Solutions for Homework 3

Bao Jinge A0214306U e0522065@u.nus.edu

1 Perceptron Algorithm as Online Convex Optimization

1.1 a

Since online gradient descent algorithm has equations as follows

$$w^{(t+1)} = w^{(t)} - \eta \nabla_{w^{(t)}} l(w^{(t)}, x_t, y_t) \tag{1}$$

When the predication is correct,

$$\nabla_{w^{(t)}} l(w^{(t)}, x_t, y_t) = \nabla_{w^{(t)}} 0 = 0 \tag{2}$$

When the predication is incorrect,

$$\nabla_{w^{(t)}} l(w^{(t)}, x_t, y_t) = \nabla_{w^{(t)}} max\{0, 1 - y_t \left\langle w^{(t)}, x_t \right\rangle\} = -y_t x_t \tag{3}$$

Plug Equation 2 and 3 into Equation 1 correspondingly, we find that

$$w^{(t+1)} = \begin{cases} w^{(t)} & \text{predication is correct} \\ w^{(t)} + \eta y_x x_t & \text{prediction is incorrect} \end{cases} \tag{4}$$

As we can see, with such loss function mentioned in problem, the perception algorithm in only doing online gradient descent when prediciton is correct.

1.2 b

The total loss of this algorithm is the expression as follows

$$\sum_{t=1}^{T} l(w^{(t)}, x_t, t_t) \tag{5}$$

where T is the total rounds. Let X_t is a indicator, which is equal to 0 when prediction is correct and 1 when prediction is incorrect in t round. Thus, the total number of mistakes is

$$\sum_{t=1}^{T} X_t \tag{6}$$

When prediction is correct,

$$l(w^{(t)}, x_t, t_t) = 0 \geq 0 = X_t \tag{7}$$

When prediction is incorrect, $y_t \left\langle w^{(t)}, x_t \right\rangle < 0$, thus

$$l(w^{(t)}, x_t, t_t) = max\{0, 1 - y_t \left\langle w^{(t)}, x_t \right\rangle\} = 1 - y_t \left\langle w^{(t)}, x_t \right\rangle \geq 1 \geq X_t \tag{8}$$

Plug Equation 6, 7 and 8 into Equation 5, we get such bound

$$\sum_{t=1}^{T} l(w^{(t)}, x_t, t_t) \geq \sum_{t=1}^{T} X_t \tag{9}$$

which means the total loss upper bounds the total number of mistakes.

1.3 c

Because there exists w^* such that $y_t \langle w^*, x_t \rangle$, with definition, the prediction is always right. Therefore, from Equation 4, the weight will always not be updated. From definition of defintion of problem (a), w^* will have zero total loss, i.e.

$$\sum_{t=1}^{T} l(w^*, x_t, t_t) = 0 \tag{10}$$

1.4 d

According to SSBD Lemma 14.1, let $v_t = \nabla_{w^{(t)}} l(w^{(t)}, x_t, y_t)$, we get

$$\sum_{t=1}^{T} \left\langle w^{(t)} - w^*, \nabla_{w^{(t)}} l(w^{(t)}, x_t, y_t) \right\rangle \leq \frac{||w^*||^2}{2\eta} + \frac{\eta}{2} \sum_{t=1}^{T} ||\nabla_{w^{(t)}} l(w^{(t)}, x_t, y_t)||^2 \tag{11}$$

From Equation 2 and 3, we know that when prediction is incorrect,

$$\nabla_{w^{(t)}} l(w^{(t)}, x_t, y_t)||^2 = ||x_t y_t||^2 = ||x_t||^2 \leq R^2 \tag{12}$$

When prediction is correct,

$$\nabla_{w^{(t)}} l(w^{(t)}, x_t, y_t)||^2 = ||0||^2 = 0 \tag{13}$$

From what is given by part (d), there are M mistakes. Thus,

$$\sum_{t=1}^{T} \left\langle w^{(t)} - w^*, \nabla_{w^{(t)}} l(w^{(t)}, x_t, y_t) \right\rangle \leq \frac{||w^*||^2}{2\eta} + \frac{\eta}{2} M R^2 \tag{14}$$

For lhs of Equation 14, we use property of convexity function, i.e.

$$l(w^{(t)}, x_t, y_t) - l(w^*, x_t, y_t) \leq l \left\langle w^{(t)} - w^*, \nabla_{w^{(t)}} l(w^{(t)}, x_t, y_{)} \right\rangle \tag{15}$$

Plug Equation 15 into 14, we get

$$\sum_{t=1}^{T} (l(w^{(t)}, x_t, y_t) - l(w^*, x_t, y_t)) \leq \frac{||w^*||^2}{2\eta} + \frac{\eta}{2} M R^2 \tag{16}$$

