Homework #1

(Due: Mar 15)

1 Theory

1.1 A "warm up" problem

Since the given distribution is uniform in the range $[-1, 1]$, $Pr(x > 0) = 0.5$ and $Pr(x \leq 0) = 0.5$
When $x > 0$, the classifier(h) labels it 1 but the actual labelling function has 0.1 fraction of data samples marked -1.Therefore the training error when $x > 0$ equals 0.1
When $x \leq 0$, the classifier(h) labels it -1 but the actual labelling function has 0.1 fraction of data samples marked 1.Therefore the training error when $x \leq 0$ equals 0.1
Therefore total training error for h $= (0.1 + 0.1)/2 = 0.1$

1.2 Bayes Optimal Predictor

For a Bayes optimal Predictor and given distribution D over X to {0,1}, we can write

$$f_D(x) = \begin{cases} 1 & \text{if } Pr[Y = 1|x] \geq 1/2 \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

Let $Pr[Y = 1|x] = t$

Therefore,

$$\begin{aligned} Pr[f_D(x) \neq Y|X = x] &= 1_{[t \geq 0.5]} Pr[Y = 0|X = x] + 1_{[t<0.5]} Pr[Y = 1|X = x] \\ &= 1_{[t \geq 0.5]}(1 - t) + 1_{[t<0.5]} t \\ &= \min(t, 1 - t) \end{aligned} \tag{2}$$

Let g be a classifier from X to {0, 1}. We have

$$\begin{aligned} Pr[g(X) \neq Y|X = x] &= Pr[g(X) = 1|X = x]Pr[Y = 0|X = x] + Pr[g(X) = 0|X = x]Pr[Y = 1|X = x] \\ &= Pr[g(X) = 1|X = x](1 - t) + Pr[g(X) = 0|X = x]t \\ &\geq Pr[g(X) = 1|X = x]\min(t, 1 - t) + Pr[g(X) = 0|X = x]\min(t, 1 - t) \\ &\geq \min(t, 1 - t) \end{aligned} \tag{3}$$

For 0-1 loss, expectation and probability have equal values, we can write expected loss for a classifier as

$$E[L_D(f_D)] = Pr[f_D(x) \neq Y|X = x] \text{ and } E[L_D(g)] = Pr[g(X) \neq Y|X = x] \tag{4}$$

Therefore combining the two results above we can write

$$E[L_D(f_D)] \leq E[L_D(g)] \tag{5}$$

1.3 Perceptron with a Learning Rate

The output label of a perceptron is given as sign($\langle w, X \rangle$).
Since the weights are now scaled by η, the output label of the modified perceptron will now be given as sign($\langle \eta w, X \rangle$). This however does not change the sign of the label, therefore the predictions are the same for both the perceptrons. As a result, they both perform the same number of iterations. In case of perceptron with a learning rate, the magnitude of the resultant vector gets scaled but the direction remains unchanged since the sign function output is the same for every misclassification in both cases.

1.4 Unidentical Distributions

For some classifier h $\in$ H, let $L_{(\bar{D}_m, f)}(h) > \epsilon$
Therefore we can write

$$\sum_{i=1}^{m} \frac{Pr_{X \sim D_i}[h(X) = f(X)]}{m} < 1 - \epsilon \tag{6}$$

$$\begin{aligned} Pr[L_S(h) = 0] &= \prod_{i=1}^{m} Pr_{X \sim D_i}[h(X) = f(X)] \\ &= \left(\left(\prod_{i=1}^{m} Pr_{X \sim D_i}[h(X) = f(X)] \right)^{\frac{1}{m}} \right)^m \end{aligned} \tag{7}$$

Applying the Arithmetic Geometric Mean Inequality we can write it as

$$\begin{aligned} Pr[L_S(h) = 0] &\leq \left(\frac{\sum_{i=1}^{m} Pr_{X \sim D_i}[h(X) = f(X)]\Big)}{m} \right)^m \\ &\leq (1 - \epsilon)^m \\ &\leq e^{-\epsilon m} \end{aligned} \tag{8}$$

Applying the union bound from probability we have that

$$\begin{aligned} Pr[L_{(\bar{D}_m, f)}(h) > \epsilon] &\leq \sum_{h \in H} Pr[L_S(h) = 0] \\ &\leq |H| e^{-\epsilon m} \end{aligned} \tag{9}$$

1.5 Vapnik-Chervonenkis (VC) Dimension

1.5.1 a

We know that the VCdim of the class of axis-aligned rectangles in 2 dimensions is 4. We now will extend this proof to n dimensions. Given parameters $a_1 \leq b_1, a_2 \leq b_2, .., a_d \leq b_d$, let h be a classifier such that

$$h_{a_1,b_1,...,a_d,b_d}(x_1, x_2, ...x_d) = \prod_{i=1}^{d} 1_{[x_i \in [a_i,b_i]]} \tag{10}$$

Therefore

$$H_d = \{h_{a_1,b_1,...,a_d,b_d}(x_1, x_2, ...x_d) : \forall i \in [d], a_i \leq b_i\} \tag{11}$$

Let there be set of data points $\{x_1, x_2, ...x_{2d}\}$ where

$$x_i = \begin{cases} e_i & \text{if } i \in [d] \\ -e_{(i-d)} & \text{if } i > d \end{cases} \tag{12}$$

Let $\{y_1, y_2, .., y_{2d}\} \in \{0, 1\}^{2d}$
Let $a_i = -2$ if $y_{i+d} = 1,$ and $a_i = 0$ otherwise.
Similarly $b_i = 2$ if $y_i = 1,$ and $b_i = 0$ otherwise.
Then

$$h_{a_1,b_1,...,a_d,b_d}(x_i) = y_i \ \forall i \in [2d] \tag{13}$$

Therefore VCdim(H^d) $\geq$ 2d.
We now show that a set of size (2d+1) cannot be shattered by H^d. By the pigeonhole principle, there exists an element x, such that for every $j \in [d]$, there exists another element v in the same set with $v_j \leq x_j$, and similarly there exists another element u in the same set with $u_j \geq x_j$. Thus the labeling in which x is negative, and the rest of the elements in the set are positive can not be obtained.
Therefore VCdim(H^d) $= 2d$.

