Ex 2.1

a) $$\epsilon = \sqrt{\frac{1}{2N} \ln \frac{2m}{\delta}}$$

$$\therefore \epsilon^2 = \frac{1}{2N} \ln \frac{2m}{\delta}$$

$$\therefore N = \frac{1}{2\epsilon^2} \ln \frac{2m}{\delta}$$

$\Rightarrow$ For $\epsilon \leq 0.05$

$$\Rightarrow N \geq \frac{1}{2\epsilon^2} \frac{2m}{\delta}$$

$m=1$, $\delta = 0.03$ & $\epsilon \leq 0.05$

$$\therefore N \geq \frac{1}{2\times(0.05)^2} \ln \frac{2\times 1}{0.03}$$

$$= 839.941$$

b) $m = 100$, $\delta = 0.03$ & $\epsilon \leq 0.05$

$$\therefore N \geq \frac{1}{2\times(0.05)^2} \ln \frac{2\times 100}{0.03}$$

$$\Rightarrow N \geq 1760.97$$

c) $m = 1000$, $\delta = 0.03$ & $\epsilon \leq 0.05$

$$\therefore N \geq \frac{1}{2\times 0.05^2} \ln \frac{2\times 1000}{0.03}$$

$$\therefore N \geq 2682.009$$

Problem 2.3

a) for positive rays the growth function is

$\Rightarrow N+1$.

The dichotomies for negative rays

$\Rightarrow N-1$

The dichotomies for positive rays

$\Rightarrow N+1$

Adding the two

Positive dictomies + Negative dictomies

$\Rightarrow N+1+N-1$

$\Rightarrow 2N$

$\therefore m_H(N) = 2N$.

The largest values of all shatter is given by by 2^N.

$N=1$.

$\therefore m(H)(1) = 2$ | $2^{(1)} = 2$.

$m(H)^{(2)} = 2 \times 2 = 4$ | $2^{(2)} = 4$

$m(H)^{(3)} = 2 \times 3 = 6$ | $2^{(3)} = 8$.

$\therefore$ The largest value for which $m(H)N = 2^N$ is 2

$\therefore d_{vc} = 2$

b) Growth function for positive intervals is given by $\frac{N^2}{2} + \frac{N}{2} + 1$

For negative intervals we will add only $N-2$ new dichotomies.

For $N=3$ we add $(+1, -1, +1)$

& For $N=4$ we add $(+1, -1, +1, +1)$
& $(+1, +1, -1, +1)$

$\therefore$ Total dichotomies $= \frac{N^2}{2} + \frac{N}{2} + 1 + N - 2.$

$$= \frac{N^2}{2} + \frac{3N}{2} - 1 \quad \text{for } N > 1$$

$\therefore$ Computing the largest value for which

$m_H(N) = 2^N$

we will not use $N=1$ as for 1 we already generated two dichotomies.

$\therefore N = 2.$

$$m_H(2) = \frac{2^2}{2} + \frac{2 \times 3}{2} - 1 \qquad 2^N \quad 2^2 = 4$$

$$= 4$$

$$m_H(3) = \frac{3^2}{2} + \frac{3 \times 3}{2} - 1 \qquad 2^3 = 8.$$

$$= 8$$

$$m(4) = \frac{4^2}{2} + \frac{4 \times 3}{2} - 1 = 13 \qquad 2^4 = 16.$$

$\therefore d_{VC} = 3$ (for $m_H(N) = 2^N$.

c) To find $m_H(N)$ we map R^d to $[0, +\infty]$.

$\therefore \phi : (x_1, \ldots x_d) \to r = \sqrt{x_1^2 + \ldots x_d^2}$

$\therefore$ This problem reduces to positive interval problem.

$\therefore$ we can conclude that:

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

$\therefore$ From (b) we have $d_{VC} = 3$

Problem 2.8

when $\forall\ m_H(N) = 2^N$ then $d_{VC} = +\infty$

if $m_H \neq 2^N$ then d_{VC} is a finite value

& $m(H)$ is bounded by d_{VC} using $N^{d_{VC}} + 1$

→ $\therefore m(H) = 1 + N$.

$m(1) = 2$ | $2^1 = 2$

$m(2) = 3$ | $2^2 = 4$

$\therefore d_{VC} = 1$

So it is bounded by $N^{(1)} + 1 = N + 1$ for all N.

$\therefore m(H) = N + 1$ is a possible growth function.

→ $m(H) = 1 + N + N(N-1)/2$

$m(1) = 2$ | $2^1 = 2$.