Because of result from part (c), i.e. Equation 10, we have

$$\sum_{t=1}^{T} l(w^{(t)}, x_t, y_t) \leq \frac{||w^*||^2}{2\eta} + \frac{\eta}{2} M R^2 \tag{17}$$

Because of result from part (b), i.e.

$$\sum_{t=1}^{T} l(w^{(t)}, x_t, y_t) \geq M \tag{18}$$

Plug Equation 18 into Equation 17, we habe

$$M \leq \frac{||w^*||^2}{2\eta} + \frac{\eta}{2} M R^2 \tag{19}$$

Here we set $\eta = \frac{||w^*||}{R\sqrt{M}}$, we get

$$M \leq ||w^*||^2 R\sqrt{M} \tag{20}$$

, i.e.

$$M \leq ||w^*||^2 R^2 \tag{21}$$

1.5 e

For each round t,

$$\begin{aligned} sign(\left\langle w^{(t)}, x_i \right\rangle) &= sign(\left\langle \sum_{i=1}^{t-1} \eta y_i x_i, x_t \right\rangle) \\ &= sign(\left\langle \eta \sum_{i=1}^{t-1} y_i x_i, x_t \right\rangle) \\ &= sign(\eta \left\langle \sum_{i=1}^{t-1} y_i x_i, x_t \right\rangle) \\ &= sign(\eta) * sign(\left\langle \sum_{i=1}^{t-1} y_i x_i, x_t \right\rangle) \end{aligned} \tag{22}$$

As what we can see as above, as long as $\eta > 0$, then the predication will not be affected, because $sign(\eta)$ will always be 1. Therefore, the mistake bound still holds when η set to 1.

2 Switching Predictors

2.1 a

Suppose there will be M mistakes during T rounds. As what is given by the problem, we predict using the label which agrees with the weighted majority of surviving hyphotheses. In other way, when mistake happens, there must be greater than half of weights of surviving hypotheses whose hypotheses did wrong prediction. Formall, suppose after t round, there will be $h \in H_t$ remaining and initializing $H_0 = H$. From analysis above, when prediction is incorrect

$$\sum_{h \in H_{t+1}} w_h \leq \frac{1}{2} \sum_{h \in H_t} w_h \tag{23}$$

When predication is right,

$$\sum_{h \in H_{t+1}} w_h = \sum_{h \in H_t} w_h \tag{24}$$

Therefore,

$$\sum_{h \in H_T} w_h \leq (\frac{1}{2})^M \sum_{h \in H_0} w_h = (\frac{1}{2})^M \tag{25}$$

Since $h*$ is a predictor that does not make any error, $h* \in H_T$. Plug it into Equation 25,

$$w_{h^*} \leq \sum_{h \in H_T} w_h \leq (\frac{1}{2})^M \tag{26}$$

i.e.

$$M \leq \log \frac{1}{w_{h^*}} \tag{27}$$

2.2 b

2.2.1 i

When $T = 1$, the k-switching predictor will have just one hypothesis, i.e. $k = 0$. When $k = 0$, as what is given by definition, the weight for each 0-switching predictor is

$$\frac{1}{|H|(1 - |H|)^0} p^0 (1-p)^{1-0-1} = \frac{1}{|H|} \tag{28}$$

Obviously, there will be $|H|$ such 0-switching predictor. Therefore, the sum of the weights of all such 0-switching predictors is 1.

2.2.2 ii

Denote the predictor sequence of length t as s_t, the number of switches in s_t as $k(s_t)$, the set of all s_t as S_t and the weight of each predictor sequence is $w(s_t)$. As an inductive basis from (i), we know that when predictor sequences of lenght $t = T - 1$, we have

$$\sum_{s_{T-1} \in S_{T-1}} \frac{1}{|H|(|H|-1)^{k(s_{T-1})}} p^{k(s_{T-1})} (1-p)^{T-2-k(s_{T-1})} = 1 \tag{29}$$

