Siyuan Zou 111639762 CSE353 Hw1

Theory Portion

I. The training error is simply 0.1 due the probability of the distribution. If x > 0 there is a 0.9 chance of the random sample being correctly labeled to 1, and if x ≤ 0 there is a 0.9 chance of the random sample being labeled as 0. From this we know the true error is also 0.1 as h(x)≠f(x) of a random instance of x has a 0.1 chance to be miss classified.

II. For this problem our goal is to proof that no other classifier, $g: X \rightarrow \{0,1\}$ has a lower error. That is, for every classifier $L_D(f_D) \leq L_D(g)$

We know that the true error is: $L_D(h) = \mathbb{P}_{X\sim D}[h(x) \neq y]$

In this case of $g: X \rightarrow \{0,1\}$, $L_D(h) = \begin{cases} \Pr[y \neq 0|x] & if(h(x) = 0 \\ \Pr[y \neq 1|x] & if(h(x) = 1 \end{cases}$

Since the function consist of probability, we can perform some changes:

$$\Pr[y \neq 0|x] \quad if(h(x) = 0 \equiv \Pr[y = 1|x] \quad if(h(x) = 0$$
$$\Pr[y \neq 1|x] \quad if(h(x) = 1 \equiv 1 - \Pr[y = 1|x] \quad if(h(x) = 1$$
$$E(X) = \begin{cases} \Pr[y = 1|x] & if(h(x) = 0 \\ 1 - \Pr[y = 1|x] & if(h(x) = 1 \end{cases}$$

Which means that if $\Pr[y = 1|x] < 1 - \Pr[y = 1|x]$, we should choose $h(x) = 0$ to minimize loss or error, and similarly we choose $h(x) = 1$ if $\Pr[y = 1|x] > 1 - \Pr[y = 1|x]$. And if they are equal well the choice doesn't matter at that point.

Rearranging the first equation we can get $2\Pr[y = 1|x] < 1 \; or \; \Pr[y = 1|x] < 1/2$

Same equation as the Bayes optimal predictor.

III. a) The change of updating step doesn't not affect the outer loop that dictates the number of iterations that the perceptron goes through. Hence changing $\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} + y_i\mathbf{x}_i$ to $\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} + \eta y_i\mathbf{x}_i$ will perform the same amount of iterations.

b) The direction of the output $\mathbf{w}^{(t)}$ will not be affected since $0 < \eta < 1$ this means that the vector $\mathbf{w}^{(t)} + \eta y_i\mathbf{x}_i$ is only adding on a portion of what is adding before in the original perceptron algorithm. This is important because as the case of if $\mathbf{w}^{(t)}$ is a positive vector ie [6, 5] and $y_i\mathbf{x}_i$ is a negative vector, ie [-2, -2] the direction will not flip due to adding a large multiple of $y_i\mathbf{x}_i$ due to a large constant η. ie let say constant $\eta = 5$, then $\mathbf{w}^{(t)} + \eta y_i\mathbf{x}_i = [-4, -5]$ which has a different direction than original. But due to this $0 < \eta < 1$ constraint this will not happen.

IV. First let us denote some variables:

Let A be the event of $\exists h \in \mathcal{H}$ s.t. $L_{(\overline{\mathcal{D}}_{m,f})}(h) > \epsilon$, and B be the event of $L_{(S,f)}(h) = 0$

Now we write the probability with A and B:

$$Pr(A \; and \; B)$$

Apply conditional probability:

$$Pr(A \; and \; B) = Pr(B|A)\, Pr(A)$$

We are applying the union bound rule:

Lets first consider the $Pr(A)$. Since the true error in corresponds with the mean distribution $\overline{D}_m = \frac{(D_1 + \cdots + D_m)}{m}$ and that $\exists h \in \mathcal{H}$, thus the $D_m\left(L_{(\overline{\mathcal{D}}_{m,f})}(h) > \epsilon\right) \leq |\mathcal{H}|$

Lets then consider the $Pr(B)$. The condition of B states that there is no training error respect to the function f. This means that for every single i^{th} sample in the sample set $h(x_i) = f(x_i)$. We know that $D(\{x_i : \; h(x_i) = \; y_i\}) = 1 \; - \; L_{(D,f)}(h) \; \leq \; 1 \; - \epsilon$ and since we have m distribution, we can combine this two knowledge and form an equation: $D_m(\{S|x : \; L_S(h) = \; 0\}) \leq (1 \; - \epsilon)^m \; \leq \; \epsilon^{-em}$

We combine the A and B we get that $\Pr\left(\exists h \in \mathcal{H} \; s.t. L_{(\overline{\mathcal{D}}_{m,f})}(h) > \epsilon \; and \; L_{(S,f)}(h) = 0\right) \leq |\mathcal{H}|\epsilon^{-em}$