Machine Learning from Data CSCI 4100
Assignment 3
Jae Park (661994900)

1. Exercise 1.13

(a)

a. y = f(x) and h $\approx$ f correctly
b. y $\neq$ f(x) and h $\approx$ f incorrectly

P[error] = P(a) + P(b) = $\mu \times \lambda + (1-\mu)(1-\lambda)$

(b)

P[error] = 0 = $\mu \times \lambda + (1-\mu)(1-\lambda)$

$= \mu\lambda + (1 - \mu - \lambda + \mu\lambda)$

$= (2\lambda - 1)\mu - \lambda + 1$

For the above equation to be independent of μ, $(2\lambda - 1) = 0$ must be true

$\therefore \lambda = 1/2$

2. Exercise 2.1

Positive rays hypothesis:

$m_H(1) = 1+1 = 2$ $m_H(2) = 2+1 = 3 < 2^2$

2 is the break point

Positive intervals hypothesis:

$m_H(1) = \frac{1}{2} + \frac{1}{2} + 1 = 2$ $m_H(2) = 2+1+1 = 4$ $m_H(3) = \frac{9}{2} + \frac{3}{2} + 1 = 7 < 2^3$

3 is the break point

Convex sets hypothesis:

$m_H(k) = 2^k = m_H(N) = 2^N$

No break point(s)

3. Exercise 2.2

(a)

i) 2 is the break point

Function $m_H(k) = k+1$ has the order of 1.

$m_H(k)=3 \leq 1 + N = 3$

So the hypothesis is bound, and the theorem holds.

ii) 3 is the breakpoint, and the function has order of 2

$m_H(k)=7 \leq \frac{1}{2}N^2 + \frac{1}{2}N + 1 = 7$

So the theorem holds.
iii) The implication of Theorem 2.4 is that if H has a break point, then there exists a polynomial bound on $m_H(N)$. However, there is no break point, so the theorem doesn't apply.

(b)
No, there is no polynomial that can bound $m_H(N) = N + 2^{[N/2]}$. For the bound to happen, there must exist a break point, but that is not the case for $m_H(N)$.

4. Exercise 2.3
i) min break point: 2
$d_{vc} = k-1 = 2-1 = 1$

ii)min break point: 3
$d_{vc} = k-1 = 3-1 = 2$

iii) min break point: none
$d_{vc} = k-1 = (\infty)-1 = \infty$

5. Exercise 2.6
(a)
Error for $E_{in}(g)$: $\sqrt{\frac{1}{2\times400} ln(\frac{2\times1000}{0.05})} = 0.115$
Error for $E_{test}(g)$: $\sqrt{\frac{1}{2\times200} ln(\frac{2\times1}{0.05})} = 0.0960$
$\therefore$ E_{in} has a higher error bar

(b) If more examples are reserved for testing, this will inevitably result in less examples for the training, which means that E_{in} will have higher error bar and g will be more likely to fail a good fitting. If g is more likely to fail, there isn't much point in using more examples to verify g.

6. Problem 1.11

$\mathrm{E_{in}(h)}$ is defined as $\frac{1}{N}\sum_{n=1}^{N}[[h(x_n) \neq f(x_n)]]$

For the case of the supermarket,

$$E_{in_{supermarket}}(h) = \frac{1}{N}\sum_{n=1}^{N} (10 \times Y[[h(x_n) =-1, f(x_n) = 1]] + 1 \times Y[[h(x_n) = 1, f(x_n) =-1])$$

For the case of CIA,

$$E_{in_{CIA}}(h) = \frac{1}{N}\sum_{n=1}^{N} (1 \times Y[[h(x_n) =-1, f(x_n) = 1]] + 1000 \times Y[[h(x_n) = 1, f(x_n) =-1])$$

, where boolean function Y is defined as
$Y[[\ condition\ ]] = \{1,\ 0 \mid \ if\ condition\ is\ true,\ Y = 1,\ otherwise\ Y = 0\}$

