Problem 2.1 : In Equation (2.1), set $\delta = 0.03$, and let $\epsilon(M, N, \delta)$

$$= \sqrt{\frac{1}{2N} \ln \frac{2M}{\delta}}$$

a) For $M = 1$, how many examples do we need to make $\epsilon \leq 0.05$?

→ $$\epsilon = \sqrt{\frac{1}{2N} \ln \frac{2M}{\delta}}$$

$$\Rightarrow N = \frac{1}{2\epsilon^2} \ln \frac{2M}{\delta} \Rightarrow N = \frac{1}{2 \times (0.05)^2} \ln \frac{2 \times 1}{0.03}$$

$$\Rightarrow N = \frac{1}{0.005} \ln (66.67) = 839.94$$

$\therefore$ for $\epsilon \leq 0.05$, we have $\underline{N \geq 839.94}$

b) For $M = 100$, how many examples do we need to make $\epsilon \leq 0.05$

→ $$N = \frac{1}{2 \times (.05)^2} \ln \frac{2 \times 100}{.03} = \frac{1}{.005} \ln (6666.67) = 1760.97$$

$\therefore$ for $\epsilon \leq 0.05$, we have $\underline{N \geq 1760.97}$

c) For $M = 10{,}000$, how many examples do we need to make $\epsilon \leq 0.05$

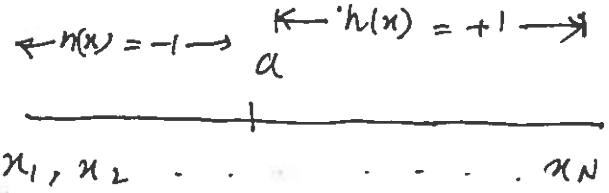
→ $$N = \frac{1}{2 \times (.05)^2} \ln \frac{2 \times 10{,}000}{.03} = \frac{1}{.005} \times \ln (666\,666.67) = 2682.009$$

$\therefore$ for $\epsilon \leq 0.05$, we have $\underline{N \geq 2682.009}$

Problem 2.3 :- Compute the max^m no. of dichotomies, $m_H(N)$, for these learning models, and consequently compute d_{VC}, the VC dimension.

a) Positive or Negative Ray : $\mathcal{H}$ contains the functions which are $+1$ on $[a, \infty)$ (for some a) together with those that are $+1$ on $(-\infty, a)$ (for some a).

→ Positive Rays



For these N points, growth func^n $m_H(N) = N+1$ (N−1 segments and for negative rays we have N−1. +1 on each end)

So, total dichotomies for 2N.

$$m_H(N) = 2N$$

Largest value of N for which $m_H(N) = 2^N$ is 2

$$d_{VC} = 2$$

For positive Rays $d_{VC} = 1$ & for negative Rays $d_{VC} = 1$

as $m_H(N) = N+1$, break point = 2 $\quad k+1 < 2$, $2+1 < 2^2$ ✓

(b) Positive or Negative Intervals :— For this $m_H(N) = \frac{N^2}{2} + \frac{N}{2} + 1$

Break point = 3 as $\quad \frac{1}{2} + \frac{1}{2} + 1 < 2$ ×

$2 + 1 + 1 < 2$ ×

$\frac{9}{2} + \frac{3}{2} + 1 < 2^3$ ✓

$\therefore d_{VC} = 3$

(c) Two concentric spheres in R^d. $a \le \sqrt{x_1^2 + \dots x_d^2} \le b$.

From the equation we can figure out that varying a and b, the circles will be centered around the origin. So the growth function will be like the positive and negative Intervals.

$$\therefore m_H = \frac{N^2}{2} + \frac{N}{2} + 1$$

$$\Rightarrow d_{VC} = 3.$$

Problem 2.8 : Which of the following are possible growth functions $m_H(N)$ for some hypothesis set

1. $1+N$:- For a finite VC dimension $m_H \leq 2^N$ and bounded by $N^{d_{VC}}+1$

$m_H(N) = 1+N \leq 2^N$

This is true for $N=2$ (breakpoint), $d_{VC} = 1$

$m_H(2) = 3 \leq 2^3$ and is bounded by $N+1$

$\therefore$ This is a possible growth function . Ex - Positive Rays

2. $1+N+N\frac{(N-1)}{2} = 1+N+\frac{N^2}{2}-\frac{N}{2} = 1+\frac{N}{2}+\frac{N^2}{2}$

