CMPT 404
Homework 2

Daniel Clark

September 27, 2016

2.1 In Equation (2.1), set $\delta = 0.03$ and let

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

2.1 (a) For M = 1, how many examples do we need to make $\epsilon \leq 0.05$?

First we will solve the equation to give N as a function of M.

$$\begin{aligned} \epsilon &= \sqrt{\frac{1}{2N} ln \frac{2M}{\delta}} \\ \epsilon^2 &= \frac{1}{2N} ln \frac{2M}{\delta} \\ N &= \frac{1}{2\epsilon} ln \frac{2M}{\delta} \end{aligned}$$

Knowing the values of ϵ and δ we can get that

$$N(M) = 10 ln \frac{200M}{3}$$

Plugging in $M = 1$ we find that $N(1) \approx 41.997$ so we need 42 samples.

2.1 (b) $For M = 100$, how many examples do we need to make $\epsilon \leq 0.05$?

Using the formula derived in part (a) we find that $N(100) \approx 88.049$ so we need 89 samples.

2.1 (c) For $M = 10000$, how many examples do we need to make $\epsilon \leq 0.05$?

Using the formula derived in part (a) we find that $N(10000) \approx 134.100$ so we need 135 samples.

2.11 Suppose $m_H(N) = N + 1$, so $d_{vc} = 1$. You have 100 training examples. Use the generalization bound to give a bound for E_{out} with confidence 90%. Repeat for $N = 10000$.

Using equation (2.12), we can give a bound for E_{out} in terms of E_{in}.

$$E_{out}(g) \leq E_{in}(g) + \sqrt{\frac{8}{N} ln \frac{4m_H(2N)}{\delta}}$$
$$E_{out}(g) \leq E_{in}(g) + \sqrt{\frac{8}{100} ln \frac{4m_H(200)}{0.1}}$$
$$E_{out}(g) \leq E_{in}(g) + \sqrt{\frac{2}{25} ln 8040}$$
$$E_{out}(g) \leq E_{in}(g) + 0.8482$$

This means that the real error will be within .8482 of the observed error with a probability of 90%. When this is repeated with $N = 10000$, we find

$$E_{out}(g) \leq E_{in}(g) + \sqrt{\frac{8}{10000} ln \frac{4m_H(20000)}{0.1}}$$
$$E_{out}(g) \leq E_{in}(g) + \sqrt{\frac{1}{1250} ln 800040}$$
$$E_{out}(g) \leq E_{in}(g) + 0.1043$$

This means that the real error will be within 0.1043 of the observed error with a probability of 90%.

2.12 For an H with $d_{vc} = 10$, what sample size do you need (as prescribed by the generalization bound) to have a 95% confidence that your generalization error is a most 0.05?

Since we are working within a 95% confidence interval, we know that $\delta = 0.05$. Additionally, out generalization error is $\epsilon = 0.05$. Using equation (2.13) we get that

$$N \geq \frac{8}{\epsilon^2} ln \frac{4((2N)^{d_{vc}} + 1)}{\delta}$$
$$N \geq \frac{8}{0.05^2} ln \frac{4((2N)^{10} + 1)}{0.05}$$
$$N \geq 3200 ln 80((2N)^{10} + 1)$$

Starting with an initial guess of $N = 10000$, we iterate until N converges on $N = 452956.895$ so we need 452957 samples.

Problem 3.1 follows