Shon Cortes
ENPM808A
Homework 2

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

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

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

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

(c) For $M = 10,000$, how many examples do we need to make $\epsilon \leq 0.05$?

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

$$\epsilon^2 = \frac{1}{2N} \ln\left(\frac{2M}{\delta}\right)$$

$$N = \frac{1}{2\epsilon^2} \ln\left(\frac{2M}{\delta}\right)$$

$\delta = 0.03$

$\epsilon \leq 0.05$

a) $M = 1$

$$N \geq \frac{1}{2(0.5)^2} \ln\left(\frac{2(1)}{(0.03)}\right)$$

$\boxed{N \geq 839.94 \text{ examples}}$

b) $M = 100$

$$N \geq \frac{1}{2(0.5)^2} \ln\left(\frac{2(100)}{(0.03)}\right)$$

$\boxed{N \geq 1760.96 \text{ examples}}$

c) $M = 10,000$ $N \geq \frac{1}{2(0.5)^2} \ln\left(\frac{2(100)}{(0.03)}\right)$

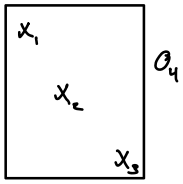
$\boxed{N \geq 2682.01 \text{ examples}}$

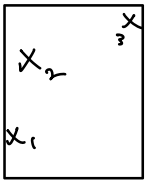
Problem 2.2 Show that for the learning model of positive rectangles (aligned horizontally or vertically), $m_{\mathcal{H}}(4) = 2^4$ and $m_{\mathcal{H}}(5) < 2^5$. Hence, give a bound for $m_{\mathcal{H}}(N)$.

$m_{\mathcal{H}}(N) \leq \sum_{i=0}^{k-1} \binom{N}{i}$ $\quad$ $m_{\mathcal{H}}(k) < 2^k$

$m_H(d_{vc}) < 2^d$

For $N=4$, $m_H(4) = 2^4$ $\quad$ For $N=5$, $m_H(5) < 2^5$





A rectangular model is defined by 4 parameters (4 corners of rectangle). Therefore $d_{vc} \geq 4$.

For $N=5$ points in R^4 there is some point that is linearly dependant on the rest. For example, there is no H (rectangle) that can classify all points inside the rectangle as positive except 1 being classified as negative.

- The $d_{vc} = 4$ for positive rectangular models, therefore $2^{d_{vc}} \leq m_H(d_{vc}) < 2^{d_{v+1}}$

Problem 2.11 Suppose $m_{\mathcal{H}}(N) = N + 1$, so $d_{\text{vc}} = 1$. You have 100 training examples. Use the generalization bound to give a bound for E_{out} with confidence 90%. Repeat for $N = 10,000$.

$m_H(N) = N+1 \qquad d_{vc} = 1 \qquad N = 100 \qquad \epsilon = 0.1 \qquad \delta = 0.1$

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

$$E_{out} \leq E_{in} + \sqrt{\frac{8}{N} \ln\left(\frac{4(2N)^{d_{vc}}+1}{\delta}\right)}$$

$$E_{out} \leq E_{in} + \sqrt{\frac{8}{100} \ln\left(\frac{4(2(100))^{(1)}+1}{(0.1)}\right)}$$

$E_{out} \leq E_{in} + 0.848 \quad$ if $N = 100$

$E_{out} \leq E_{in} + 0.104 \quad$ if $N = 10,000$

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

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

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

$d_{vc} = 10$

$\delta = 0.05$

$\epsilon \leq 0.05$

$N = ?$

Guess $N = 1000$

$\epsilon \geq 0.862$

$N = 10{,}000$

$\epsilon \geq 0.268$

$N = 600000$

$\epsilon \geq 0.0478$

$N = 460{,}000$

$\epsilon \geq 0.049$

Problem 2.23 Consider the learning problem in Example 2.8, where the input space is $\mathcal{X} = [-1, +1]$, the target function is $f(x) = \sin(\pi x)$, and the input probability distribution is uniform on $\mathcal{X}$. Assume that the training set $\mathcal{D}$ has only two data points (picked independently), and that the learning algorithm picks the hypothesis that minimizes the in-sample mean squared error. In this problem, we will dig deeper into this case.

For each of the following learning models, find (analytically or numerically) (i) the best hypothesis that approximates f in the mean-squared-error sense (assume that f is known for this part), (ii) the expected value (with respect to $\mathcal{D}$) of the hypothesis that the learning algorithm produces, and (iii) the expected out-of-sample error and its bias and var components.

(a) The learning model consists of all hypotheses of the form $h(x) = ax + b$ (if you need to deal with the infinitesimal-probability case of two identical data points, choose the hypothesis tangential to f).

