Assignment 2

Alberto Mejia
RIN: 661514960

CSCI 4100 - Machine Learning from Data

September 14, 2019

1. (50) LFD Exercise 1.8

The Binomial Distribution Formula is

$$P(x) = \binom{n}{x} p^x q^{n-x}$$

Where
n = the number of trials
x = the number of successes
p = the probability of success
q = the probability of failure (1-p)

In our problem:
v = The fraction of red marbles within the sample
μ = probability of red marbles
$v \leq 0.1$
$\mu = 0.9$

Since our v is less than or equal to 0.1(10%) in a sample size of 10, we can either get 9 green marbles and 1 red marble or just 10 green marbles.

n = 10 (Our sample number)
$x = \begin{cases} 0 & \textit{if no red marbles} \\ 1 & \textit{if one red marble} \end{cases}$
p = 0.9
q = 0.1

$$P[v \leq 0.1] = \binom{10}{0} 0.9^0 0.1^{10-0} + \binom{10}{1} 0.9^1 0.1^{10-1} = 0.0000000091 = 9.1 \times 10^{-9}$$

2. (50) LFD Exercise 1.9

The Hoeffding Inequality states that for any sample size N

$$P[|v - \mu| > \varepsilon] \leq 2e^{-2\varepsilon^2 N} \quad \text{for any } \varepsilon > 0.$$

In our problem:
$v \leq 0.1$
$\mu = 0.9$
$P[v \leq 0.1]$ so $|v - 0.9| \geq 0.8$ so $\varepsilon = 0.8$
$P[v \leq 0.1] = P[|v - 0.9| > 0.8] \leq 2e^{-2(0.8)^2 10} = 5.52159 \times 10^{-6}$

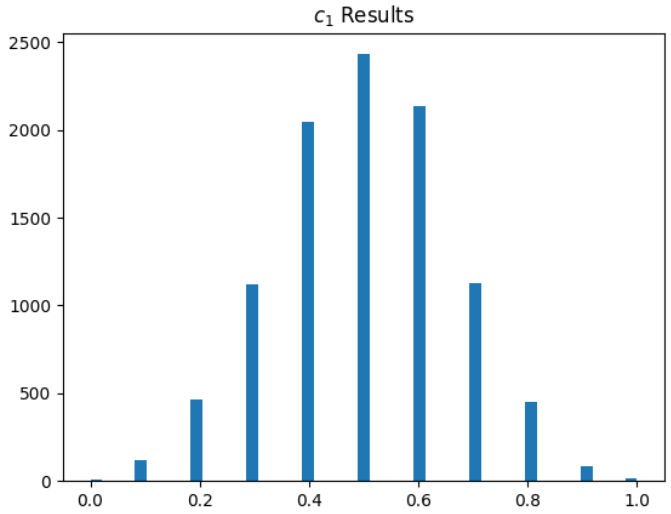
1.9 vs 1.8 = $5.52159 \times 10^{-6} > 9.1 \times 10^{-9}$
As we can observe, the probability calculated using Hoeffding's Inequality gives us a larger upper bound probability than what was calculated in the previous exercise

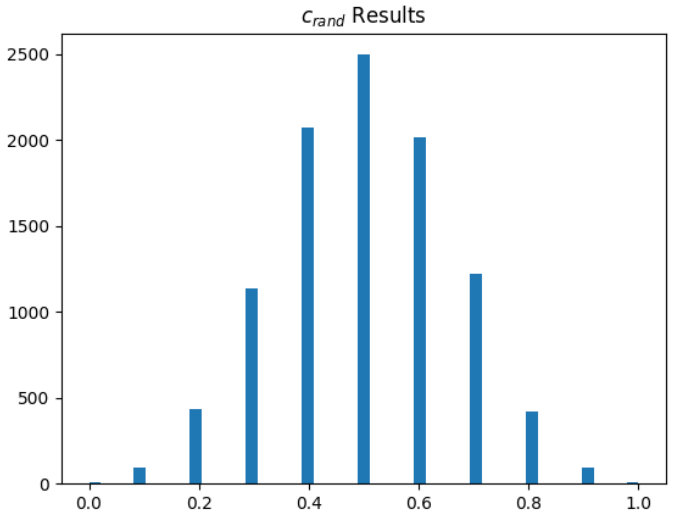
using Binomial Distribution. So these two methods can lead to different bounds of varying tightness.

