Assignment 5

Alberto Mejia
RIN: 661514960

CSCI 4100 - Machine Learning from Data

October 6, 2019

1. (200) LFD Exercise 2.8

(a)
$\bar{g} \approx \frac{1}{K}\sum_{k=1}^{K} g_k(x)$, is denoted as the 'average function', meaning that it is a form of a linear combination. $\bar{g}$ can be interpreted as a final hypothesis output for a dataset, meaning that is belongs to the hypothesis set H. Therefore, if H is closed under a linear combination, $\bar{g} \in H$.

(b)
Let us consider a binary classification model in which there is only 2 hypotheses in our hypothesis set H, $\{H_1 = +1, H_2 = -1\}$, where $g_1 = H_1 = +1$ and $g_2 = H_2 = -1$. We know that $\bar{g} \approx \frac{1}{K}\sum_{k=1}^{K} g_k(x)$ and with our dataset, we have $\bar{g} = (g_1(x) + g_2(x)) = 0$. Since our expectation of g is zero, we know that $\bar{g} \notin H$ because $\nexists\ H(x) = 0$.

(c)
No, I do not expect $\bar{g}$ to be a binary function. Say we have a binary classification model again where $x > 0 = H_1$ and $x < 0 = H_2$. Let our $g_1 = \{H_1 = +1, H_2 = -1\}$ and it's inverse $g_2 = \{H_1 = -1, H_2 = 1\}$. The average function will be 0 for all x, thus $\bar{g}$ is not a binary classification as we can see in this case, it can't be a binary classification (it can only be 0).

2. (200) LFD Problem 2.14

(a) Show $d_{vc}(H) < K(d_{vc} + 1)$
We know that the VC dimension for each H is d_{vc}, therefore our set of hypotheses includes $\{1,\ldots,k\}$ with a breakpoint $k = d_{vc} + 1$. Since there are cases for each i where H cannot shatter $d_{vc} + 1$ points, thus H can shatter at most d_{vc} points; that means in the union of all of the hypothesis, H cannot possibly shatter $K(d_{vc} + 1)$ points.
Therefore,
$d_{vc}(H) < K(d_{vc} + 1)$

(b)
From part (a), we know that
$d_{vc}(H) < K(d_{vc} + 1)$
If we have l data points for a hypothesis set H with d_{vc}, we can get at most $l^{dvc} + 1$ dichotomies. This gives us $m_{H_k}(l)(H) \leq l^{dvc} + 1$ and $m_H(l)(H) \leq K(l^{dvc} + 1)$
If H is the union of the set of Hypotheses, we have $m_H(l)(H) \leq K(l^{dvc} + K)$ or just $m_H(l)(H) \leq 2Kl^{dvc}$
Assuming that l satisfies $2^l > 2kl^{dvc}$ give us
$m_H(l)(H) \leq 2Kl^{dvc} \leq 2^l$
By definition, $d_{vc} \leq l$ because $m_H(l)(H) \leq 2^l$,

Therefore, we can conclude that $d_{vc} \leq 1$.

3. (200) LFD Problem 2.15

The monotonically increasing hypothesis set is

$$H = \{h \mid x_1 \geq x_2 \text{ -> } h(x_1) \geq h(x_2)\}$$

Where $x_1 \geq x_2$ if and only if the inequality is satisfied for every component.

(a) Here is an example of a monotonic classifier with +1 and -1 regions labeled.

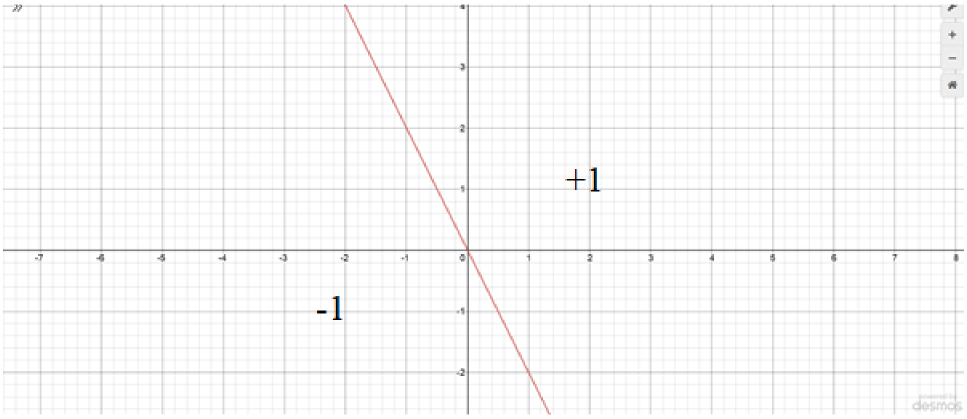


Figure 1: Example of a 2D monotonic classifier

(b)

Assume we have generated N points by choosing the next point to be larger in y than the previous but also smaller in x. So, we are essentially monotonically increasing one component of the points while also monotonically decreasing the other component. Based on this method of generating random points, there does not exist a point that is strictly greater than another. This leads us to see that H can always shatter our N points, thus giving us a $m_H(N) = 2^N$ and $d_{vc} = \infty$ Def. 2.5 & (2.10)

