Machine Learning from Data CSCI 4100
Assignment 5
Jae Park
661994900

Exercise 2.8

(a) $\bar{g}$ can be rewritten as $\bar{g}(\mathbf{x}) = \frac{1}{N}\sum_{i=1}^{N} g_i(\mathbf{x}) = \frac{1}{N}g_1(\mathbf{x}) + \frac{1}{N}g_2(\mathbf{x}) + \cdots + \frac{1}{N}g_k(\mathbf{x})$. This shows that $\bar{g}(\mathbf{x})$ is indeed a linear combination of all the $\bar{g}_i(\mathbf{x})$'s. And $\mathcal{H}$ is closed under linear combination. Therefore, $\bar{g} \in \mathcal{H}$.

(b) Consider the function in Exercise 2.7 (b) where the binary target functions return either 1 or -1 for some dataset. Then, we can say that the average function $\bar{g}$, whose expected value(average) is 0, is not in the model's hypothesis set.

(c) I could have $\bar{g}$ to be a binary function but there is never a guarantee.
Counterexample. Consider the case from earlier in (b) where a binary function returns either 1 or -1 for some data set. It does binary classification. Now, if you had i number of such functions, according to LFD pg.63, $\bar{g}(\mathbf{x}) = \frac{1}{N}\sum_{i=1}^{N} g_i(\mathbf{x})$ and this $\bar{g}(\mathbf{x}) = 0$. However, $\bar{g}(\mathbf{x}) = 0 \,\forall\, x$ so not a binary function.

Problem 2.14

(a) When you have K of $\mathcal{H}_i$'s where $\mathcal{H}_1 \cup ... \cup \mathcal{H}_K = \mathcal{H}$ and each $\mathcal{H}_i$ has VC dimension of d_{vc}, $(d_{vc}+1)$ is the size of data points that $\mathcal{H}_i$ cannot shatter. Then for the $\mathcal{H}$, the number of all possible dichotomies is given as $\mathcal{H} < (2^{\boldsymbol{d_{vc}+1}})^{\boldsymbol{K}}$. So, $d_{vc}(\mathcal{H}) < (d_{vc}+1)K$, or $d_{vc}(\mathcal{H}) < K(d_{vc}+1)$.

(b) By the theorem 2.10 in LFD pg.50, $m_{\mathcal{H}}(l) \leq K(l^{d_{vc}}+1)$, and the coefficient K is given as a multiplier because of the number of hypotheses in the set. By the condition $(K > 1)$, it must be true that $K(l^{d_{vc}}+1) \leq 2K \cdot l^{d_{vc}}$, and $l^{\,d_{vc}}$ is together positive because both the base and exponent are positive ints.
The question also provides one more condition: $2Kl^{d_{vc}} \leq 2^l$. Combining all the inequalities, we get $m_{\mathcal{H}}(l) \leq K(l^{d_{vc}}+1) \leq 2Kl^{d_{vc}} \leq 2^l$. $m_{\mathcal{H}}(l) \leq 2^l$. Therefore, $d_{vc}(\mathcal{H}) \leq l$

(c) Proving $x \leq \min(y,z)$ is equivalent as proving two separate inequalities $x \leq y \wedge x \leq z$. In (a) we already proved that $d_{vc}(\mathcal{H}) \leq K(d_{vc}+1)$. Now, if we use (b) and let l be equal to the second part of the min, that is, $l = 7(d_{vc}+K)log_2(d_{vc}K)$

$2^{7(d_{vc}+K)log_2(d_{vc}K)} > 2K \cdot 7(d_{vc}+K)log_2(d_{vc}K)^{d_{vc}}$.

Apply log_2 on both sides

$$7(d_{vc}+K)log_2(d_{vc}K) > 1 + log_2K + log_2 7(d_{vc}+K) + log_2(d_{vc}K)^{d_{vc}}$$

$$7(d_{vc}+K)log_2(d_{vc}K) > 1 + log_2K + log_2 7 + log_2(d_{vc}+K) + d_{vc}\,log_2\,d_{vc}K$$

Because the inequality above holds true for K = 2, 3, ...
$d_{vc}(\mathcal{H}) \leq \min(K(d_{vc}+1), 7(d_{vc}+K)\log_2(d_{vc}K))$.

