Homework 5
CSCI4100

Han Hai
Rin:661534083
haih2@rpi.edu

October 9th 2018

1. **[200 points] Exercise 2.8**

   (a) Show that if H is closed under linear combination (any linear combination of hypothesess in H is also a hypothesis in H), then g^{-} is in set H.
   $g^{-}(x) = \frac{1}{K}\sum_{k=1}^{K} gk(x)$ this function is a linear combination, so if H is closed under linear combination, then $g^{-} \in H$

   (b) Give a model for which the average function g^{-} is not in the model's hypothesis set.[$Hint$: Use a very simple model.]
   Let the hypothesis H contains two hyptohesis, H_1 and H_2
   Let H_1 be $g_1(x) = 0$ for any x, and let H_2 be $g_2(x) = 1$ for any x, $g^{-}(x) = \frac{1}{2}(g1(x)+g2(x)) = 0.5$.
   $g^{-}(x)$ is not in the hypothesis set

   (c) For binary classification, do you expect $g^{-}(x)$ to be a binary function?
   No, because if it's a binary function, it will only produce +1, or -1 which is not the case for $g^{-}(x)$.

2. **[200 points] Problem 2.14** Let $H_1, H_2 ... H_K$ be K hypothesis sets (K>1), each with the same 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 $d_{vc}(H) < K(d_{vc} + 1)$.
   Let's try to find a bound described in problem 2.13
   I made a guess from other people's comments on the forum
   $d_{vc}(\cup_{k=1}^{K} H_K) \leq K - 1 + \sum_{k=1}^{K} d_{vc}(H_K)$
   Let's prove this inequality
   First, $m_{H1\cup H2} \leq m_{h1} + m_{h2}$
   Because for every dicatomny in $m_{H1\cup H2}$ it's either part of m_{h1} or m_{h2}, thus the inequality holds
   Using Sauer's Lemma
   $m_{H1\cup H2} \leq \sum_{i=0}^{d_1} NCi + \sum_{i=0}^{d_2} NCi$
   $\leq \sum_{i=0}^{d_1} NCi + \sum_{i=0}^{d_2} NC(N-i)$
   $= 2^k$
   for all n such that$d1 + 1 \leq N - D2 - 1$ implies $N \geq d1 + d2 + 1$
   Therefore, we can conclude that $d_{vc}(H1 \cup H2) \leq d_1 + d_2 + 1$
   Now prove h(K): $d_{vc}(\cup_k = 1^K) \leq K - 1 + \sum_{k=1}^{K} d_{vc}(H_K)$ by induction
   Base case: K=2
   $d_vcH_1 \cup H_2 \leq 1 + \sum_{k=1}^{2} d_vcH_k$
   which is ture
   Induction step:
   Assume K-1 is correct for original hypothesis
   $d_{vc}(\cup_{k=1}^{K} H_K) = d_{vc}((\cup_{k=1}^{K-1} H_K) \cup H_K)$
   $\leq 1 + d_{vc}((\cup_{k=1}^{K-1} H_K) \cup d_{vc}H_K)$

$\leq 1 + K - 2 + \sum_{k=1}^{K-1} d_v c H_k + d_{vc} H_K$
$\leq K - 1 + \sum_{k=1}^{K} d_{vc}(H_K)$

Now we prove the bound by induction
Since every element in the set has the same dvc, we can further derive the bound
$d_{vc}(H) \leq K - 1 + K d_{vc}$
$< K(d_{vc} + 1)$

