Homework 02 LaTeX Sheet

Cameron Brenner

3/6/2019

1 Exercise 2.1

In Exercise 2.1, we are given equation 2.1 which is

$$\epsilon(M, N, \delta) = \sqrt{\frac{1}{2N} ln \frac{2M}{\delta}} \tag{1}$$

and we are given the value $\delta = 0.03$

1.1 Part A.

In part A. of Exercise 2.1, given equation 2.1, we are asked to find the number of examples, N, to make $\epsilon \leq 0.05$, and we are given $M = 1$. This can be done by solving for N.

$$0.05 \geq \sqrt{\frac{1}{2N} ln \frac{2}{0.03}} \tag{2}$$

$$ln \frac{2}{0.03} = 4.1997 \tag{3}$$

$$0.05 \geq \sqrt{\frac{4.1997}{2N}} \tag{4}$$

$$0.05^2 \geq \frac{4.1997}{2N} \tag{5}$$

$$.0025 \geq \frac{4.1997}{2N} \tag{6}$$

$$\frac{.0025}{4.1997} \geq \frac{1}{2N} \tag{7}$$

$$\frac{4.1997}{.0025} \geq 2N \tag{8}$$

$$1679.88 \geq 2N \tag{9}$$

$$839.94 \geq N \tag{10}$$

$$\tag{11}$$

So $N \leq 839.94$, meaning that is how many examples are needed to make $\epsilon \leq 0.05$ when $M = 1$.

1.2 Part B.

In part A. of Exercise 2.1, we are asked to do the same thing as part A, but instead we are given $M = 100$.

$$0.05 \geq \sqrt{\frac{1}{2N} ln \frac{200}{0.03}} \tag{12}$$

$$ln \frac{200}{0.03} = 8.8049 \tag{13}$$

$$0.05 \geq \sqrt{\frac{8.8049}{2N}} \tag{14}$$

$$0.05^2 \geq \frac{8.8049}{2N} \tag{15}$$

$$.0025 \geq \frac{8.8049}{2N} \tag{16}$$

$$\frac{.0025}{8.8049} \geq \frac{1}{2N} \tag{17}$$

$$\frac{8.8049}{.0025} \geq 2N \tag{18}$$

$$3521.96 \geq 2N \tag{19}$$

$$1760.98 \geq N \tag{20}$$

$$\tag{21}$$

So $N \leq 1760.98$, meaning this is the upper bound of examples needed to get $\epsilon \leq 0.05$ when $M = 100$.

1.3 Part C.

In part C. of Exercise 2.1, we are asked to do the same thing as the above parts, but we are given $M = 10,000$.

$$0.05 \geq \sqrt{\frac{1}{2N} ln \frac{20000}{0.03}} \tag{22}$$

$$ln \frac{20000}{0.03} = 13.41 \tag{23}$$

$$0.05 \geq \sqrt{\frac{13.41}{2N}} \tag{24}$$

$$0.05^2 \geq \frac{13.41}{2N} \tag{25}$$

$$.0025 \geq \frac{13.41}{2N} \tag{26}$$

$$\frac{.0025}{13.41} \geq \frac{1}{2N} \tag{27}$$

$$\frac{13.41}{.0025} \geq 2N \tag{28}$$

$$5364 \geq 2N \tag{29}$$

$$2682 \geq N \tag{30}$$

$$\tag{31}$$

So $N \leq 2682$, meaning this is the upper bound of examples needed to get $\epsilon \leq 0.05$ when $M = 10000$.

2 Exercise 2.11

In Exercise 2.11, we are asked, given $m_H(N) = N + 1$ so the $d_{vc} = 1$, to use the generalization bound to give a bound for E_{out} with confidence of 90%.

The generalization bound is given as equation 2.12, which is

$$E_{out}(g) \leq E_{in}(g) + \sqrt{\frac{8}{N} ln \frac{4m_H(N)}{\delta}} \qquad (32)$$

In the Exercise, we are told we have 100 training examples, so $N = 100$, and because the confidence is 90%, $\delta = 0.1$.

$$E_{out}(g) \leq E_{in}(g) + \sqrt{\frac{8}{100} ln \frac{4m_H(100)}{0.1}} \qquad (33)$$

$$E_{out}(g) \leq E_{in}(g) + \sqrt{\frac{8}{100} ln \frac{4(101)}{0.1}} \qquad (34)$$

$$E_{out}(g) \leq E_{in}(g) + \sqrt{\frac{8}{100} ln \frac{404}{0.1}} \qquad (35)$$

$$ln \frac{404}{.1} = 8.304 \qquad (36)$$

$$E_{out}(g) \leq E_{in}(g) + \sqrt{\frac{66.432}{100}} \qquad (37)$$

$$E_{out}(g) \leq E_{in}(g) + \sqrt{.664} \qquad (38)$$

$$E_{out}(g) \leq E_{in}(g) + 0.815 \qquad (39)$$

$$\qquad (40)$$

We are then asked to repeat finding the bound, but when there are 10,000 training examples, so now $N = 10000$.

