LFD Problem Set 5

John Cohen

October 7, 2024

Exercise 2.8

We can estimate the average function for any x by $g^{-}(x) \approx \frac{1}{K}\sum_{k=1}^{K} g_k(x)$. Essentially we are viewing g(x) as a random variable, with the randomness coming from the randomness in the data set; g^{-} is the expected value of this random variable (for a particular x), and g^{-} is a function, the average function, composed of these expected values. g^{-} need not be in the model's hypothesis set, even though it is the average of the functions that are.

(a) Show that if H is closed under linear combination (any linear combination of hypothesis in H is also a hypothesis in H), then $g^{-} \in H$.
Since H is "closed under linear combination":

(1) it is also closed under scalar multiplication. For any value $h(x) \in H$ and scalar α, then $\alpha h(x) \in H$.
(2) it is closed under addition. For two $h_1(x), h_2(x) \in H$, then $(h_1(x) + h_2(x) \in H$.
Apply (1):
For all $k = 1...K$, if $g_i(x) \in H$, then $\frac{1}{K} g_i(x) \in H$.
Apply (2):
For all $k = 1...K$, if $g_i(x) \in H$, then $\sum_{k=1}^{K} g_k(x) \in H$.
Combine:
Therefore, if $\sum_{k=1}^{K} g_k(x) \in H$ and $\frac{1}{K} g_i(x) \in H$ as we have shown, then $\sum_{k=1}^{K} \frac{1}{K} g_k(x)$ or $\frac{1}{K}\sum_{k=1}^{K} g_k(x) \in H$. Because $g^{-}(x) = \frac{1}{K}\sum_{k=1}^{K} g_k(x)$, then $g^{-}(x) \in H$.

(b) Give a model for which the average function g^{-} is not in the model's hypothesis set. [Hint: Use a very simple model.]

Consider the model: $H = \{h | h(x) \in \{0, 1\}\}$.
this can give us a hypothesis set $H : g_1(x) = 0, g_2(x) = 1$.
$g^{-}(x) = \frac{1}{K}\sum_{k=1}^{K} g_k(x) = \frac{1}{2}(1) + \frac{1}{2}(0) = \frac{1}{2}$. $g^{-}(x) = \frac{1}{2} \notin H$

(c) For binary classification, do you expect g^{-} to be a binary function

No. Binary classification gives two different outputs, usually {0,1} or {-1,+1}. Average

of instances of these binary outputs will provide a value between the two, usually [0,1] or [-1,1]. It will be neither most of the time unless every instance is the same as the others, average value of all 1's, -1's, 0's etc. Thus $g^{-}(x)$ is not restricted to binary outputs as the rest of the $g_k(x)$ values in H are. Example:
Consider the model: $H = \{h | h(x) \in \{0, 1\}\}$.
this can give us a hypothesis set $H : g_1(x) = 0, g_2(x) = 1$.
$g^{-}(x) = \frac{1}{K} \sum_{k=1}^{K} g_k(x) = \frac{1}{2}(1) + \frac{1}{2}(0) = \frac{1}{2}$. $g^{-}(x) = \frac{1}{2} \notin H$. 0.5 is not a binary value, it is not {0,1}.

Problem 2.14

Let $H_1, H_2, ..., H_K$ be K hypothesis sets with finite VC dimension d_{VC}. Let $H = H_1 \cup H_2 \cup ... \cup H_K$ be the union of these models.

(a) Show that $d_{VC}(H) < K(d_{VC} + 1)$.

For all H_i hypotheses have a d_{VC} that means each can shatter its d_{VC} points. Combine all and at most their union can shatter $K * d_{VC}$ because H_1 can shatter the first d_{VC} points and H_i can shatter the ith d_{VC} points and so on until K. $d_{VC}(H) \leq K * d_{VC}$. However $K * d_{VC} + 1$ cannot work because that is one more point than the maximum all can possibility shatter. Therefore, $d_{VC}(H) < K * d_{VC} + 1$ and thus $d_{VC}(H) < K(d_{VC} + 1)$.

(b) Suppose that l satisfies $2^l > 2Kl^{d_{VC}}$. Show that $d_{VC}(H) \leq l$.