1.5.2 b

We know that the sine function ranges in [-1,1], and so the hypothesis class

$$H = \{x \mapsto \lceil sin(\theta x) \rceil : \theta \in R\} \tag{14}$$

always outputs either 0 or 1. Consider n points (1..n) with point i in the range $[2(i-1)\pi, 2(i-1)\pi + 2\pi]$. We can give each such point labels 0 or 1. Consider the first point, if we keep it between $[0, \pi]$ we get label 1 and if we keep it between $[\pi, 2\pi]$ we get label 0. Since sine function is a repeating function with interval 2π radians, this holds true for all the n points. Therefore H shatters the n points.
Note that this is true for any n, as a result we don't have a n which H cannot shatter. Therefore we conclude VCdim(H) $= \infty$

1.6 Boosting

Let m be the number of instances in the training set.
Error of the current weak learner h_t on current distribution is

$$\epsilon_t = \sum_{i=1}^{m} D_i^t 1_{y_i \neq h_t(x_i)}$$

Let $w_t = \frac{1}{2} \ln \left(\frac{1}{\epsilon_t} - 1 \right)$
The update definition of the distribution gives the underlying equation

$$D_i^{(t+1)} = \frac{D_i^{(t)} e^{-w_t y_i h_t(x_i)}}{\sum_{j=1}^{m} D_j^{(t)} e^{-w_t y_j h_t(x_j)}} \tag{15}$$

Therefore applying summation on both sides we get,

$$\sum_{i=1}^{m} D_i^{(t+1)} 1_{y_i \neq h_t(x_i)} = \frac{\sum_{i=1}^{m} D_i^{(t)} e^{-w_t y_i h_t(x_i)} 1_{y_i \neq h_t(x_i)}}{\sum_{j=1}^{m} D_j^{(t)} e^{-w_t y_j h_t(x_j)}} \tag{16}$$

In case of incorrect predictions, $y_i h_t(x_i) = -1$
Therefore,

$$\sum_{i=1}^{m} D_i^{(t+1)} 1_{y_i \neq h_t(x_i)} = \frac{e^{w_t} \sum_{i=1}^{m} D_i^{(t)} 1_{y_i \neq h_t(x_i)}}{\sum_{j=1}^{m} D_j^{(t)} e^{-w_t y_j h_t(x_j)}} \tag{17}$$

$$\sum_{i=1}^{m} D_i^{(t+1)} 1_{y_i \neq h_t(x_i)} = \frac{e^{w_t} \sum_{i=1}^{m} D_i^{(t)} 1_{y_i \neq h_t(x_i)}}{e^{w_t} \sum_{j=1}^{m} D_j^{(t)} 1_{y_j \neq h_t(x_j)} + e^{-w_t} (1 - \sum_{j=1}^{m} D_j^{(t)} 1_{y_j \neq h_t(x_j)})} \tag{18}$$

$$\sum_{i=1}^{m} D_i^{(t+1)} 1_{y_i \neq h_t(x_i)} = \frac{e^{w_t} \epsilon_t}{e^{w_t} \epsilon_t + e^{-w_t} (1 - \epsilon_t)} \tag{19}$$

$$\sum_{i=1}^{m} D_i^{(t+1)} 1_{y_i \neq h_t(x_i)} = \frac{1}{\left(1 + \frac{1 - \epsilon_t}{e^{2 w_t} \epsilon_t} \right)} \tag{20}$$

$$\sum_{i=1}^{m} D_i^{(t+1)} 1_{y_i \neq h_t(x_i)} = \frac{1}{\left(1 + \frac{1 - \epsilon_t}{1 - \epsilon_t} \right)} \tag{21}$$

$$\sum_{i=1}^{m} D_i^{(t+1)} 1_{y_i \neq h_t(x_i)} = \frac{1}{2} \tag{22}$$

Therefore we see that the error of the current weak learner h_t on the next iteration's distribution $D^{(t+1)}$ is exactly 0.5.

1.7 Cross-validation

Since the $Pr[y = 1] = Pr[y = 0] = 0.5$ and the classifier is a constant predictor(for all instances it either predicts 0 or 1) we know that the true error will always be 0.5.
Lets consider 2 cases for LOO(leave-one-out) estimate.

- The number of 1s in the training set labels is odd.

  It means the training set labels has even number of 1s on test labels with label 1 and so the classifier labels them all incorrectly, and the training set has odd number of 1s on the test labels with label 0 and so the classifier labels them all incorrectly as well. Therefore the average error is 1.

- The number of 1s in the training set labels is even.

  It means the training set labels has odd number of 1s on test labels with label 1 and so the classifier labels them all correctly, and the training set has even number of 1s on the test labels with label 0 and so the classifier labels them all correctly as well. Therefore the average error is 0.

In both cases the absolute difference between the true error and LOO estimate turns out to be exactly 0.5