_1 x_i + b_1)$$

$$= 2 \sum_{i=1}^{3} \left(f_\theta(x_i) - y_i\right) \cdot \sigma(w_1 x_i + b_1)$$

Similarly $\dfrac{\partial l}{\partial \alpha_2} = 2 \sum_{i=1}^{3} (f_\theta(x_i) - y_i) \cdot \sigma(w_2 x_i + b_2)$

$$\frac{\partial l}{\partial w_1} = \sum_{i=1}^{3} 2\left(f_\theta(x_i) - (y_i)\right) \cdot \frac{\partial}{\partial w_1}\left(\alpha_1 \sigma(w_1 x_i + b_1) + \alpha_2 \sigma(w_2 x_i + b_2)\right)$$

$$= 2 \sum_{i=1}^{3} \left(f_\theta(x_i) - y_i\right) \cdot \alpha_1 x_i \cdot 1(\underbrace{w_1 x_i + b_1}_{z} > 0)$$

because $\dfrac{\partial \sigma(z)}{\partial z} = \begin{cases} 1 & \text{if } z > 0 \\ 0 & \text{else} \end{cases}$

$$\frac{\partial \sigma(w_1 x + b)}{\partial w_1} = 1 . x_i \, (w_1 x_i + b > 0)$$

Similarly $\dfrac{\partial l}{\partial w_2} = 2 \sum_{i=1}^{3} \left(f_\theta(x_i) - y_i\right) \alpha_2 x_i \cdot 1 \, (w_2 x_i + b_2 > 0)$

$$\frac{\partial l}{\partial b_1} = 2 \sum_{i=1}^{3} \left(f_\theta(x_i) - y_i\right) \frac{\partial}{\partial b_1}\left(\alpha_1 \sigma(w_1 x_i + b_1) + \alpha_2 \sigma(w_2 x_i + b_2)\right)$$

$$= 2 \sum_{i=1}^{3} \left(f_\theta(x_i) - y_i\right) \alpha_1 . 1(w_1 x_i + b_1 > 0)$$

$$\frac{\partial l}{\partial b_2} = 2 \sum_{i=1}^{3} \left(f_\theta(x_i) - y_i\right) \alpha_2 . 1 \, (\underbrace{w_1 x_i + b_2}_{z} > 0)$$

Foundation of Data Science

1 Data is the foundation of any kind of science, study or information. The patterns, relevant informations obtained from those data has helped us to understand any kind of fact we currently know. This is possible only because of the concepts in data science. It helps us to understand, work with data and retrieve relevant information out of the raw data.

2 From the course I get an impression like probability and regression concepts in statistical learning are like the central components of data science. Because whenever we go deep into the machine learning models we ultimately understand this is all built upon some basic ideas supplemented with the most modern computational capabilities. I found probability and regression concepts as these basic components on which the modern Machine learning is built upon.

3 The most valuable take away for me from this course is that me being a data science student now knows the basic mathematical concepts behind the machine learning models

4 The courses introduction to data science and foundation to data science takes completely two different paths. Foundation being a second cycle of the course was expected to be like a continuation but felt like a completely different thing. It would be really helping if both the courses are structured in a way to supplement each other.

5 The teaching format and structure of the course is just right

6 6 / 7 / 8

Being a student from non math background I have had challenges understanding the lecture notes easily. It took me a while to understand each thing. Because of the same reasons the home assignments were a bit challenging for me. Mainly because statistics is a vast subject and me someone new to statistics. But this has indeed helped me to read more about each problem in the assignments.

AI tools like ChatGPT and Deepseek have helped me in achieving good understanding of concepts and also breaking down problems and making them easier to understand.

Obviously uncontrolled use of ~~ed~~ AI tools may lead to underutilized brains but its ultimately dependent on the quality of the user. A driven user will use the tools to supplement the knowledge.