Problem 2.15

(a)

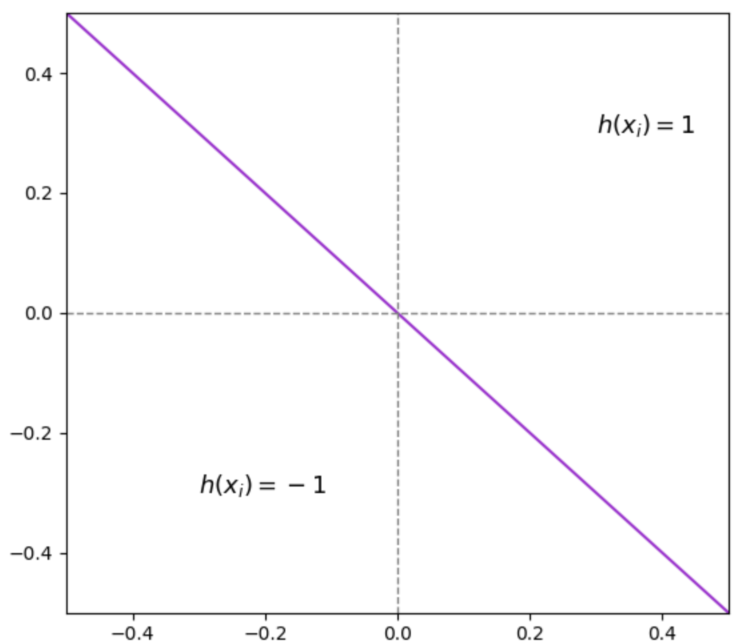

Figure 1: monotonic classifier in *2D*

(b) Consider a set of N points generated by first choosing one point and, then generating the next point by increasing the first component and decreasing the second component until N points are obtained. This monotonic classifier is capable of labeling all data points as either 1 or -1 regardless of other points. The only case when this would break is if $x_1 \geq x_2$ and $h(x_1) < h(x_2)$ was true. However, this is not possible given the construction of the classifier. And now, N data points can be shattered by $\mathcal{H}$, or as a matter of fact, any number of data points (there is no limit in N). Thus, $m_{\mathcal{H}}(N) = 2^N$ and $d_{vc} = \infty$

Problem 2.24

(a)

$$g(x) = \frac{x_2^2 - x_1^2}{x_2 - x_1}(x - x_1) + x_1^2$$

Expand and reorganize, and we get

$$= (x_1 + x_2)x - x_1 x_2$$

The average of this function in range [-1, 1]

$$\bar{g}(x) = \frac{1}{2} \cdot \frac{1}{2} \int_{-1}^{1} \int_{-1}^{1} (x_1 + x_2)x - x_1 x_2 \; dx_1 dx_2$$

$$= \frac{1}{4} \int_{-1}^{1} \frac{x_1^2}{2} x + x_2 x - \frac{x_1^2}{2} x_2 \Big|_{-1}^{1} dx_2$$

$$= \frac{1}{4} \int_{-1}^{1} x_2 x \; dx_2$$

$$= \frac{1}{4} \cdot \frac{x_2^2}{2} x \Big|_{-1}^{1} = 0$$

(b) First, generate the dataset of size N. For N times, we select two random numbers x_1, x_2 from the range [-1,+1] and determine fit a line thorough two points $(x_1, x_1^2)$and (x_2, x_2^2). From averaging the E_{out} values, we get $\mathbb{E}_{\mathcal{D}}[E_{out}(g^{(\mathcal{D})})]$. Now, we calculate

$$\begin{aligned}
\mathsf{bias} &= \mathbb{E}_x[\mathsf{bias(x)}] = \mathbb{E}_x[(\bar{g}(x) - f(x))^2] \\
\mathsf{var} &= \mathbb{E}_x[\mathsf{var(x)}] = \mathbb{E}_x[\mathbb{E}_{\mathcal{D}}[(g^{(\mathcal{D})}(x) - \bar{g}(x))^2]] \\
\mathsf{E_{out}} &= \mathbb{E}_{\mathcal{D}}\left[\mathbb{E}_x[(g^{(\mathcal{D})}(x) - f(x))^2]\right] = \mathbb{E}_x\left[\mathbb{E}_{\mathcal{D}}[(g^{(\mathcal{D})}(x) - f(x))^2]\right]
\end{aligned}$$

…LFD pg. 63, 64

(c) For the experiment, 5000 different random points from uniform distribution in range [-1,+1] were used.

$$\bar{g}(x) = -0.00283612x - 0.007965496$$

with bias = 0.203111445, var = 0.3465579906, E_{out} = 0.551598358472. bias + var ≈ 0.5496694360188, which is pretty close.