(b) The learning model consists of all hypotheses of the form $h(x) = ax$. This case was not covered in Example 2.8.

(c) The learning model consists of all hypotheses of the form $h(x) = b$.

$f(x) = \sin(\pi x)$

$h_1(x) = ax + b$

$h_2(x) = ax$

$h_3(x) = b$

$h_1(x) = ax + b$ $\quad m = \frac{y_2 - y_1}{x_2 - x_1} = \hat{a}$

$\text{bias} = \frac{1}{|X|} \sum_{x \in X} (\bar{g}(x) - f(x))$

$\text{Variance} = \frac{1}{|X|K} \sum_{x \in X} \sum_{k=1}^{K} (\hat{a}^{(D_k)} x - \bar{g}(x))^2$

$y = mx + b$

$b = y - mx \Rightarrow \hat{b} = y_1 - \hat{a} x_1$, $b = y_2 - \hat{a} x_2$

$\bar{g}(x) = 0.792x$

bias = 0.21

variance = 0.76

$h_2(x) = ax$

$$E_{in} = \sum_{i=1}^{2} (y_i - ax_i)^2$$

$$\frac{\partial}{\partial a} E_{in} = -2 \sum_{i=1}^{2} x_i (y_i - ax_i) = 0$$

$$0 = \sum x_i y_i - \sum a x_i^2$$

$$x_1 y_1 + x_1 y_2 = a(x_1^2 + x_2^2)$$

$$\hat{a} = \frac{x_1 y_1 + x_1 y_2}{x_1^2 + x_2^2}$$

$$\bar{g}(x) = 1.446x$$

bias = 0.27
variance = 0.23

$h_3(x) = b$

$$\hat{b} = \frac{y_1 + y_2}{2}$$

$$\bar{g}(x) = 0.005x$$

bias = 0.5
Variance = 0.08

(6) To show that k is a break point for H

a) • Show a set of k points $x_1, , x_k$ which H can shatter.

b) • Show H can shatter any set of k points.

c) • Show a set of k points $x_1, , x_k$ which H cannot shatter.

d) • Show H cannot shatter any set of k points $x_1, , x_k$.

Select appropriate choice but explain all statements no matter your choice.

(7) Show that the VC dimension of

K is a break point if no data set of size K can be shattered by H.

Options (a) and (b) do not show K is a break point of H because they simply prove H CAN shatter a set or any set of points.

Options (c) and (d) show that K is a break point however (d) is considered overkill.

(7) Show that the VC dimension of a perceptron in R^d is $d+1$

$d_{vc} = d+1$ for perceptron in R^d

- There must exist a set of N points that can be shattered by H for $d_{vc} \geq N$ (d_{vc} is at least N)
- There is no set of N points that can be shattered by H for $d_{vc} < N$ (d_{vc} is at most N)

$X = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_{d+1} \end{bmatrix} = \begin{bmatrix} | \longrightarrow \\ | \longrightarrow \\ \vdots \longrightarrow \\ | \longrightarrow \end{bmatrix}$ X is invertible, $X \in R^{d+1}$ for $N = d+1$ points.

$y = \begin{bmatrix} \pm 1 \\ \vdots \\ \pm 1 \end{bmatrix}$ $w = \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_{d+1} \end{bmatrix}$

- if $Xw = y$ then we can classify all points correctly.

$w = X^{-1}y$, for this set of points we can generate a weight vector that properly classifies the data.

Therefore $d_{vc} \geq d+1$ where $N = d+1$

$d+2$ points

$X = \begin{bmatrix} x_1 & x_2 & x_3 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$

$X = \begin{bmatrix} x_1 \longrightarrow x_4 \\ 1 \; 0 \; 0 \; 1 \\ 0 \; 1 \; 0 \; 1 \\ 0 \; 0 \; 1 \; 1 \end{bmatrix}$ $x_1 \rightarrow x_4$ linearly dependent

x_1, x_2, x_3 are linearly independent

$\sum_{i=1}^{d+2} a_i x_i = 0$ $\quad a_j x_j = \sum_{i \neq j} a_i x_i$ $\quad$ if for x_j, $y_j = -1$ $\quad sig(a_i) = y_i$

$a_j w^T x_j = \sum_i a_i w^T x_i$

$a_j \underbrace{sig(w^T x_j)}_{\text{can be negative}} = \sum_i \underbrace{a_i \, sig(w^T x_i)}_{\text{always positive}}$ $\Rightarrow$ this is a set that can not be shattered for $N = d+2$ points

- For any $d+2$ points in a $d+1$ space will be linearly dependent. Therefore $d_{vc} < d+2$ where $N = d+2$

If $d+1 \leq d_{vc} < d+2$ then $d_{vc} = d+1$ for a perceptron in R^d.