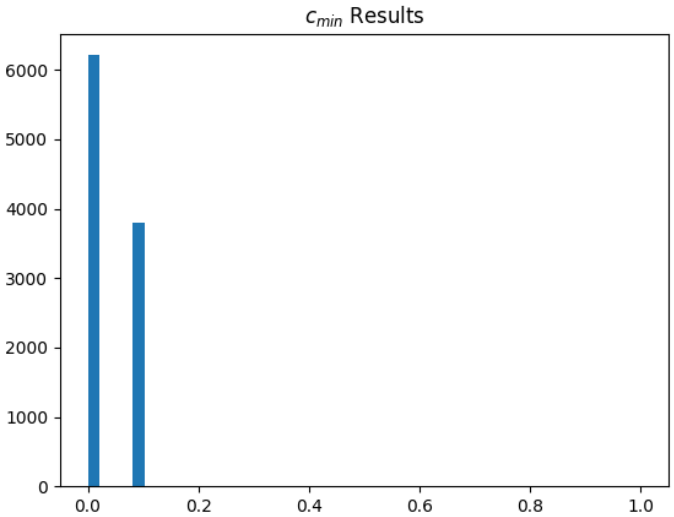
3. (100) LFD Exercise 1.10

(a) $\mu = 0.5$ for all three coins. Since the coins are fair coins, the probability of getting heads for each coin is 0.5

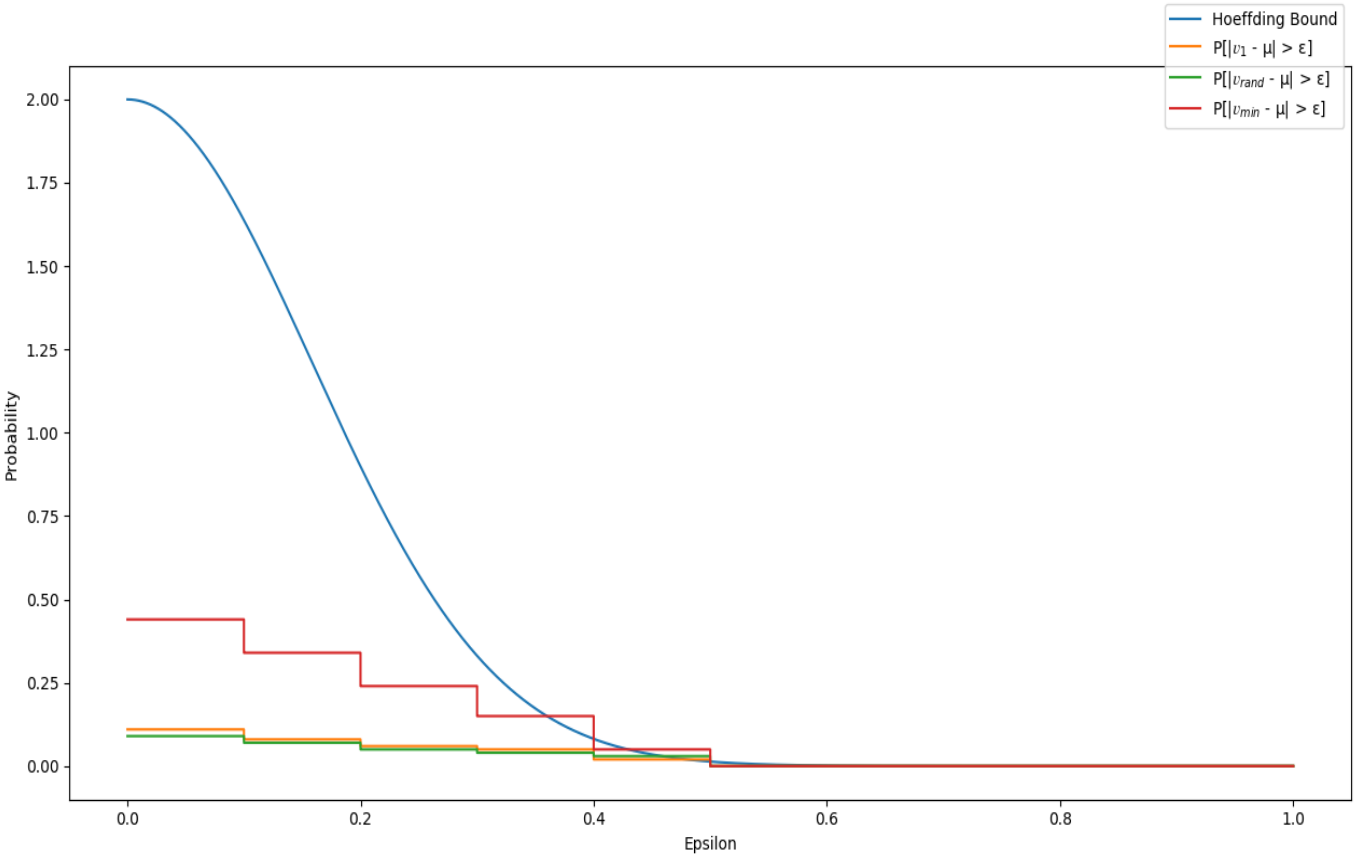
(b)