$m(2) = 4$ | $2^2 = 4$.

$m(3) = 7$ | $2^3 = 8$.

$\therefore d_{VC} = 2$.

$\therefore m(H)$ must be bounded by $N^2 + 1$ for all N.

$\therefore m(H) = 1 + N + N(N-1)/2$ is a possible growth function.

→ $m_H(N) = 2^{(\sqrt{N})}$

$m(1) = 2$ | $2^1 = 2$.

$m(2) = 2^{1.41} \approx 2$ | $2^2 = 4$

$\therefore d_{VC} = 1$

∴ It must be bounded by $N+1$

Computing for squares.

$N = 9$	$N+1$
$2^{\sqrt{9}} = 2^3 = 8$	$9+1 = 10$
$2^{\sqrt{16}} = 2^4 = 16$	$16+1 = 17$
$2^{\sqrt{25}} = 2^5 = 32$	$25+1 = 26$. ← This violates the bound condition.

Hence $m_H(N) = 2^{\sqrt{N}}$ is not a possible growth function.

→ $m_H(N) = 2^{N/2}$

$m_H(1) = 2^{1/2} \approx 1.41$ | $2^1 = 2$.

∴ $d_{VC} = 0$.

∴ It must be bounded by $N^0+1 = 1+1 = 2$. for all N.

Computing for different Ns

$N = 2$.

	d_{VC} $N+1$
$m_H(2) = 2^{2/2} = 2$	2
$m_H(4) = 2^{4/2} = 4$	2 ← This violates bound condition

∴ $m_H(N) = 2^{N/2}$ is not a possible growth function.

Problem 2.12

Sample complexity is given by.

$$N \geq \frac{8}{\epsilon^2} \ln\left(\frac{4 m_H(2N)}{\delta}\right)$$

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

$dvc = 10, \quad \epsilon = 0.05, \quad \delta = 0.05$

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

Trying an initial guess of $N = 1000$

we get

$$N \geq \frac{8}{0.05^2} \ln\left[\left(\frac{4[2\times 1000]^{10}+1}{0.05}\right)\right]$$

$$\approx 2.557251 \times 10^5$$

Now we try $N = 2.557251 \times 10^5$ in above eqn. we get $N = 4.595572 \times 10^5$. which is rapidly converging.

Problem 2.22

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

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

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

where using Fubini's theorem.

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

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

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

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

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

$$= \text{var}(x) + \text{bias}(x) + \epsilon^2 - 2(\bar{g}(x) - f(x))\epsilon$$

Taking expectation relative to (x, y)

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

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

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

$$= var + bias + var_\epsilon[\epsilon] - 2E_x[(\bar{g}(x) - f(x))E_\epsilon[\epsilon]]$$

$$= var + bias + \sigma^2 \quad (\text{since } E_\epsilon(\epsilon) = 0)$$

Quest 6 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.

Proof - The term likelihood is that we will get output y from input x given. the target distribution P(y/x) was indeed captured by hypothesis h(x)

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

using $h(x) = \theta w^T x$. we have.

$$P(y/x) = \theta(y w^T x) \rightarrow (1)$$

Now the method of maximum likelihood selects the hypothesis h which maximizes probability given by.

$$\prod_{n=1}^{N} P(y_n/x_n)$$

We can equivalently minimize it by putting

$$-\frac{1}{N} \ln \left(\prod_{n=1}^{N} P(y_n/x_n) \right)$$

$$= \frac{1}{N} \sum_{n=1}^{N} \ln \left(\frac{1}{P(y_n/x_n)} \right)$$

$\frac{1}{N}$ $\ln(\cdot)$ is a monotonically decreasing function & substituting $P(y_n/x_n)$ from eqn (1)

we get

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

Since we are trying to minize the term with respect to w allows us to call it as an error measure. Substituting

$$\frac{1}{\theta(y_n w^T x)} = 1 + e^{-y_n w^T x_n}$$

we have

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

∴ the pointwise error measure is

$$e(h(x_n), y_n) = \ln\left(1 + e^{-y_n w^T x_n}\right)$$

It has a negative exponential hence it will be small if $y_n w^T x_n$ is large and vice versa.

This implies that $\text{sign}(w^T x_n) = y_n$.

∴ Error indeed varies w to classify each x_n correctly. which is nothing but maximizing the likehood of $\prod_{n=1}^{N} P(y_n/x_n)$.

Thus the hypothesis h that maximizes the likehihood. $\prod_{n=1}^{N} P(y_n/x_n)$ is equivalent to minizing the cross entropy error.

