CS446 / ECE 449: Machine Learning, Fall 2020, Homework 3

Name: Saud Alrasheed (Sauda2)

Worked individually

Problem (2):

Problem 2.1:

It is clear that for any two distinct data points $x_1, x_2 \in \mathbf{R}$, $s.t \quad x_1 < x_2$. We will separate the data for any given label configuration. Suppose both x_1, x_2 have label 0. Then, the classifier $f(x) = x - (x_2 + 1)$ will perfectly separate the data. Now, suppose both x_1, x_2 have label 1, then the classifier $f(x) = -x + (x_2 + 1)$ will perfectly separate the data. The remaining two cases are (x_1 has a label 0, x_2 has a label 1) and (x_1 has a label 1, x_2 has a label 0). Then, the classifiers $f(x) = x - \frac{(x_1+x_2)}{2}$ and $f(x) = -x + \frac{(x_1+x_2)}{2}$ perfectly classify the remaining two cases respectively.

Now, consider three distinct data points $x_1, x_2, x_3 \in \mathbf{R}$, $s.t\ x_1 < x_2 < x_3$. Also, consider the label configuration: (x_1, x_3 has label 1) and (x_2 has label 0). Thus, for any linear classifier that perfectly classifies any two points, it fails to classify the third (That's a consequence of linearity, proof is trivial).

Thus, $VC(\mathcal{F}_{affline}) = 2$.

Problem 2.2:

We will start by showing that we can shatter $k+1$ points. We begin by picking set of pints X. We will reflect the bias term in both the set of data and weights. Thus,

$$X = \begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ 0 & 0 & \ddots & 0 \\ \vdots & & & \vdots \\ 1 & 1 & \dots & 1 \end{bmatrix} = \begin{bmatrix} & & & 0 \\ & I_k & & 0 \\ & & & \vdots \\ 1 & 1 & \dots & 1 \end{bmatrix}$$

Here the last row reflects the bias term $X \in \mathbf{R}^{k+1} \times \mathbf{R}^{k+1}$, thus,

$$w^T x_i + b = (\begin{bmatrix} w_1 & w_2 & \dots & w_k & b \end{bmatrix} X)_i$$

Now, one can notice that X is invertible (Proof is trivial). So, we know that $y = sgn(w^T X)$, and by the invertibality of X, we have that, $w^T = yX^{-1}$. Therefore, for any label configura-

tion, there exist a set of weights that perfectly classifies the data points.

Now, we show that our class of linear functions can't shatter $k+2$ data points. Again, let X be the matrix of data points with last row of ones.

$$w^T x_i + b = (\begin{bmatrix} w_1 & w_2 & \dots & w_k & b \end{bmatrix} X)_i$$

Here, it's clear that $X \in \mathbf{R}^{k+1} \times \mathbf{R}^{k+2}$. Thus, X consist of at least one column v_i that is linearly dependent. So, we have that,

$$v_i = \sum_{i \neq j} a_j v_j$$

for some $a_j \neq 0$. Thus, it is clear that the label assignment of v_i depends on $\sum_{i \neq j} a_j v_j$. So, we cant perfectly classify all possible label configurations for any $k+2$ data points. Thus, $VC(\mathcal{F}_{affline}) = k+1$.

Problem 2.3:

We will show that there always exist a set of k that can be shattered by $\mathcal{F}$. Let X be the set of data points such that, $X = I_k$, where I is the identity matrix. Then, simply consider the naive decision tree with following split decisions, $\{x_1 \leq 0, x_2 \leq 0, x_3 \leq 0, \dots, x_k \leq 0\}$. Here, it is clear that after the k decision splits we have that all leafs are pure. Thus, the tree shatter k data points. Since our choice of k is arbitrary, it follows that decision trees have infinite VC dimension.