$m_H(N) = 1+\frac{N}{2}+\frac{N^2}{2} \leq 2^N$

This is true for $N=3$ (breakpoint), $d_{VC} = 2$

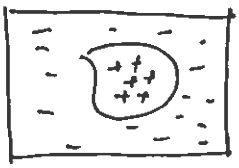
$m_H(3) = 1+\frac{3}{2}+\frac{9}{2} = 7 \leq 8$ and is bounded by N^2+1

$\therefore$ This is a possible growth function . Ex - Positive Intervals

3. 2^N :- For an infinite VC-dimension, $m_H = 2^N$

For convex sets, there are infinite break points and $m_H = 2^N$

Ex :-



$\therefore$ This is a possible growth function

4. $2^{\sqrt{N}}$:- $m_H = 2^{\sqrt{N}} \leq 2^N$

This is true for $N=2$, $d_{VC} = 1$

But, the bound $N+1$ fails for $N=25$. $\therefore$ This is not a possible growth function.

5. $2^{N/2}$:- $m_H = 2^{N/2} \leq 2^N$. This is true for $N=0$ as $1 \leq 2$

But, the bound $N^0+1=2$, fails for $N=4$. $\therefore$ This is not a possible growth function

~~Exercise~~

Problem 2.12 :- For an $\mathcal{H}$ with $d_{VC} = 10$, what sample size do you need to have 95% confidence that your generalization error is at most .05?

→ The VC bound gives the equation :-

$$E_{out}(g) \leq E_{in}(g) + \sqrt{\frac{8}{N} \ln \frac{4((2N)^{d_{VC}}+1)}{\delta}}$$

here, $\epsilon = \sqrt{\frac{8}{N} \ln \frac{4((2N)^{d_{VC}}+1)}{\delta}}$

Confidence = 95% means $\delta = 0.05$

$\epsilon = 0.05$

for error to be at most .05

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

$$\Rightarrow N \geq \frac{8}{(0.05)^2} \ln 4\left(\frac{(2\times N)^{10}+1}{0.05}\right)$$

The general Rule of thumb is $N \geq 10\, d_{VC}$

we take a random guess of $N = 1000$ initially in R.H.S

$$\Rightarrow N \geq \frac{8}{(.05)^2} \ln 4 \cdot \frac{((2000)^{10}+1)}{.05} \approx 2.57 \times 10^5$$

Iteratively $N \geq \frac{8}{(.05)^2} \ln\left(4\left(\frac{(2 \times 2.57\times10^5)^{10}+1}{.05}\right)\right) \approx 5.6\times10^5$

Iteratively $N \geq \frac{8}{(.05)^2} \ln\left(\frac{4(2\times 5.6\times10^5)^{10}+1)}{.05}\right) \approx 4.59\times10^5$

Iteratively $N \geq \frac{8}{(.05)^2} \ln\left(\frac{4\times(2\cdot 4.59\times10^5)^{10}+1)}{.05}\right) \approx 4.53\times10^5$

Iteratively $N \geq \frac{8}{(.05)^2} \ln\left(\frac{4(2\cdot 4.53\times10^5)^{10}+1)}{.05}\right) \approx 4.529\times10^5$

Iteratively, $N \geq \frac{8}{(.05)^2} \ln\left(\frac{4((2\times 4.529\times 10^5)^{10}+1)}{.05}\right) \approx 4.53\times 10^5$

$\therefore$ The estimate of $N \approx \underline{\underline{4.529\times 10^5}}$

Problem 2.22 :- When there is noise in the data, $E_{out}(g^{(D)}) = E_{x,y}[(g^{(D)}(x) - y(x))^2]$, where $y(x) = f(x) + \epsilon$. If ϵ is a zero mean noise random variable with variance σ^2, show that the bias variance decomposition becomes $E_D[E_{out}(g^{(D)})] = \sigma^2 + bias + var$.