Assume the contradiction. $2^l > 2Kl^{d_{VC}}$ implies $d_{VC}(H) > l$. That is at the very closest $d_{VC}(H) = l + 1$. Substitute $d_{VC}(H) - 1$ for l in $2^l > 2Kl^{d_{VC}} \implies 2^{d_{VC}(H)-1} > 2K(d_{VC}(H) - 1)^{d_{VC}}$. This cannot be true and thus me must conclude l satisfies $2^l > 2Kl^{d_{VC}} \implies d_{VC}(H) \leq l$ and not $d_{VC}(H) > l$.

(c) Hence, show that

$$d_{VC}(H) = \min(K(d_{VC} + 1), 7(d_{VC} + K)\log_2(d_{VC}K))$$

That is, $d_{VC}(H) = O(\max(d_{VC}, K)\log_2 \max(d_{VC}, K))$ is not too bad.

From part a we know $d_{VC}(H) < K(d_{VC} + 1)$.
From part b we know if l satisfies $2^l > 2Kl^{d_{VC}}$ then l is in $7(d_{VC} + K)\log_2(d_{VC}K)$ s.t. $7(d_{VC} + K)\log_2(d_{VC}K) \geq l$. $d_{VC}(H) \leq l \leq 7(d_{VC} + K)\log_2(d_{VC}K)$ from part b and c.
Simplified: $d_{VC}(H) \leq 7(d_{VC} + K)\log_2(d_{VC}K)$.
With both bounds $d_{VC}(H) < K(d_{VC} + 1)$ and $d_{VC}(H) \leq 7(d_{VC} + K)\log_2(d_{VC}K)$ we can combine them to get: $d_{VC}(H) = \min(K(d_{VC} + 1), 7(d_{VC} + K)\log_2(d_{VC}K))$

Problem 2.15

The monotonically increase hypothesis set is

$$H = \{h | x_1 \geq x_2 \implies h(x_1) \geq h(x_2)\},$$

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

(a) Give an example of a monotonic classifier in two dimensions, clearly showing the +1 and -1 regions.

With Two Dimensions, it follows that

$$x_1 \geq y_1, x_2 \geq y_2 \implies x \geq y \implies h(x) \geq h(y)$$

With binary classification where $h(x) \in \{-1, +1\}$, $h(x) \geq h(y)$ allows us to make conclusions about certain cases:
If $x \geq y$ and $h(y) = +1 \implies h(x) = +1$
If $x \geq y$ and $h(x) = -1 \implies h(y) = -1$
Our example is as follows:

$$h(x) = \begin{cases} +1 & \text{for } x_2 \geq -x_1 + 5 \\ -1 & \text{otherwise} \end{cases}$$

This means all data points above the line $x_2 = -x_1 + 5)$ are +1 and all points bellow are -1. x_2 is a component of data point x graphed on the y-axis and x_1 is a component of data point x graphed on the x-axis.
This function is a monotonic classifier as whenever a data point x^i is greater than another x^j, $h(x^i) = h(x^j)$ or $h(x^i) > h(x^j)$. [Insert Graph]

(b) Compute $m_H(N)$ and hence the VC dimension. [Hint: 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.]

Consider the set described in the hint above. Take a point x^i and the next point x^{i+1}. By the definition of the set outlined above $x_1^i > x_1^{i+1}$ while $x_2^i < x_2^{i+1}$. We see that no two points are comparable in the coordinate-wise order of $\geq$ or $\leq$. As no two points satisfy the the inequality, the monotonicity condition cannot constrain the labels assigned to each of the points. This means any combination of labels are allowed under H. In other words, we can shatter any dichotomy for any set of N points constructed this way! From this we can conclude the growth function $m_H(N) = 2^N$ for any number N and thus $d_{VC} = \infty$.

Basic Proof: take the hypothesis $h(x) = [x_2 \geq -x_1]$ (+1 for true, -1 for false). described in part a. Starting with a random point x^1, we can make this evaluate to either -1 or +1. We

can construct x^2 s.t. $x_1^2 = x_1^1 + a$ and $x_2^2 = x_2^1 - b$ for some arbitrary constants a and b.
If $h(x^1) = -1$, we can make $h(x^2) = -1$ with $a \leq b$.
If $h(x^1) = -1$, we can make $h(x^2) = +1$ with $a \geq b + c$ for some constant c.
If $h(x^1) = +1$, we can make $h(x^2) = -1$ with $a \leq b + c$ for some constant c.
If $h(x^1) = +1$, we can make $h(x^2) = +1$ with $a \geq b$.
We can label the next point +1 or -1 regardless of the previous point.