c_1 Results
2500
2000
1500
1000
500
0
0.0
0.2
0.4
0.6
0.8
1.0


c_{rand} Results
2500
2000
1500
1000
500
0
0.0
0.2
0.4
0.6
0.8
1.0


c_{min} Results
6000
5000
4000
3000
2000
1000
0
0.0
0.2
0.4
0.6
0.8
1.0

(c)



(d) According the following plots produced, c_1 and c_{rand} follow Hoeffding's bound as they don't go out of the bound like c_{min} does. They also decrease in a similar fashion as the bound so they are forever contained where as c_{min} doesn't so it eventually protrudes out of the bound. Also, it can be seen that c_1 and c_{rand} produce pretty much identical lines; the reson for this is that they are both randomly chosen. c_{min} was not random since we chose it beforehand.

(e) As we examined before, c_1 and c_{rand} follow Hoeffding's bound because they are both fixed before the experiments. c_{min} is dependent on the data since we did not choose it randomly, we chose it on the premise of being the minimum coin and thus we violated Hoeffding's Bound.

4. (100) LFD Exercise 1.11

(a) No. As told by Prof. Malik, the whole purpose of this class is to find if we CAN learn anything outside of the data set. But as shown in lecture, we can only estimate (E_{out}) how a hypothesis may perform outside of the data set. One thing we cannot do is guarantee that it will perform better than random on points outside of our data set.

(b) Yes. As we just covered in part (a), we cannot guarantee ANYTHING about how our hypothesis will do outside of our dataset. We can only estimate with proved

theorems how they will perform. Because of this, if we cannot guarantee how algorithm S or C will perform outside of the dataset, then it is theoretically possible for C to perform better than S. It's also possible for S to perform amazing on the dataset then just perform poorly outside of the dataset. To reiterate, nothing is guaranteed.

(c) $p = 0.9$
S agrees the most with D so
$P(S = f) = p = 0.9$
While C chooses the other hypothesis (-1) so
$P(C = f) = (1\text{-}p) = 0.1$
So $P[P(S = f) > P(C = f)] = P[0.9>0.1] = 1$
S performs 90% better than C. 90% always performs better than the 10% C performs so there is 100% probability that S will perform better than C.

(d) Only if our p-value is small. If our p-value is < 0.5, that means that there is strong evidence against the null hypothesis, thus you reject the null hypothesis. Here, S is the null hypothesis since C always directly contradicts the null hypothesis (alternate hypothesis).

5. (100) LFD Exercise 1.12

Definitely not (a) as explained in exercise 1.11

Again, we can't promise anything so I can't promise my friend that I will come back with a g with high probability that it will perform well out of sample. So not (b) either.

(c) Is the best answer because I can promise to either get one of the following: fail in finding a good g or find a good g. Gives me flexibility.