➡ Expected value of the error over the entire space may be represented as-

$$E_D[E_{out}(g^{(D)})] = E_D[E_{x,y}[(g^{(D)}(x) - y(x))^2]]$$
$$= E_{x,y}[E_D[(g^{(D)}(x) - y(x))^2]]$$

To evaluate this, we need the average hypothesis,

$$\bar{g}(x) = E_D[g^{(D)}(x)]$$

$$\Rightarrow E_D[E_{out}(g^{(D)})] = E_{x,y}[E_D[g^{(D)}(x)^2] - 2\bar{g}(x)y(x) + y(x)^2]$$

Now, $E_D[g^{(D)}(x)^2] - 2\bar{g}(x)y(x) + y(x)^2$

$$= E_D[g^{(D)}(x)^2] - 2\bar{g}(x)(f(x)+\epsilon) + (f(x)+\epsilon)^2$$

$$= E_D[g^{(D)}(x)^2] - \bar{g}(x)^2 + \bar{g}(x)^2 - 2\bar{g}(x)f(x) - 2\bar{g}(x)\epsilon + f(x)^2 + \epsilon^2 + 2f(x)\epsilon$$

$$= \left(E_D[g^{(D)}(x)^2] - \bar{g}(x)^2\right) + \left(\bar{g}(x)^2 - 2\bar{g}(x)f(x) + f(x)^2\right) + \epsilon^2 - 2(\bar{g}(x) - f(x))\epsilon$$

$$= E_D\left[\left(g^{(D)}(x)\right)^2 - 2E_D\left[g^{(D)}(x)\right]\bar{g}(x) + \bar{g}(x)^2\right) + \left(\bar{g}(x) - f(x)\right)^2 + \epsilon^2 + -2\left(\bar{g}(x) - f(x)\right)\epsilon$$

as $\underline{E_D\left(g^{(D)}(x)\right) = \bar{g}x}$

$$= E_D\left[g^{(D)}(x) - 2g^{(D)}x\,\bar{g}(x) + \bar{g}(x)^2\right] + \left(\bar{g}(x) - f(x)\right)^2 + \epsilon^2 - 2\left(\bar{g}(x) - f(x)\right)\epsilon$$

$$= \underbrace{E_D\left[\left(g^{(D)}(x) - \bar{g}(x)\right)^2\right]}_{var(x)} + \underbrace{\left(\bar{g}(x) - f(x)\right)^2}_{bias(x)} + \epsilon^2 - 2\left(\bar{g}(x) - f(x)\right)$$

Now,

$$E_D\left[E_{out}(g)\right] = E_{x,y}\left[var(x) + bias(x) + \epsilon^2 - 2\bar{g}(x)f(x)\epsilon\right]$$

$$= E_{x,y}\left[var(x)\right] + E_{x,y}\left[bias(x)\right] + E_{x,y}\left[\epsilon^2\right] - 2E_{x,y}\left[\bar{g}(x) - f(x)\epsilon\right]$$

$$\rightarrow \boxed{E_D\left[E_{out}(g)\right) = var + bias + \sigma^2}$$

Q6: Prove that selecting the hypothesis h that maximizes the likelihood $\prod_{n=1}^{N} P(y_n|x_n)$ is equivalent to minimizing the cross-entropy error $E_{in}(w) = \frac{1}{N}\sum_{n=1}^{N} \ln\left(1 + e^{-y_n w^T x_n}\right)$

→ likelihood is the probability of getting the $y_1, y_2 \ldots y_N$ in D from the corresponding $x_1, x_2 \ldots x_N$.

$$P\left((y_1, y_2, \cdots y_N)|(x_1, x_2, \cdots x_N)\right) = \prod_{n=1}^{N} P(y_n|x_N)$$

Likelihood for one point,

$$P[y|x] = \begin{cases} h(x) & \text{for } y = +1 \\ 1 - h(x) & \text{for } y = -1 \end{cases}$$

Substitute $h(x) = \theta(w^T x)$

From the equation of logistic function $\theta(s) = \frac{e^s}{1+e^s}$

$$\therefore \theta(-s) = 1 - \theta(s)$$

where $s = \vec{w}^T \vec{x}$

$$\Rightarrow P(y|x) = \theta(y\, w^T x) \quad \text{where } y = \{-1, +1\}$$

multiplying the likelihood for all the points,

$$\rightarrow P(y_1, y_2 \cdots y_N | x_1, x_2 \cdots x_N) = \prod_{n=1}^{N} \theta(y_n w^T x_n)$$

Maximizing the likelihood,

$$\frac{1}{N} \ln \left(\prod_{n=1}^{N} \theta(y_n w^T x_n) \right)$$

$\therefore$ the algorithm is increasing, so the log of this is also maximized

multiplying by $\frac{1}{N}$ is in line with the algorithm growing proportionally.