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

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

Taking gradient we have.

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

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

$$= \frac{1}{N} \sum_{n=1}^{N} \frac{1}{1 + e^{-y_n w^T x_n}} \cdot e^{-y_n w^T x_n} \cdot -y_n x_n$$

$$= \frac{1}{N} \sum_{n=1}^{N} \frac{-y_n x_n}{e^{+y_n w^T x_n} + e^{-y_n w^T x_n} \cdot e^{y_n w^T x_n}}$$

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

Exercise 3.6

a) E_{in} is given by.

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

we have $h(x) = \theta(w^T x)$ → ①

& we have $\theta(-s) = 1 - \theta(s)$ → ②

$$\therefore E_{in}(w) = \frac{1}{N}\sum_{n=1}^{N} \ln \frac{1}{\theta(+1 \times w^T x_n)} [[y_n = +1]]$$

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

Using eqⁿ ②

~~$E_{in}(w) = \frac{1}{N}\sum_{n=1}^{N} \ln \frac{1}{1}$~~

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

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

Using eqⁿ ①

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

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

3.6

b) $h(x) = \theta(w^T x)$.

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

$$+ y_n = -1 \ln \left(\frac{1}{1 - h(x_n)} \right)$$

Putting $h(x_n) = \theta(w^T x_n)$

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

$$+ \left[y_n = -1 \ln \frac{1}{1 - \theta w^T x_n} \right]$$

We know, $1 - \theta(w^T x_n) = \theta(-w^T x_n)$

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

$$+ \left[y_n = -1 \right] \ln \left(\frac{1}{\theta(-w^T x_n)} \right).$$

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

$$\text{or } \ln \left(\frac{1}{\theta - y w^T x_n} \right)$$

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

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

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

Exercise 3.13

a) second order polynomial is given by

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

Solving the eqn of parabola.

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

$$x_1^2 - 6x_1 + 9 + x_2 = 0.$$

Equating with (1) we have weight vector.

$$\vec{w} = [\underset{w_0}{9}, \underset{w_1}{-6}, \underset{w_2}{1}, \underset{w_3}{1}, \underset{w_4}{0}, \underset{w_5}{0}]$$

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

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

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

Equating with eqn (1) we have

$$\vec{w} = [\underset{w_0}{24}, \underset{w_1}{-6}, \underset{w_2}{-8}, \underset{w_3}{1}, \underset{w_4}{0}, \underset{w_5}{1}]$$

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

$$2[x_1^2 - 6x_1 + 9] + [x_2^2 - 8x_2 + 16] = 1$$

$$2x_1^2 - 12x_1 + 18 + x_2^2 - 8x_2 + 16 = 1$$

$$33 + 2x_1^2 - 12x_1 - 8x_2 + x^2 = 0.$$

Equating with eqn (1) we have

$$\vec{w} = [\underset{w_0}{33}, \underset{w_1}{-12}, \underset{w_2}{-8}, \underset{w_3}{2}, \underset{w_4}{0}, \underset{w_5}{1}]$$

Problem 3.16

a) $g(x) = P[y = +1 / x]$

		true classification	
		+1 (correct person)	-1 (intruder)
you say	+1	0	c_a
	-1	c_r	0

$$\text{cost (accept)} = 0 \cdot P(y=+1/x) + c_a P(y=-1/x)$$

~~$= c_a t$~~

Now $g(x) = P[y = +1/x]$.

$\therefore$ For negative probability.

$1 - P = 1 - g(x)$

$\therefore$ cost accept $= c_a g(x)$.

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

$$= c_r g(x).$$

b) g(x) for accepting the person is

$$\text{cost(accept)} = \text{cost(reject)}$$

$$\Rightarrow C_a(1-g(x)) = C_r\, g(x)$$

$$\therefore g(x) = \frac{C_a}{C_a + C_r}$$

$\therefore$ our threshold will be

$$k = \frac{C_a}{C_a + C_r}$$

c) For supermarket

$C_r = 10$

$C_r = 1$

$\therefore k = \frac{1}{1+10} = \frac{1}{11}$ makes sense since we are penalizing more false reject, when $g(x) < k$ which is very close to 0 we rejecting as we want to avoid false rejects.

For CIA

$C_r = 1$

$C_a = 1000$.

$\therefore k = \frac{1000}{1000+1} = \frac{1000}{1001}$ which makes sense since are penalizing more for false accept. when $g(x) \geq k$ which very close to 1 we are accepting since we want to avoid false accept.