6. (300) LFD Problem 1.3

(a) $p = min_{1\leq n\leq N}\ y_i({w*}^T x_n)$
As defined, w^* is an optimal set of weights that separates the data. This means what w^* correctly classifies each data point x_i where $i \in \{1,...,N\}$. As a result, we have

$$y_i = \text{sign}({w*}^T x_i)$$

However, based on the update rule, correctly classified data means we always have

$$y_i({w*}^T x_i) > 0$$

$$\text{So } p > 0$$

This fact was also reasoned in Exercise 1.3 part (a)

(b) Show $w^T(t)w^* \geq w^T(t\text{-}1)w^* + p$
The update rule is:

$$w(t+1) = w(t) + y(t)x(t)$$

So to compute the L.H.S (Left hand side)
$w^T(t)w^* = [w^T(t\text{-}1) + y(t\text{-}1)x(t\text{-}1)]w^*$ Subtract 1 from t in the update rule
$[w(t\text{-}1) + y(t\text{-}1)x(t\text{-}1)]^T w^*$
$[w^T(t\text{-}1) + y(t\text{-}1)x^T(t\text{-}1)]w^*$ y is a scalar
$w^T(t\text{-}1)w^* + y(t\text{-}1)x^T(t\text{-}1)w^*$ Distribute in w^*

To computer the R.H.S (Right hand side)
From part (a), we know p the minimum of $y_i(w^{*T}x_n)$
$p = \boldsymbol{min}_{1 \leq n \leq N}\, y_i(w^{*T}x_n) > 0$

$w^T(t)w^* \geq w^T(t\text{-}1)w^* + p$
$w^T(t)w^* \geq w^T(t\text{-}1)w^* + \boldsymbol{min}_{1 \leq n \leq N}\, y_{(t\text{-}1)}(w^{*T}x_{(t\text{-}1)})$ Substitute in p

Comparing the results of the LHS to the RHS
$w^T(t\text{-}1)w^* + y(t\text{-}1)x^T(t\text{-}1)w^* \geq w^T(t\text{-}1)w^* + \boldsymbol{min}_{1 \leq n \leq N}\, y_{(t\text{-}1)}(w^{*T}x_{(t\text{-}1)})$
$y(t\text{-}1)x^T(t\text{-}1)w^* \geq \boldsymbol{min}_{1 \leq n \leq N}\, y_{(t\text{-}1)}(w^{*T}x_{(t\text{-}1)})$ Subtract from both sides
So it looks like $y(t\text{-}1)x^T(t\text{-}1)w^* \geq p$
$w^T(t)w^* \geq w^T(t\text{-}1)w^*$

Thus,
$w^T(t)w^* \geq w^T(t\text{-}1)w^* + p$

PROVE $w^T(t)w^* \geq tp$
<u>Base case:</u>

Let $t = 0$
Assume $w(0) = 0$
That means $w^T(0)w^* = 0$
$tp = (0)p = 0$
$0 \geq 0$
So $w^T(t)w^* \geq tp$ is true when $t = 0$.

Now we assume that $w^T(t)w^* \geq tp$ holds true for any $t > 0$
Let's prove for t+1 now.

<u>Induction step:</u>

Let $q := t$
Show $w^T(t+1)w^* \geq (t+1)p$

Remember, $w^T(q)w^* \geq qp$

RHS

$w^T(q)w^* + p = qp + p$

$w^T(q+1)w^* \geq qp + p$

$w^T(q+1)w^* \geq (q + 1)p$ Factor out p

$w^T(t+1)w^* \geq (t + 1)p$ Substitute t back in

PROVED

(c) Show $||w(t)||^2 \leq ||w(t\text{-}1)||^2 + ||x(t\text{-}1)||^2$

Recall the update rule:

$$w(t + 1) = w(t) + y(t)x(t)$$

$w(t) = w(t\text{-}1) + y(t\text{-}1)x(t\text{-}1)$

$w(t)^2 = w^T(t)w(t) = w^T(t)[w(t\text{-}1) + y(t\text{-}1)x(t\text{-}1)]$

$= w^T(t)w(t\text{-}1) + w^T(t)y(t\text{-}1)x(t\text{-}1)$

$y(t\text{-}1)(w^T(t\text{-}1)x(t\text{-}1)) \leq 0$ because x(t-1) was misclassified by w(t-1)

$w^T(t)w(t\text{-}1) \leq ||w(t\text{-}1)||^2 + ||x(t\text{-}1)||^2$

Thus

$||w(t)||^2 \leq ||w(t\text{-}1)||^2 + ||x(t\text{-}1)||^2$

(d) From part (c) we have

$||w(t)||^2 \leq ||w(t\text{-}1)||^2 + ||x(t\text{-}1)||^2$

Prove $||w(t)||^2 \leq tR^2$

Base case:

Let $t = 0$

Assume that $||w(t)||^2 \leq tR^2$ is true for $t = 0$

$||w(0)||^2 \leq (0)R^2$

$0 \leq 0$

Trivially, this holds

Now we assume that $||w(t)||^2 \leq tR^2$ holds true for any t

More specifically, lets show that Show $||w(t+1)||^2 \leq tR^2$

Induction step:

So we now that it holds true for t when $t = 0$.