7. Problem 1.12

(a)

$$E_{in}(h) = \sum_{n=1}^{N} (h - y_n)^2$$

$$E_{in}(h) = \sum_{n=1}^{N} h^2 - 2hy_n + y_n{}^2 \text{ and h is a constant}$$

$$E_{in}(h) = \sum_{n=1}^{N} h^2 - 2hy_n + y_n{}^2$$

$$E_{in}(h) = Nh^2 - 2h\sum_{n=1}^{N} y_n + \sum_{n=1}^{N} y_n{}^2$$

$$= Nh^2 - 2h\sum_{n=1}^{N} y_n + [\frac{1}{N}\sum_{n=1}^{N} y_n - \frac{1}{N}\sum_{n=1}^{N} y_n] + \sum_{n=1}^{N} y_n{}^2$$

$$= N(h - \frac{1}{N}\sum_{n=1}^{N} y_n)^2 - \frac{1}{N}\sum_{n=1}^{N} y_n + \sum_{n=1}^{N} y_n{}^2$$

Notice that the $-\frac{1}{N}\sum_{n=1}^{N} y_n + \sum_{n=1}^{N} y_n{}^2$ term is constant within a given *in*-sample where the data are $y = \{y_1 \dots y_n\}$.

So, for $E_{in}(h) = N(h - \frac{1}{N}\sum_{n=1}^{N} y_n)^2 - \frac{1}{N}\sum_{n=1}^{N} y_n + \sum_{n=1}^{N} y_n{}^2$ to become minimum, $N(h - \frac{1}{N}\sum_{n=1}^{N} y_n)^2$ term should go to zero, i.e., $(h - \frac{1}{N}\sum_{n=1}^{N} y_n) = 0$

$$\therefore\ h = \frac{1}{N}\sum_{n=1}^{N} y_n = h_{mean}$$

(b)

From N points of data $y_1 \cdots y_{med} \cdots y_n$, three different cases are possible for h_{med} to exist within it.

1. $h_{med} < y_{med}$
2. $h_{med} = y_{med}$
3. $h_{med} > y_{med}$

For all cases, $E_{in}(h)$ could be defined as such.

$$\begin{aligned} E_{in}(h) &= \sum_{i=1}^{indexof\ h_{med}} |h_{med} - y_i| + \sum_{indexof\ h_{med}+1}^{N} |h_{med} - y_i| \\ &= \sum_{i=1}^{indexof\ h_{med}} (h_{med} - y_i) - \sum_{indexof\ h_{med}+1}^{N} (h_{med} - y_i) \end{aligned}$$

The right term has -1 as a coefficient because $h_{med} < y_i$ from index $(h_{med} + 1)$ to N. Rewritten,

$$E_{in}(h) = \sum_{i=1}^{indexof\ h_{med}} (h_{med} - y_i) + \sum_{indexof\ h_{med}+1}^{N} (y_i - h_{med}) \qquad \text{...for case 1 \& 3}$$

For case 2, $h_{med} = y_{med}$, so

$$E_{in}(h) = \sum_{i=1}^{indexof\ y_{med}} (y_{med} - y_i) + \sum_{indexof\ y_{med}+1}^{N} (y_i - y_{med}) \qquad \text{...for case 2}$$

Now, we assess the three different cases. In comparison to case 2, case 1 has smaller left-term and bigger right-term because $h_{med} < y_{med}$ (y_i is being subtracted by or subtracted from a smaller value).
However, the amount it gets bigger by the right-term is much greater than the amount it gets smaller by the left-term because of the indices. Left-term has index from 1 to (*index of* h_{med}), and right-term has index from (*index of* $h_{med}+1$) to N, which means now there are more bigger numbers contributing to the sum than the smaller numbers decreasing the sum. So case 1 has a bigger $E_{in}(h)$ than case 2.
The same reasoning applies between case 3 and case 2. Case 3 has a bigger $E_{in}(h)$ than case 2 because case 3 might have a smaller right-term but a much greater left-term than case 2, not just from the greater (h_{med}-y_i) values but also from the wider indices of the left-term.

(c)
The median value h_{med} will stay the same even if the rightmost data goes from y_N to infinity. However, this will increase the mean and make h_{mean} value meaningless as it will get pulled to the right and absurdly big (also infinity).