$\therefore$ The above equation is the maximized likelihood.

To minimize, we have error measure,

$$E_{in}(w) = -\frac{1}{N} \ln \left(\prod_{n=1}^{N} \theta(y_n w^T x_n) \right)$$

$$= \frac{1}{N} \sum_{n=1}^{N} \ln \left(\frac{1}{\theta(y_n w^T x_n)} \right)$$

$\because \theta(s) = e^s / 1+e^s$

$\Rightarrow \theta(s) = 1 / 1+e^{-s}$

$$\Rightarrow \boxed{E_{in}(w) = \frac{1}{N} \sum_{n=1}^{N} \ln \left(1 + e^{-y_n w^T x_n} \right)}$$

↑ In sample error of Logistic Regression or Cross Entropy Error.

Q7. Derive the gradient of the in-sample error $\nabla E_{in}(w(t))$ used in the gradient descent algorithm.

→ The in-sample error measure for logistic regression,

$$E_{in}(w(t)) = \frac{1}{N} \sum_{n=1}^{N} \ln \left(1 + e^{-y_n w^T x_n} \right)$$

The gradient of this can be calculated as follows

$$\nabla E_{in}(w(t)) = \frac{1}{N} \sum_{n=1}^{N} \frac{1}{1+e^{-y_n w^T x_n}} \times \left(e^{-y_n w^T x_n} \right) \times (-y_n x_n)$$

as we are differentiating w.r.t $w(t)$

$$\Rightarrow \nabla E_{in}(w(t)) = \frac{1}{N} \sum_{n=1}^{N} \left(\frac{e^{-y_n w^T x_n}}{1+e^{-y_n w^T x_n}} \right) (-(y_n x_n))$$

$$= \frac{1}{N} \sum_{n=1}^{N} \theta(-y_n w^T x_n) \left(-y_n x_n \right)$$

$$= -\frac{1}{N} \sum_{n=1}^{N} y_n x_n \, \theta(-s)$$

$\because \theta(-s) = \frac{1}{1+e^{-s}}$

$$\therefore \boxed{\nabla E_{in}(w(t)) = -\frac{1}{N}\sum_{n=1}^{N} \frac{y_n x_n}{1+e^{y_n w^T x_n}}}$$

Exercise 3.13 (a) $z = \phi_2(x)$. How can we use a hyperplane $\tilde{w}$ in z to represent the following boundaries in x?

The Parabola $(x_1-3)^2 + x_2 = 1$

$$\Rightarrow x_1^2 + 9 - 6x_1 + x_2 = 1$$

$$\Rightarrow -6x_1 + x_2 + x_1^2 + 8 = 0$$

Now $\phi_2(x) = (1, x_1, x_2, x_1^2, x_1x_2, x_2^2)$

here the weights are $(-6, 1, 1, 0, 0)$ and intercept is 8. This satisfies the plane in the feature space (x_1, x_2, x_1^2). This shows that the boundary is linear in the feature space $\phi_2(x)$

(b) The circle $(x_1-3)^2 + (x_2-4)^2 = 1$

$$\Rightarrow x_1^2 + 9 - 6x_1 + x_2^2 + 16 - 8x_2 = 1$$

$$\Rightarrow -6x_1 - 8x_2 + x_1^2 + x_2^2 + 24 = 0$$

weights are $(-6, -8, 1, 0, 1)$ corresponding to $\phi_2(x)$ and the intercept is 24. This shows that a circular boundary is linear in this feature space (x_1, x_2, x_1^2, x_2^2)

(c) the ellipse $2(x_1-3)^2 + (x_2-4)^2 = 1$

$$\Rightarrow 2(x_1^2 + 9 - 6x_1) + x_2^2 + 16 - 8x_2 = 1$$

$$\Rightarrow -12x_1 - 8x_2 + 2x_1^2 + x_2^2 + 23 = 0$$

Here we see that the weights are $(-12, -8, 2, 0, 1)$ and the intercept is 23 for the feature space (x_1, x_2, x_1^2, x_2^2). This shows that the ellipse is linearly separable in $\phi_2(x)$.