$||w(t+1)||^2 \leq (t+1)R^2$

$||w(t+1)||^2 \leq tR^2+R^2$

However, let us make our basis step t-1 and show t is also true.

$||w(t+1)||^2 \leq tR^2+R^2$

$||w(t)||^2 \leq (t\text{-}1)R^2+R^2$

But we know from part (c)

$||w(t)||^2 \leq ||w(t\text{-}1)||^2 + ||x(t\text{-}1)||^2$

$||w(t)||^2 \leq ||w(t\text{-}1)||^2 + ||x(t\text{-}1)||^2 \leq (t\text{-}1)R^2 + R^2$

$||w(t)||^2 \leq ||w(t\text{-}1)||^2 + ||x(t\text{-}1)||^2 \leq tR^2\text{-}R^2 + R^2$

$||w(t)||^2 \leq ||w(t\text{-}1)||^2 + ||x(t\text{-}1)||^2 \leq tR^2$

Thus $||w(t)||^2 \leq tR^2$

(e) Show $\frac{w^T(t)}{||w(t)||} w^* \geq \sqrt{t}\,\frac{p}{R}$ proving $t \leq \frac{R^2||w^*||^2}{p^2}$

part (d) gave us $||w(t)||^2 \leq tR^2$

which is equivalent to $||w(t)|| \leq \sqrt{tR^2}$ -> $||w(t)|| \leq \sqrt{t}R^2$

part (b) gave us $w^T(t)w^* \geq tp$

So $\frac{w^T(t)}{||w(t)||} w^* \geq \frac{tp}{\sqrt{t}R} = \sqrt{t}\,\frac{p}{R}$ Subtract t

$\frac{R}{p}\frac{w^T(t)}{||w(t)||} w^* \geq \frac{tp}{\sqrt{t}R} = \sqrt{t}$ Multiply by the reciprocal

$\frac{R}{p}\frac{w^T(t)}{||w(t)||} w^* \geq \sqrt{t}$

But given the hint

$\frac{w^T(t)w^*}{||w(t)||\,||w^*||} \leq 1$

And using Cauchy–Schwarz inequality that states

$|\langle u,v\rangle|^2 \leq \langle u,u\rangle\langle v,v\rangle$

$|\,w^T(t)w^*|^2 \leq ||w(t)||^2\,||w^*||^2$

We simply $\frac{R}{p}\frac{w^T(t)}{||w(t)||} w^* \geq \sqrt{t}$ based on the inequality

$\frac{Rw^*}{p} \geq \sqrt{t}$ Get rid of square root

Thus,

$t \leq \frac{R^2||w^*||^2}{p^2}$

7. (300) LFD Problem 1.7

(a)

$$P[k \mid N, \mu] = \binom{N}{k}\mu^k(1-\mu)^{N-k}$$

$v = \frac{k}{N}$

$k = (v)(N)$

$v = 0$

$k = (0)N = 0$

$\mu = 0.05$

$N = 10$

The number of coins that have μ=0.05 is itself a binomial distribution with probability p.

$\binom{10}{0}(0.05)^0 = 1$ so we can ignore this expression

$P[\text{Atleast 1 coin}] = \binom{10}{0}(0.05)^0(1-0.05)^{10-0} = 0.59873 = p$

$P[\text{Atleast 1,000 coin}] = 1 - \left(1 - (0.5987)\right)^{1000} = 1$

$P[\text{Atleast 1,000,000 coin}] = 1 - (1 - (0.5987))^{1000000} = 1$

$\mu = 0.8$

P[Atleast 1 coin] $= (1 - 0.8)^{10-0} = 1.024 \times 10^{-7} = p$
P[Atleast 1,000 coin] $= 1 - \left(1 - (1.024 \times 10^{-7})\right)^{1000} = 1.024 \times 10^{-4}$
P[Atleast 1,000,000 coin] $= 1 - \left(1 - (1.024 \times 10^{-7})\right)^{1000000} = 0.09733$

(b)