MLD HW2

yasir Mohamed

September 2024

1 Exercise 1.8

P[red $\leq$ 1] = P[red = 1] + P[red = 0]

$$= \binom{10}{1} u * v^9 + v^{10} \ = 10 * 0.9 * 0.1^9 + 0.1^{10} = 9.1 * 10^{-9}$$

2 Exercise 1.9

According to Hoeffding

$$P[|v - \mu| > \epsilon] \leq 2e^{-2\epsilon^2 N}$$

N = 10 = Number of samples $\mu = 0.9$ v = .1 $\epsilon = 0.7$ since an error of more than 0.7 will result in v $\leq$ 0.1

$$P[v \leq 0.1] \quad \leq \quad P[|v - \mu| > \epsilon] \quad \leq \quad 2e^{-2e^2 N} \approx 0.000110903198864$$

3 Exercise 1.10

3.1 A

μ = 0.5 as fair coins have an equal probability of resulting in a head or tails when flipped

3.2 B

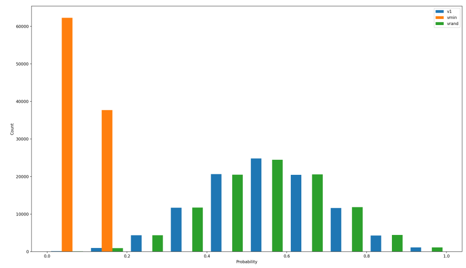

Figure 1: Conut of v_1 ,v_{rand}, and v_{min} across experiments

3.3 C

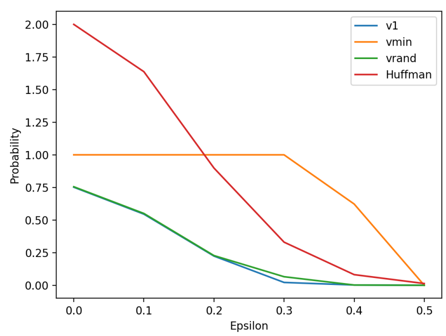

Figure 2: $P[|v - \mu] > \epsilon$ for $0 \leq \epsilon \leq 5$

3.4 D

since the randomly selected coin v_{rand} is selected completely on random it obeys Hoeffding. The first coinv_1 obeys Hoeffding since it is selected before we look at the data, the probabilities of other coins, making it random. The min coinv_{vmin} does not obey Hoeffding as we selectively choose the coin with the smallest probability by looking at the probabilities of other coins.

3.5 E

In our case, we have 100,000 different bins or hypotheses($|H| = 100,000$) that we are selecting from to approximate the outer bin function,f, and when we choose v_{min} we are selecting a single hypothesis,h_min, from our hypothesis set after sampling our data to represent f. This is similar to the learning algorithm in which we select a hypothesis from our hypothesis set to estimate f. As for V_{min} and v_1 they are chosen before sampling since they are selected before looking at the data set or the probability of each coin flip.

4 Exercise 1.11

4.1 a

No, since the sample,D, can incorrectly resemble the outside of D. If outside of D has a probability of 0.1 of +1 and .9 probability of -1, the sample D may get extremely unlucky and contain more +1 than -1 resulting in a function,S, always outputting +1 which will result in a larger error in D_out compared to random

4.2 B

In the same scenario above if we select C instead of S then the out-of-sample error will decrease, while our in=sample error increase

4.3 C

for S to produce a better hypothesis then we need to have at least 13 +1 in the sample data. The probability of selecting at least 13 '+1' data points in D :

$$p[+1 \geq 13] = p[13]+p[14]+..+p[25] = \sum_{k=13}^{25} \binom{25}{k} 0.9^k(.1)^{25-k} = 0.9999998379165841 \approx 1$$

4.4 D

NO since in most cases, the in-sample data will resemble the out-of-sample function as such selection of the hypothesis that more closely resembles in-sample data will most likely result in the hypothesis that most resemble out-of-sample data

5 Exercise 1.12

I will promise her that I will either produce a hypothesis g or declare that I failed. we are guaranteed 4000 data points so the only variable that could change is our hypothesis set.

We will set a hypothesis set of fixed size and if the function is too complex then none of the functions in the hypothesis set will agree with the data and we will declare failure. If the function is simple enough and can be approximated with one of the functions in the hypothesis set then the algorithm will eventually converge to that function and return it as g which should approximate f well due to the large number of data points.

6 Problem 1.3

7 A

they have the same sign

$$sign(W^{*T}X_n) = y_n$$

therefore multiplying them will always result in $p \geq 0$

7.1 B

At t = 1:

$$w^T(t)w^* \geq w^T(t-1)w^* + \rho$$

$$w^T(1)w^* \geq w^T(0)w^* + \rho$$

$$w^T(0)w^* + y_i(x_i^T w^*) \geq (0)w^* + \rho$$

$$y_i(x_i^T w^*) \geq \rho$$

which is given in part a for t+1 we assume the previous:

$$w^T(t+1)w^* \geq w^T(t)w^* + \rho$$

$$w^T(t+1)w^* = w^T(t)w^* + y_i(x_i^T w^*) \geq w^T(t)w^* + \rho$$

this also shows that"

$$w^T(t) \geq w^T(t-1)w^* + \rho \geq w^T(t-2) + 2\rho \geq ... \geq t\rho$$

7.2 C

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

$$||w(t)||^2 = ||w(t-1) + y(t-1)x(-1)||^2 =$$

$$||w(t-1) + y(t-1)x(-1)||||w(t-1) + y(t-1)x(-1)||^T$$

since y(t-1) will either be a +1 or -1 then it will simply become a +1 when squared:

$$= ||w(t-1)||^2 + ||x(t-1)||^2 + 2y(t-1)x(t-1)w(t-1)$$

since we know that it y(t-1) was misclassified we know that $y(t-1)x(t-1)w(t-1)$ will be negative therefore

$$||w(t-1)||^2 + ||x(t-1)||^2 + 2y(t-1)x(t-1)w(t-1) \leq ||w(t-1)||^2 + ||x(t-1)||^2$$

