(1) problem 2.1 : $\epsilon(M, N, \delta) = \sqrt{\frac{1}{2N} \ln \frac{2M}{\delta}}$

given $\delta = 0.03$

(a) $M = 1; \ \epsilon \leq 0.05$

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

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

$$N \geq \frac{1}{2 \cdot (0.05)^2} \ln (66.666)$$

$$= 200 * 4.19970507$$

$$= 839.941015576$$

(b) $M = 100; \ \epsilon \leq 0.05$

$$N \geq \frac{1}{2\epsilon^2} \ln \frac{2M}{\delta} = \frac{1}{2(0.05)^2} \ln \frac{2(100)}{0.0\cancel{5}3}$$

$$= 200 \ln (6,666.66667)$$

$$= 1760.97505277736$$

(c) $M = 10000; \ \epsilon \leq 0.05$

$$N \geq \frac{1}{2\epsilon^2} \ln \frac{2M}{\delta} = \frac{1}{2\cancel{00}(0.05)^2} \ln \frac{2(1000)}{0.03}$$

$$= 200 \ln (666666.66666)$$

$$13.4100454499$$

$$= 200 \times \cancel{4}.\cancel{10774603569}$$

$$= 2682.009089712$$

(2) problem 2.3: compute maximum number of dichotomies, $m_{\mathcal{H}}(N)$ for below and compute d_{VC}.

ⓐ For positive rays growth function is equal to $N+1$. If we enumerate the dichotomies added by negative rays would result $N-1$ $(N+1+N-1)$

$$\therefore \quad m_{\mathcal{H}}(N) = 2N$$

largest value of N for which $m_{\mathcal{H}}(N) = 2^N$ is 2.

$$\therefore \quad d_{VC} = 2.$$

ⓑ For positive intervals growth function is equal to $N^2/2 + N/2 + 1$. After adding new dichotomies generated by negative intervals, we get $N-2$ and this valid for $N>1$, similarly if $N=1$ since we already generated ~~the~~ two dichotomies with positive intervals alone.

$$\therefore \quad m_{\mathcal{H}}(N) = \frac{N^2}{2} + \frac{3N}{2} - 1 \quad [\text{if } N>1 \ \& \ 2 \text{ if } N=1]$$

we can infer the largest value of N for which $m_{\mathcal{H}}(N) = 2^N$ is 3 $(m_{\mathcal{H}}(4) = 13)$, we have $d_{VC} = 3$

ⓒ For growth function of concentric circles, mapping the problem from $\mathbb{R}^d$ to $[0, +\infty]$. before to that map ϕ_x should be defined as

∴ Problem of concentric circles in R^d is similar to the problem of positive intervals in R and consequently we can infer that

$$m_{\mathcal{H}}(N) = \frac{N^2}{2} + \frac{N}{2} + 1$$

which is independent of d.

∴ Largest value of N for which $m_{\mathcal{H}}(N) = 2^N$ is 2 ($m_{\mathcal{H}}(3) = 7$), $d_{VC} = 2$.

③ problem 2.12 : we have the following implicit bound for the sample complexity N (with $d_{VC} = 10$, $\epsilon = 0.05$ and $\delta = 0.05$)

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

To find N, consider N = 1000 in the R.H.S.

$$N \geq \frac{8}{0.05^2} \ln\left(\frac{4[(2\cdot 1000)^{10}+1]}{0.005}\right) \simeq 2.57251 \times 10^5$$

∴ when we try the new value $N = 2.57251 \times 10^5$ in R.H.S and iterate the process, rapidly converging to an estimate of $N \simeq 4.52957 \times 10^5$

(4) 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

Sol/ 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

$$P(y/x) = \theta(y\, w^T x) \quad \text{--- (1)}$$

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 placing

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

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

$\frac{1}{N} \ln(\cdot)$ is a monotonically decreasing function substitute $p\left(\frac{y_n}{x_n}\right)$ from equation ①

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

∴ we're trying to minimize the term with expect to w allows to call it as an error measure. substituting

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

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

∴ Positive error measure is

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

It has a negative exponential hence it will be small iff $y_n w^T x_n$ is large & viceversa.

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

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

⑤ Derive the gradient of in-sample error $\nabla E_{in}(w(t))$ used in gradient descent algorithm.

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

Taking gradient we have

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

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

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

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

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

⑥ Exercise 3.13

ⓐ 2^{nd} order polynomial is given by

$$z(x) = (1, x_1, x_2, x_1^2, x_1 x_2, x_2^2) \quad \text{——} ①$$

solving the equation 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 as below

$$\begin{array}{cccccc} w_0, & w_1, & w_2, & w_3, & w_4, & w_5 \end{array}$$
$$\vec{w} = [9, -6, 1, 1, 0, 0]$$

(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 (1) we have weight vector as

$$\begin{array}{cccccc} w_0 & w_1 & w_2 & w_3 & w_4 & w_5 \end{array}$$
$$\vec{w} = [24, -6, -8, 1, 0, 1]$$

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

$$2[x_1^2 - 6x_1 + 9] + [x_2^2 - 8x_2 + 16] = 1$$
$$2x_1^2 - 12x_1 + 18 + x_2^2 - 8x_2 + 16 = 1$$
$$33 + 2x_1^2 - 12x_1 - 8x_2 + x_2^2 = 0$$

equating with eq (1) we have weight vector as

$$\begin{array}{cccccc} w_0 & w_1 & w_2 & w_3 & w_4 & w_5 \end{array}$$
$$\vec{w} = [33, -12, -8 \quad 2 \quad 0 \quad 1]$$

(10) problem 3.16.

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

		True classification	
		(+) correct person	-1 (intruder)
you say	+1	0	C_a
	-1	C_r	0

(a) cost (accept) $= 0 \cdot P(y = +1/x)$

$+ C_a P(y = -1/x)$

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

$\therefore$ for negative probability.

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

$\therefore$ cost accept $= C_a g(x)$

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
cost (accept) = cost (reject)

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

$\therefore g(x) = C_a / C_a + C_r$ $\therefore$ out threshold will be twice

$x = C_a / C_a + C_r$

(c) For Supermarket $C_r = 1$, $C_a = 10$.

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