$$E_{out}(g) \leq E_{in}(g) + \sqrt{\frac{8}{10000} ln \frac{4m_H(10000)}{0.1}} \qquad (41)$$

$$E_{out}(g) \leq E_{in}(g) + \sqrt{\frac{8}{10000} ln \frac{4(10001)}{0.1}} \qquad (42)$$

$$E_{out}(g) \leq E_{in}(g) + \sqrt{\frac{8}{10000} ln \frac{40004}{0.1}} \qquad (43)$$

$$ln \frac{40004}{.1} = 12.899 \qquad (44)$$

$$E_{out}(g) \leq E_{in}(g) + \sqrt{\frac{103.192}{10000}} \qquad (45)$$

$$E_{out}(g) \leq E_{in}(g) + \sqrt{0.01} \qquad (46)$$

$$E_{out}(g) \leq E_{in}(g) + 0.1 \qquad (47)$$

$$\qquad (48)$$

3   Exercise 2.12

In exercise 2.12 we have a hypothesis set H with a d_{vc} of 10. We are asked to find the sample size given 95% confidence and a generalization error no more than 0.05. The equation we will use is equation 2.13, which is:

$$N \geq \frac{8}{\epsilon^2} ln \frac{4((2N)^{d_{vc}} + 1)}{\delta} \qquad (49)$$

We then substitute 0.05 for δ and ϵ and multiply out the 4 to simplify the equation a little. We also substitute the $d_{vc} = 10$.

$$N \geq \frac{8}{0.05^2} ln \frac{4(2N)^{10} + 4}{0.05} \qquad (50)$$

We start by making an estimate for N, and then plugging the result into N again, and repeating until it converges to a sample size. Assume $N = 10,000$

$$N \geq \frac{8}{0.05^2} ln \frac{4(2(10000)^{10} + 4}{0.05} \tag{51}$$

$$N \geq \frac{8}{0.0025} ln \frac{4(20000)^{10} + 4}{0.05} \tag{52}$$

$$N \geq \frac{8}{0.0025} ln \frac{4(abignumber) + 4}{0.05} \tag{53}$$

$$N \geq \frac{8}{0.0025} * 100.421 \tag{54}$$

$$N \geq 3200 * 100.421 \tag{55}$$

$$N \geq 321347.2 \tag{56}$$

$$\tag{57}$$

So now we plug this number back into the equation.

$$N \geq \frac{8}{0.05^2} ln \frac{4(2(321347.2)^{10} + 4}{0.05} \tag{58}$$

$$N \geq \frac{8}{0.0025} ln \frac{4(642694.4)^{10} + 4}{0.05} \tag{59}$$

$$N \geq \frac{8}{0.0025} ln \frac{4(1.2023895e + 58) + 4}{0.05} \tag{60}$$

$$N \geq \frac{8}{0.0025} * 138.116 \tag{61}$$

$$N \geq 3200 * 138.116 \tag{62}$$

$$N \geq 441972.073 \tag{63}$$

$$\tag{64}$$

We are getting closer so now we plug this number in for N.

$$N \geq \frac{8}{0.05^2} ln \frac{4(2(441972.073)^{10} + 4}{0.05} \tag{65}$$

$$N \geq \frac{8}{0.0025} ln \frac{4(883944.146)^{10} + 4}{0.05} \tag{66}$$

$$N \geq \frac{8}{0.0025} ln \frac{4(2.9123813e + 59) + 4}{0.05} \tag{67}$$

$$N \geq \frac{8}{0.0025} * 141.304 \tag{68}$$

$$N \geq 3200 * 141.304 \tag{69}$$

$$N \geq 452172.8 \tag{70}$$

$$\tag{71}$$

The difference is getting noticeably smaller so we will try one more time before giving an estimate

$$N \geq \frac{8}{0.05^2} ln \frac{4(2(452172.8)^{10} + 4}{0.05} \tag{72}$$

$$N \geq \frac{8}{0.0025} ln \frac{4(904345.6)^{10} + 4}{0.05} \tag{73}$$

$$N \geq \frac{8}{0.0025} ln \frac{4(3.6588474e + 59) + 4}{0.05} \tag{74}$$

$$N \geq \frac{8}{0.0025} * 141.531 \tag{75}$$

$$N \geq 3200 * 141.532 \tag{76}$$

$$N \geq 452902.4 \tag{77}$$

$$\tag{78}$$

This answer is very similar to the last one, so we can now give an estimate for N, which is $N \approx 450,000$, meaning with a confidence of 95% and a generalization error of at most 0.05, we need a sample size around 450,000.