7.3 D

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

at t = 0:

$$||w(0)||^2 \leq (0)R^2$$
$$02 \leq 0$$

for t = 1:

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

$$||w(0) + y_i(0)x_i(0)||^2 \leq R^2$$

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

for t + 1:

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

$$||w(t+1)||^2 = ||w(t) + x_j(t)y_j(t)||^2 = ||w(t)||^2 + ||x_j(t)||^2 + ||y_j(t)x_j(t)^T w(t)$$

use $||w(t)||^2 \leq tR^2$ and the definition of R :

$$||w(t)||^2 + ||x_j(t)||^2 + ||y_j(t)x_j(t)^T w(t) <= tR^2 + R^2 + ||y_j(t)x_j(t)^T w(t)$$

we know that $x_j(t)^T w(t)$ has the opposite sign to $y_j(t)$ meaning it will be ≤ 0:

$$tR^2 + R^2 + ||y_j(t)x_j(t)^T w(t) \leq tr^2 + R^2$$

7.4 E

$$\frac{w^T(t)w^*}{||w(t)||} \geq \frac{t\rho}{\sqrt{tR^2}} = \frac{\sqrt{t}\rho}{R}$$

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

now if we start to rewrite the equation in terms of t:

$$\sqrt{t} \leq \frac{Rw^T(t)w^*}{\rho||w(t)||}$$

$$t \leq \frac{R^2||w(t)||^2||w^*||^2}{\rho^2||w(t)||^2}$$

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

8 Problem 1.7

8.1 A

For 1 coin with $\mu = 0.05$

$$P[1] = \binom{10}{0} 0.05^0(1-0.05)^{10-0} = 0.598736939238$$

the probability of n coins having at least one head is 1 $P[n]$ - the probability of all coins having 0 heads$p[n_0]$. the probability of $P[n_0] = (1-p[1])^n$ Therefore for 1,000 coins the probability of at least one head is

$$P[1000] = 1 - P[1000_0] = 1 - (1-p[1])^{1000} = 1 - (0.401263060762)^{1000} \approx 1$$

for 1,000,000 coins

$$P[1,000,000] = 1 - P[,000,1000_0] = 1 - (1-p[1])^{1,000,000} = 1 - (0.401263060762)^{1,000,000} \approx 1$$

With $\mu = 0.8$:

$$P[1] = \binom{10}{0} 0.8^0(1-0.8)^{10-0} = 1.024 * 10^{-7}$$

$$p[1,000] = 1 - p[1]^{1000} = 0.000102394762576$$

$$p[1,000,000] = 1 - p[1]^{1,000,000} = 0.0973315926832$$

8.2 B

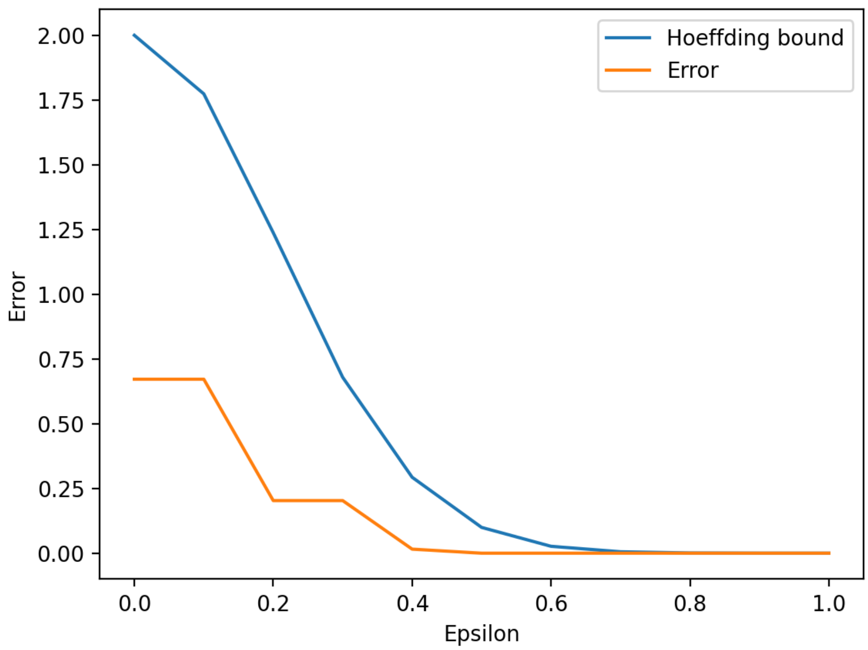

Figure 3: Error Probability vs Epislon