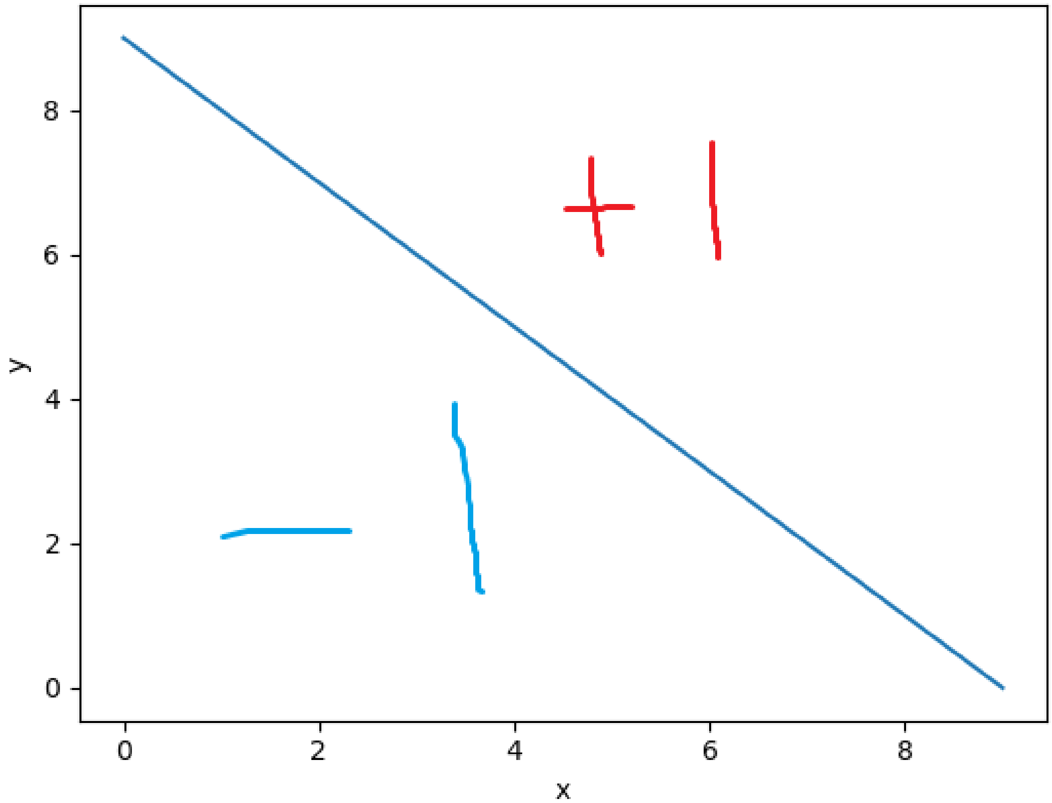
(b) Suppose that l satisfies $2^l > 2Kl^{d_{vc}}$. Show that $d_{vc}(H) \leq l$
. E2,10 states that $m_H(N) \leq N^{d_{vc}} + 1$
so $m_H(l) \leq l^{d_{vc}} + 1$
for every element k in Hypothesis set, we have $H_k(l) \leq l^{d_{vc}} + 1$
As we proved in part a $m_{H1 \cup H2} \leq m_{h1} + m_{h2}$
So we can extend that $\sum_{k=1}^{K} mH_k(l) \leq K(l^{d_{vc}} + 1)$ which $\leq 2Kl^{d_{vc}}$ since $l^{d_{vc}} > 1$
therefore $m_H(l) \leq 2Kl^{d_{vc}} < 2^l$
so H cannot shatter l points
therefore, $d_{vc} < l$

(c) Hence show that $d_{vc}(H) \leq min(K(d_{vc} + 1), 7(d_{vc} + K)log_2(d_{vc}K))$
we already proved $d_{vc}(H) \leq K(d_{vc} + 1)$
if we want to prove $d_{vc}(H) \leq 7(d_{vc} + K)log_2(d_{vc}K)) = n$ for convenient, we need to prove $2^n > 2Kn^{d_{vc}}$ from conclusion of part b
take log both sides
$n > 1 + log_2 K + d_{vc} log_2 n$
Let x be d_{vc} and y be K
LHS becomes $7(x + y)log_2(xy)$