Problem 2.24

Consider a simplified learning scenario. Assume that the input dimension is one. Assume that the input variable x is uniformly distributed in the interval [-1,+1]. The data set consists of 2 points $\{x_1, x_2\}$ and assume that the target function is $f(x) = x^2$. Thus, the full data set is $D = \{(x_1, x_1^2), (x_2, x_2^2)\}$. The learning algorithm returns the line fitting these two points as g(H consists of function of the form $h(x) = ax+b$). We are interested in the test performance (E_{out}) of our learning system with respect to the squared error measure, the bias and the var.

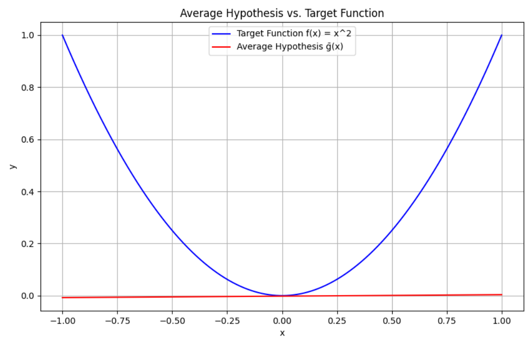
(a) Give the analytic expression for the average function $g^{-}(x)$.

I would expect $g^{-}(x) = 0$. Remember the learning algorithm returns the line fitting these two points as g(H consists of function of the form $h(x) = ax + b$). Every pair of points in a parabola has exactly 1 unique mirror pair of points. These 2 mirror pairs form 2 mirror lines. The average of each line created and its mirror is $g(x) = 0$. The set all lines $g_i(x)$ created by all pairs of points on the target function $f(x) = x^2$ can be represented as a set of all mirror line pairs. Because the average of the entire set of $g_i(x)$ lines, $g^{-}(x)$, is also average of all line pairs and the average of all pairs is 0, then $g^{-}(x) = \frac{1}{N}\sum_{i=0}^{N} g_i(x) = \frac{1}{N/2}\sum_{i=0}^{N/2} 0 = 0$ for N pairs of data points. This result comes from the symmetry of the uniform distribution and the independent sampling of x_1, x_2.

(b) Describe an experiment that you could run to determine (numerically) $g^{-}(x)$, E_{out}, bias, and var.

I am going to run a Monte Carlos simulation. I will randomly generate 2 points x_1, x_2 as values between [-1,1]. Square each point to find their y-axis values. Now we have 2 points in the 2-D space . Calculate their g(x) by finding the line that passes through both points and record it. This will be $g_i(x) = \frac{x_2^2 - x_1^2}{x_2 - x_1}x + (x_1^2 - x_2^2 - x_1^2 x_2 - x_1)$. Repeat this experiment many times, I will do so 10,000 times. Finally take the average of all 10,000 recorded hypotheses to calculate $g^{-}(x)$. This is $g^{-}(x) = \frac{1}{K}\sum_{k=1}^{K} g_k(x)$. We then calculate E_{out}, bias, and var using all g, f, g^{-}, and a test dataset T of 1000 random points between [-1,+1].

(c) Run your experiment and report the results. Compare E_{out} with bias+var. Provide a plot of your $g^{-}(x)$ and $f(x)$ (on the same plot).



Bias: 0.2021
Variance: 0.3396
E_{out}: 0.5417
Bias + Variance: 0.5417
E_{out} = Bias + Variance

(d) Compute analytically that E_{out}, bias and var should be.

a $= \frac{x_2^2 - x_1^2}{x_2 - x_1} = x_2 + x_1$ and b $= x_1^2 - (x_2 + x_1)x_1 = -x_2 x_1$.