4   Exercise 3.1

In Exercise 3.1 we are given two semi-circles from Dr. Rivas's code and told to set $rad = 10, thk = 5, sep = 5$ for the exercise.

4.1   Part A.

In Part A, we are asked to run the PLA on the semi circles and plot our results.
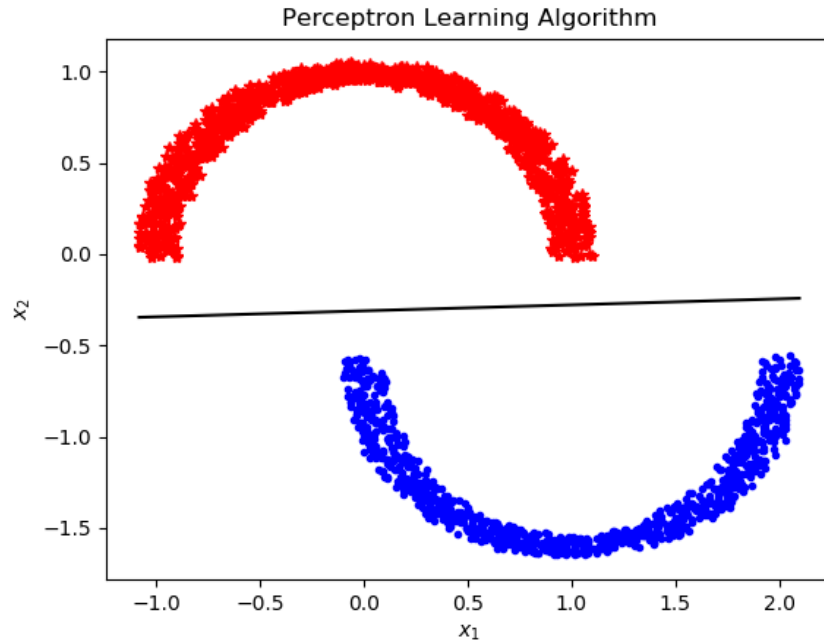
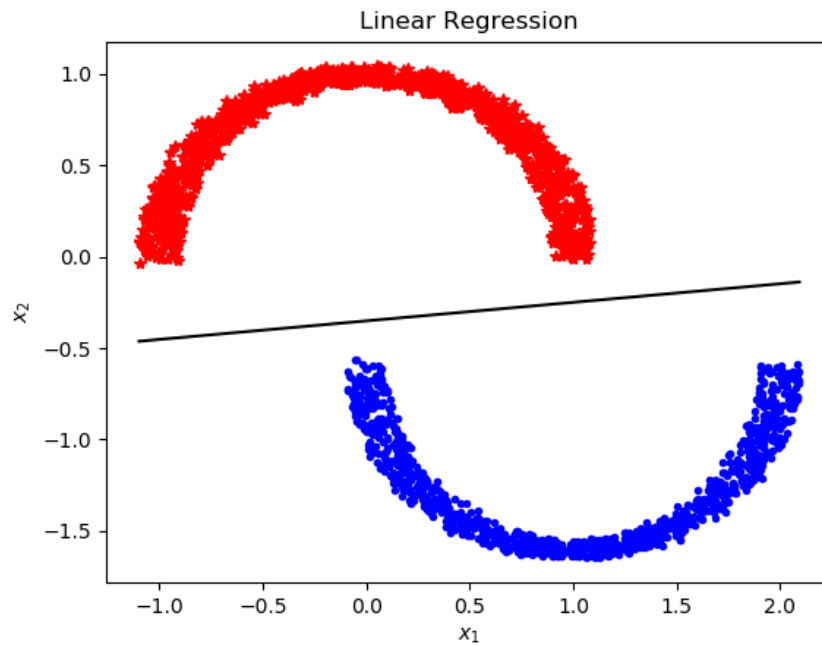Here are the misclassified points for the PLA.

```
2000
2000
253
253
16
16
238
238
99
99
0
```

4.2   Part B.

In Part B, we are asked to do the same thing but with linear regression, using the code given from the lab.

Here are the data points for the linear regression.

```
2000
2000
340
340
196
196
13
13
235
235
5
5
151
151
0
```

The difference in the PLA and Lin Reg is interesting. In the LinReg model, the closer inside parts of the semi circles are closer to the solution line that linear regression gives, and the ends on the farther end are farther apart from the solution, while on the PLA, all ends of the semicircles seem to be equally far apart from the PLA. Although the linear regression model seemed to have run a little longer, it seems the more accurate of the two.