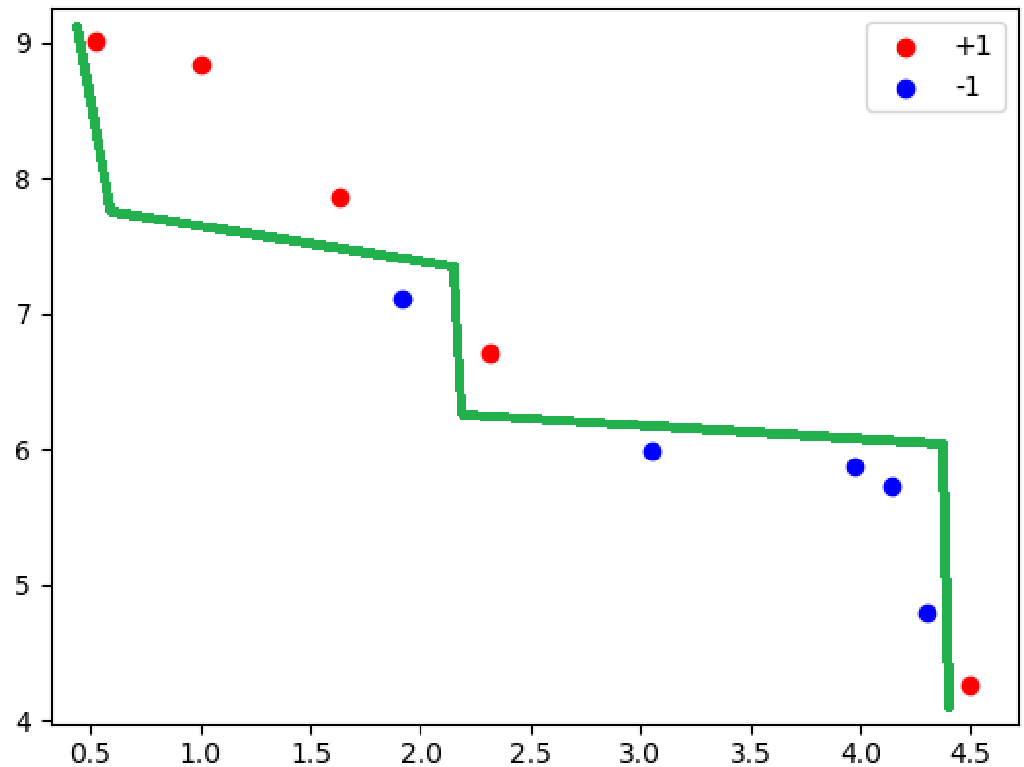
RHS becomes $1 + log_2 y + xlog_2(7(x + y)log_2(xy))$
$\leq log_2(2y) + xlog_2(7(x + y)) + xlog_2 log_2(xy)$
$< log2(xy) + xlog_2(7(x + y)) + xlog_2(xy)$ for $x >= 2$
$\leq log2(xy) + 6xlog_2(xy) + xlog_2(xy)$
since $7(x + y) < (xy)^6 when xy \geq 2$
$\leq 7ylog2(xy) + 7x(log2(xy)) = LHS$
thus we prove that n is a qualified l from part b, $d_{vc}(H) \leq 7(d_{vc} + K)log_2(d_{vc}K)$, and the claim holds

3. **[200 points] Problem 2.15** The monotonically increasing hypothesis set is
$H = \{h | x_1 \geq x_2 \rightarrow h(x_1) \geq h(x_2)\}$
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.
The below graph is a monotonic classifier in 2d with +1 above the classifier and -1 below.



(b) Compute $m_H(N)$ and hence the VC dimension.



The above graph is an example of 10 randomly generated points and a random dichotomy as hinted in the book, we can use a monotonic classifier like this green line to shatter them. We can see that a similar classifier can always shatter any dichotomy of 10 such points, thus $m_H(N) = 2^N$ and thus $d_{vc}(H) = \infty$

4. **[400 points] 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 x1,x2 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. We are interested in the test performance ($E[E_{out}]$) 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 $g^{-}(x)$.

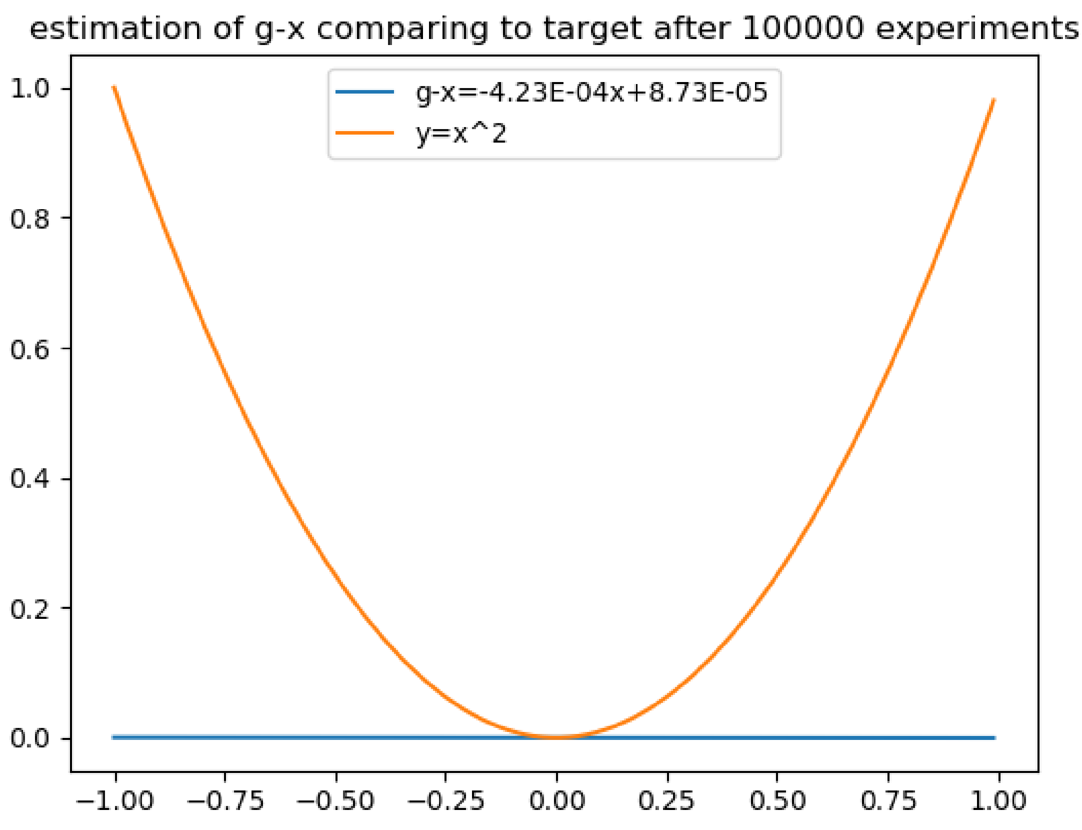
   $g(x) = ax + b$
   $x_1^2 = ax_1 + b$
   $x_2^2 = ax_2 + b$
   $a = \frac{x_2^2 - x_1^2}{x_2 - x_1} = x_1 + x_2$
   $b = x_1^2 - x_1 \frac{x_2^2 - x_1^2}{x_2 - x_1} = -x_1 x_2$
   thus $g(x) = (x_1 + x_2)x - x_1 x_2$