E_{out}:

$E_{out} = E_x[(g(x) - f(x))^2] = E_x[(ax + b - x^2)^2] = E_x[x^4] - 2aE_x[x^3] + (a^2 - 2b)E_x[x^2] + 2abE_x[x] + b^2$

$$E_{out} = \frac{1}{2}\int_{-1}^{1} x^4 dx - \frac{2a}{2}\int_{-1}^{1} x^3 dx + \frac{a^2 - 2b}{2}\int_{-1}^{1} x^2 dx + \frac{2ab}{2}\int_{-1}^{1} x dx + b^2$$

$$E_{out} = \frac{1}{5} + \frac{a^2 - 2b}{3} + b^2$$

$$E_D[E_{out}] = \frac{1}{5} + \frac{1}{3}E_D[(a)^2 - 2b] + E[b^2]$$

$$E_D[E_{out}] = \frac{1}{5} + \frac{1}{3}E_D[(x_2 + x_1)^2 - 2(-x_2 x_1)] + E[(-x_2 x_1)^2]$$

$$E_D[E_{out}] = \frac{1}{5} + \frac{1}{3}E_D[x_2^2 + x_1^2 + 2x_2 x_1 + 2x_2 x_1)] + E[x_2^2 x_1^2]$$

$$E_D[E_{out}] = \frac{1}{5} + \frac{1}{3}E_D[x_2^2 + x_1^2 + 4x_2 x_1)] + E[x_2^2 x_1^2]$$

$$E_D[E_{out}] = \frac{1}{5} + \frac{1}{3}\frac{1}{4}\int_{-1}^{1}\int_{-1}^{1}(x_2^2 + x_1^2 + 4x_2 x_1)dx_1 dx_2 + \frac{1}{4}\int_{-1}^{1}\int_{-1}^{1}(x_2^2 x_1^2)dx_1 dx_2$$

$$E_D[E_{out}] = \frac{1}{5} + \frac{1}{3}\frac{1}{4}\frac{8}{3} + \frac{1}{4}\frac{4}{9}$$
$$E_D[E_{out}] = \frac{8}{15}$$

Bias:
$$bias(x) = (g^{-}(x) - f(x))^2 = 0 + 0 + f(x)^2 = (x^2)^2 = x^4$$
$$bias = E_x[bais(x)] = E_x[x^4] = \frac{1}{2}\int_{-1}^{1} x^4 dx = \frac{1}{2}\frac{2}{5} = \frac{1}{5}$$

Var:
$$var(x) = E_D[(g(x) - g^{-}(x))^2] = E_D[(g(x) - 0)^2] = E_D[(ax+b)^2] = E_D[a^2x^2 + 2abx + b^2)^2]$$
$$var(x) = E_D[a^2]x^2 + 2E_D[ab]x + E_D[b^2]$$
$$var(x) = E_D[(x_2 + x_1)^2]x^2 + 2E_D[(x_2 + x_1)(-x_2x_1)]x + E_D[(-x_2x_1)^2]$$
$$var(x) = E_D[x_2^2 + 2x_2x_1 + x_1^2]x^2 + 2E_D[-x_2^2x_1 - x_2x_1^2]x + E_D[x_2^2x_1^2]$$
$$var(x) = E_D[x_2^2 + 2x_2x_1 + x_1^2]x^2 - 2E_D[x_2^2x_1 + x_2x_1^2]x + E_D[x_2^2x_1^2]$$
$$var(x) = \frac{1}{4}\int_{-1}^{1}\int_{-1}^{1}(x_2^2+2x_2x_1+x_1^2)dx_1dx_2*x^2 - \frac{2}{4}\int_{-1}^{1}\int_{-1}^{1}(x_2^2x_1+x_2x_1^2)dx_1dx_2*x + \frac{1}{4}\int_{-1}^{1}\int_{-1}^{1}x_2^2x_1^2dx_1dx_2$$
$$var(x) = \frac{1}{4}(\frac{4}{3} + 0 + \frac{4}{3})x^2 - 0x + \frac{1}{4}(\frac{4}{9}) = \frac{2}{3}x^2 + \frac{1}{9}$$
$$var = E_x[var(x)] = E_x[\frac{2}{3}x^2 + \frac{1}{9}] = \frac{2}{3}\frac{1}{2}\int_{-1}^{+1} x^2 + \frac{1}{9} = \frac{2}{3}\frac{1}{2}\frac{2}{3} + \frac{1}{9} = \frac{1}{3}$$

The tested results are very similar to the analytical results:
var: $\frac{1}{3} \approx 0.3396$
bias: $\frac{1}{5} \approx 0.2021$
E_{out}: $\frac{8}{15} \approx 0.5417$