4. (400) LFD Problem 2.24

(a) Given our data set, $D = \{(x_1, x_1^2), (x_2, x_2^2)\}$

$$g(x) = kx + b$$

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

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

According to the average function, which is

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

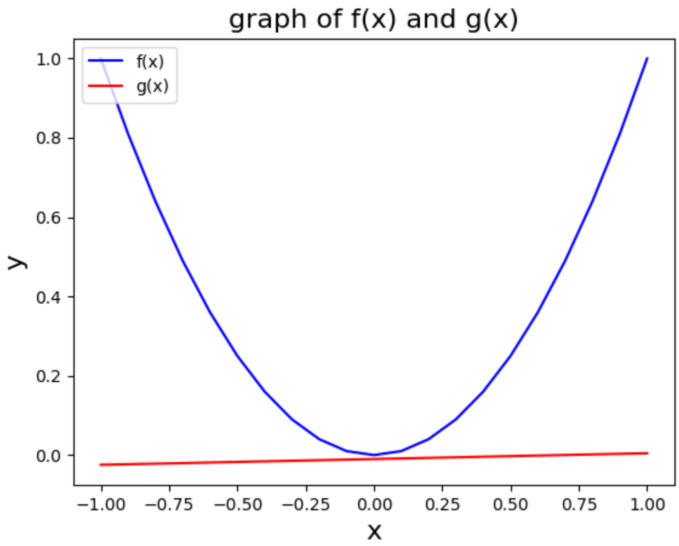
$\bar{g} \approx \frac{1}{K}\sum_{k=1}^{K} g_k(x)$

$\bar{g} = \frac{1}{K}\sum_{k=1}^{K}((x_1 + x_2)x - x_1x_2)$
$\bar{g} = 0$

(b)
Given a dataset of two points, one experiment that we could run is where we randomly generate two points via a random binomial function (which returns a random number between −1 and +1) and then calculate g(x) for these two points. We can do this for 5,000 iterations and after the experiment, we can calculate:

- Our average function, or $\bar{g} = E_D[g^{(D)}(x)]$
- E_{out} for each iteration using, $E_{out}(g^{(D)}) = E_x[(g^{(D)}(x) - f(x))^2]$
- Bias using, $bias(x) = (\bar{g}(x) - f(x))^2$
- Variance using, $var(x) = E_D[(g^{(D)}(x) - \bar{g}(x))^2]$

(c)



After the experiment, we have the following:

- $E_{out} = 0.5298 \approx 0.53$
- $Bias = 0.206 \approx 0.2$
- $Variance = 0.3293 \approx 0.33$

$E_{out} \approx bias + var$
$E_{out} \approx 0.2 + 0.33 = 0.53$
(d)

- $bias(x) = (\bar{g}(x) - f(x))^2 = \frac{1}{2}\int_{-1}^{1}(\sum_{k=1}^{K}(\bar{g}(x) - f(x))^2 \, d_x$
$= \frac{1}{2}\int_{-1}^{1}(x^2)^2 d_x$
$= \frac{1}{2}\int_{-1}^{1}(x^4) d_x$

$= \frac{1}{2}(0.4)$

$= 0.2$

- $\mathrm{var}(x) = E_D[(g^{(D)}(x) - \bar{g}(x))^2] = \frac{1}{2}\int_{-1}^{1}(\frac{1}{K}\sum_{k=1}^{K}(\bar{g}(x) - \bar{g}(x))^2)\, d_x$

$= \frac{1}{2}\int_{-1}^{1}(\frac{1}{2})(\frac{1}{2})\int_{-1}^{1}\int_{-1}^{1}((x_1 + x_2)x\text{-}x_1x_2)^2 d_{x_1} d_{x_2} d_x$

$= \frac{1}{2}\int_{-1}^{1}(\frac{1}{4})(\int_{-1}^{1}\int_{-1}^{1}((x_1 + x_2)x\text{-}x_1x_2)^2 d_{x_1} d_{x_2} d_x$

$= \frac{1}{2}(\frac{1}{4})\int_{-1}^{1}\int_{-1}^{1}\int_{-1}^{1}((x_1 + x_2)x\text{-}x_1x_2)^2 d_{x_1} d_{x_2} d_x$

$= \frac{1}{8}\int_{-1}^{1}\int_{-1}^{1}\int_{-1}^{1}((x_1 + x_2)x\text{-}x_1x_2)^2 d_{x_1} d_{x_2} d_x$

$= \frac{2}{3}x^2 + \frac{1}{9}$

$= E_D[(\frac{2}{3}x^2 + \frac{1}{9})^2]$

$= \frac{1}{2}\int_{-1}^{1}(\frac{2}{3}x^2 + \frac{1}{9})\, d_x$

$= 1/3$

$\therefore E[Eout] = 0.2 + 1/3 = 0.2 + 0.33 = 0.53$