   $g^{-}(x) = E_D[g(x)] = \int_{-1}^{1} \int_{-1}^{1} \int_{-1}^{1} ((x_1 + x_2)x - x_1 x_2) dx_1 dx_2 dx$=0

(b) Describe an experiment that you could run to determine (numerically) g(x), $E[E_{out}], biasandvariance$

We can pick two random points between [-1,1]to be x_1, x_2, and pick the line which go through these two points to be the g for this experiment, repeat the experiment many times (K times) to obtain a hypothesis set H of size K
and then calculate $g^{-}(x)$ using $g^{-}(x) approximately = \frac{1}{K}\sum_{k=1}^{K} g_k(x)$, in this case, y=ax+b a we can take the average of each x1+x2 and b to be the average of x1x2
we can then calculate Eout using $E_{out} = E_D[E_x[(g^{(D)}(x) - f(x))^2]]$ and bias(x) using bias(x)=$(g^{-}(x) - f(x))^2$
and bias= $E_x[bias(x)]$
var(x)to be $E_D[(g^{(D)}(x) - g^{-}(x))^2]$ and variance to be $E_x[var(x)]$
and finally $E[E_{out}] = bias + variance$

(c) Run you experiment and report the results. Compare $E[E_{out}]$ with bias +variance. Provide a plot of your $g^{-}(x)$ and $f(x)$
I ran the experiment in python I chose K=100000, and got the g-x to be $g - (x) = 4.23 * 10^{-4}x + 8.73 * 10^{-5}$ However, as the K grows even larger this should approach zero.



I then calculated E_{out} using the method described in part b and got around 0.53
bias using the method described in part b and got around 0.2 and variance to be 0.33
we can see that bias+varaince=Eout 0.33+0.2=0.53 in this case

(d) Compute analytically what $E[E_{out}]$, bias and var should be
$E[E_{out}] = E_D[E_x[(g^{(D)}(x) - f(x))^2]]$
$= \frac{1}{8}\int_{-1}^{1}\int_{-1}^{1}\int_{-1}^{1}((x_1 + x_2)x - x_1x_2 - x^2)^2 dx_1 dx_2 dx$
omitted long steps to solve this integral...

$= \frac{1}{3} + \frac{1}{5} \approx 0.53$
for bias we have solved that the expected value for $g-(x)$ is zero in part a, therefore,
bias=$E_x[g-(x) - f(x)]^2$= $E_x[(0 - x^2)^2]$
=$E_x[x^4] = \frac{1}{2}\int_{-1}^{1} x^4 = 0.2$
variance=$\frac{1}{8}\int_{-1}^{1}\int_{-1}^{1}\int_{-1}^{1}[(x_1 + x_2)x - x_1x_2]^2 dx_1dx_2dx$
omitted long steps to solve this integral...
=$\frac{1}{3} \approx 0.33$
We can see that analytically the result is consistent with our experiment