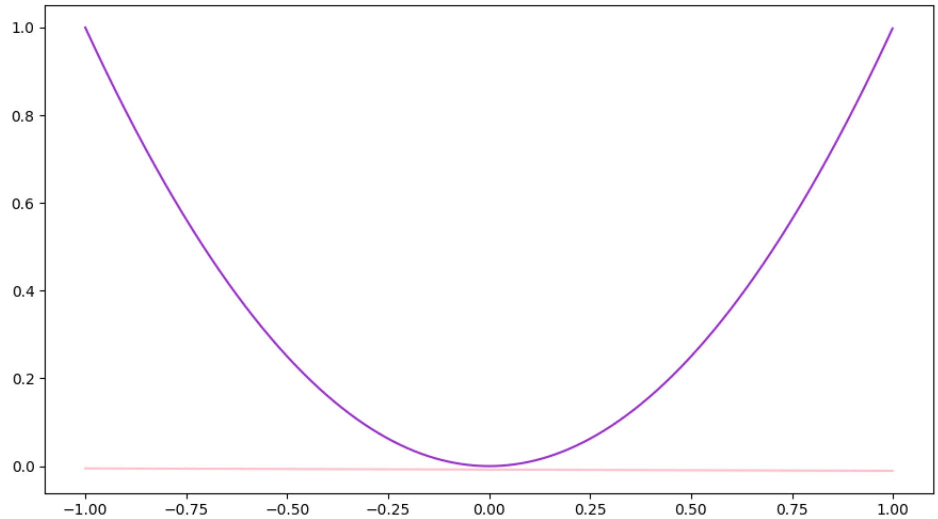

Figure 2: f(x) and $\bar{g}(x)$

d)

$$\text{bias} = \mathbb{E}_x[\text{bias(x)}] = \mathbb{E}_x[(\bar{g}(x) - f(x))^2] = \frac{1}{2}\int_{-1}^{1}(x^2)^2 \ \mathrm{d}x$$
$$= \frac{1}{2}\int_{-1}^{1} x^4 \ dx = \frac{1}{5}$$

$$\text{var} = \mathbb{E}_x[\text{var(x)}]$$
$$\text{var(x)} = \mathbb{E}_{\mathcal{D}}[(g(x)^2 - \bar{g}(x))^2] = \mathbb{E}_{\mathcal{D}}[a^2x^2 + 2abx + b^2]$$

$$= \mathbb{E}_{\mathcal{D}}[(x_1 + x_2)^2x^2 + 2(x_1 + x_2)(-x_1x_2)x + (-x_1x_2)^2]$$

$$= \mathbb{E}_{\mathcal{D}}\left[(x_1 + x_2)^2\right] \cdot x^2 - 2\mathbb{E}_{\mathcal{D}}\left[(x_1 + x_2)\, x_1x_2\right] \cdot x + \mathbb{E}_{\mathcal{D}}\left[x_1^2x_2^2\right]$$

$$= \frac{1}{4}\int_{-1}^{1}\int_{-1}^{1}\left(x_1^2 + 2x_1x_2 + x_2^2\right) dx_1dx_2 \cdot x^2 - 2 \times \frac{1}{4}\int_{-1}^{1}\int_{-1}^{1}\left(x_1^2x_2 + x_1x_2^2\right) x_2^2 \ dx_1dx_2 \cdot x + \frac{1}{4}\int_{-1}^{1}\int_{-1}^{1} x_1^2x_2^2 \ dx_1dx_2$$

$$= \frac{1}{4}\left(\frac{8}{3}\right) \cdot x^2 - 0 \cdot x + \frac{1}{4}\left(\frac{4}{9}\right) = \frac{2}{3}x^2 + \frac{1}{9}$$

$$\therefore \text{var} = \mathbb{E}_x\left[\frac{2}{3}x^2 + \frac{1}{9}\right] = \frac{1}{2}\int_{-1}^{1}\left(\frac{2}{3}x^2 + \frac{1}{9}\right) dx = \frac{1}{3}$$

$$\therefore E_{out} = \text{bias} + \text{var} = \frac{8}{15}$$