When $t = T$ we focus on the last hypthosis that will be pushed back to s_{T-1}. There will be two cases. Suppose the last hypothesis in each predictor sequence of s_{T-1} is h_{T-1}. For the T-th sample, if we choose the hypothesis h_{T-1} to append, then we have $k(s_T) = k(s_{T-1})$. Denote this event as E_1, the sum of weight that in such case is

$$\begin{aligned}
&\sum_{s_T \in S_T : E_1} \frac{1}{|H|(|H|-1)^{k(s_T)}} p^{k(s_T)} (1-p)^{T-1-k(s_T)} \\
=&\sum_{s_{T-1} \in S_{T-1}} \frac{1}{|H|(|H|-1)^{k(s_{T-1})}} p^{k(s_{T-1})} (1-p)^{T-1-k(s_{T-1})} \\
=&\sum_{s_{T-1} \in S_{T-1}} \frac{1}{|H|(|H|-1)^{k(s_{T-1})}} p^{k(s_{T-1})} (1-p)^{T-2-k(s_{T-1})} * (1-p) \\
=&1-p
\end{aligned} \tag{30}$$

if we choose to append a hypothesis which is not h_{T-1}, then we have $k_{s_T} = k_{s_{T-1}} + 1$ and there will be $|H| - 1$ choices from H. Denote this event as E_2, the sum of weight that in such case is

$$\begin{aligned}
&\sum_{s_T \in S_T : E_2} \frac{1}{|H|(|H|-1)^{k(s_T)}} p^{k(s_T)} (1-p)^{T-1-k(s_T)} \\
=&\sum_{s_{T-1} \in S_{T-1}} \frac{1}{|H|(|H|-1)^{k(s_{T-1})+1}} p^{k(s_{T-1})+1} (1-p)^{T-1-(k(s_{T-1})+1)} * (|H|-1) \\
=&\sum_{s_{T-1} \in S_{T-1}} \frac{1}{|H|(|H|-1)^{k(s_{T-1})}} p^{k(s_{T-1})} (1-p)^{T-2-k(s_{T-1})} * p \\
=&p
\end{aligned} \tag{31}$$

Since $E1$ and $E2$ are involed all cases that will happen, thus

$$\begin{aligned}
&\sum_{s_T \in S_T} \frac{1}{|H|(|H|-1)^{k(s_T)}} p^{k(s_T)} (1-p)^{T-1-k(s_T)} \\
=&\sum_{s_T \in S_T : E1} \frac{1}{|H|(|H|-1)^{k(s_T)}} p^{k(s_T)} (1-p)^{T-1-k(s_T)} + \sum_{s_T \in S_T : E@} \frac{1}{|H|(|H|-1)^{k(s_T)}} p^{k(s_T)} (1-p)^{T-1-k(s_T)} \\
=&(1-p) + p \\
=&1
\end{aligned} \tag{32}$$

Thus the statement is true for predictor sequence of length T.

2.3 c

From problem (b), we know that each k-switching predictor has weight

$$\frac{1}{|H|(|H|-1)^k} p^k (1-p)^{T-1-k} \tag{33}$$

Plug equation above into Equation 27, that give us

$$M \leq \log \frac{1}{w_{h^*}} = \log |H| + k \log(|H|-1) + k \log(1/p) + (T-k-1) \log(1/(1-p)) \tag{34}$$

2.4 d

Focus on the inductive proof in part (b). What we focus on is the last hypothesis of the sequence. Suppose now the last hypothesis is h^* and we need to predict for x_t and y_t. We consider sum of the weights of all sequence ending with h^* as w_{h^*}. If the prediction is wrong, then we can not use this sequence any more. Thus we set the weight of $w_{h^*} = 0$. If the prediction is right, then we can keep use h^*, which means append h^* to sequence that ends with h^* and sequence that does not end with h^*. From inducation from part (b), we know the weight updated rule for the former is $w_{h^*}^{(t)} = w_{h^*}^{(t-1)} * (1-p)$. the wight updeted rule for the latter is $w_{h^*}^{(t)} = \sum_{h,h \neq h*} w_h^{(t-1)} * \frac{p}{|H|-1}$. As we can see, the update in each iteration is $O(|H|)$. The specified algorithm is as follows

Algorithm 1 K-swtiching Predicator Modified Halving Algorithm

Input: a finite class H of expert binary predictors, T predication sample of pair (x_t, y_t)
Output: k-switching predication

```
for all $h = 1$ to $H$ do
    $w_h = \frac{1}{|H|}$
end for
for all $t = 1$ to $T$ do
    $\hat{y} = \arg\max_{r \in \{0,1\}} \sum_{h \in H, h(x_t) = r} w_h$
    Output the prediction $\hat{y}$
    $w\_sum = 0$
    for all $h = 1$ to $H$ do
        $w\_sum = w\_sum + w_h$
    end for
    for all $h = 1$ to $H$ do
        if $h(x_t) \neq y_t$ then
            $w_h = 0$
        else
            $w_h \leftarrow w_h * (1-p) + (w\_sum - w_h) * \frac{p}{|H|-1}$
        end if
    end for
end for
```

As we can see as above, the running time for each iteracion is $O(|H|)$