Problem 3.16 :- $g(x) = P[y = +1 | x]$

Cost Matrix

You say \ True Classification	+1	−1
+1	0	c_a
−1	c_r	0

(a) $\text{Cost (accept)} = 0 \cdot P[y = +1|x] + c_a \cdot P[y = -1|x]$

$= c_a \, P[y = -1 | x]$

$= c_a \, (1 - g(x))$

$\text{cost (reject)} = c_r \cdot P[y = +1|x] + 0 \cdot P[y = -1|x]$

$= c_r \, g(x)$

(b) $\text{cost (accept)} = \text{cost (reject)}$ (As initially we consider that both events are equally likely)

$\Rightarrow c_a (1 - g(x)) = c_r (g(x))$

$\Rightarrow c_r g(x) = c_a - c_a g(x)$

$\Rightarrow g(x) = \dfrac{c_a}{c_r + c_a} = k$, where k = Threshold

(c) Supermarket example

h \ b	+1	−1
+1	0	1
−1	10	0

CIA Example

h \ b	+1	−1
+1	0	1000
−1	1	0

In the Supermarket example, we dont want any False rejects.

So $k = 1/11$. Hypothesis is rejected when $g(x) < k \approx 0$

In the CIA example, we don't want any false accepts.

So $k = 1000/1001$ Hypothesis is rejected when $g(x) > k \approx 1$.

Exercise-3.6 : -(a) Generally, if we are learning from ±1 data to predict a noisy target $P(y|x)$ with candidate hypothesis h, show that the maximum likelihood method reduces to the task of finding h that minimizes $E_{in}(w) = \sum_{n=1}^{N} [[y_n = +1]] \ln \frac{1}{h(x_n)} + [[y_n = -1]] \ln \frac{1}{1-h(x_n)}$

→ The method of maximum likelihood selects the hypothesis h which maximises the probability $P(y|x)$

$$\min \quad E_{in}(w) = \frac{1}{N} \sum_{n=1}^{N} \ln \frac{1}{P(y_n|x_n)} \quad \text{as} \quad \prod_{n=}^{N} P(y_n|x_n)$$

$$\Downarrow \text{minimize}$$

$$= \frac{-1}{N} \ln \prod_{n=1}^{N} P(y_n|x_n$$

For pointwise error

$$P(y|x) = \begin{cases} h x & \text{for } y = +1 \\ 1 - h(x) & \text{for } y = -1 \end{cases}$$

For ±1 data we have

$$E_{in}(w) = \sum_{n=1}^{N} [[y_n = +1]] \ln \frac{1}{P(y_n|x_n)} + \sum_{n=1}^{N} [[y_n = -1]] \ln \frac{1}{P(y_n|$$

$$\Rightarrow E_{in}(w) = \sum_{n=1}^{N} [[y_n = +1]] \ln \frac{1}{(h(x_n))} + [[y_n = -1]] \ln \frac{1}{(1 - h(x_n))}$$

(k) For the case $h(x) = \theta(w^T x)$, argue that minimizing the in sample error in part (a) is equivalent to minimizing the one in 3.9

→ $h(x) = \theta(w^T x)$

In sample error,

$$E_{in}(w) = \sum_{n=1}^{N} [y = \pm 1] \ln \frac{1}{h(x)} + [y = -1] \ln \frac{1}{(1-h(x))}$$

$$= \sum_{n=1}^{N} [y = \pm 1] \ln \frac{1}{\theta(w^T x)} + [y = -1] \ln \frac{1}{1-\theta(w^T x)}$$

from the question, we see that for the two probability distributions $\{p, 1-p\}$ and $\{q, 1-q\}$, the cross entropy error is $p = [y_n = +1]$ and $q = h(x)$

$$\Rightarrow \text{Cross Entropy Error} = p \log \frac{1}{q} + (1-p) \log \frac{1}{1-q} \quad \text{—(i)}$$

From Equation 3.9,

$$E_{in}(w) = \frac{1}{N} \sum_{n=1}^{N} \ln(1 + e^{-y_n w^T x_n})$$

$$= \frac{1}{N} \sum_{n=1}^{N} \ln \left(\frac{1}{\theta(y_n w^T x_n)} \right)$$

$$= [y = \pm 1] \ln \frac{1}{\theta(w^T x_n)} + [y = -1] \ln \frac{1}{1-\theta(w^T x)}$$

$$= [y = \pm 1] \ln \frac{1}{h(x)} + [y = -1] \ln \frac{1}{1-h(x)}$$

$\Leftrightarrow$ $p \log \frac{1}{q} + (1-p) \log \frac{1}{1-q}$ ∴ Minimizing the 2 equations is equivalent.