CMSC 35900 (Spring 2008) Learning Theory **Lecture: 2**

Perceptron and Winnow

Instructors: Sham Kakade and Ambuj Tewari

1 The Perceptron Algorithm

Algorithm 1 PERCEPTRON

$w_1 \leftarrow \mathbf{0}$
for $t = 1$ to T **do**
  Receive $x_t \in \mathbb{R}^d$
  Predict $\mathrm{sgn}(w_t \cdot x_t)$
  Receive $y_t \in \{-1, +1\}$
  if $\mathrm{sgn}(w_t \cdot x_t) \neq y_t$ **then**
    $w_{t+1} \leftarrow w_t + y_t x_t$
  else
    $w_{t+1} \leftarrow w_t$
  end if
end for

The following theorem gives a dimension independent bound on the number of mistakes the PERCEPTRON algorithm makes.

Theorem 1.1. *Suppose Assumption M holds. Let*

$$M_T := \sum_{t=1}^{T} \mathbf{1}\left[\mathrm{sgn}(w_t \cdot x_t) \neq y_t\right]$$

denote the number of mistakes the PERCEPTRON *algorithm makes. Then we have,*

$$M_T \leq \frac{\|x_{1:T}\|^2 \cdot \|w^*\|^2}{\gamma^2} \ .$$

Proof. The key idea of the proof is to look at how the quantity $w^* \cdot w_t$ evolves over time. We first provide an lower bound for it. Define $m_t = \mathbf{1}\left[\mathrm{sgn}(w_t \cdot x_t) \neq y_t\right]$. Note that $w_{t+1} = w_t + y_t x_t m_t$ and $M_T = \sum_t m_t$. We have,

$$\begin{aligned} w^* \cdot w_{t+1} &= w^* \cdot w_t + y_t x_t m_t \\ &= w^* \cdot w_t + y_t (w^* \cdot x_t) m_t \\ &\geq w^* \cdot w_t + \gamma m_t \ . && \text{(Assumption M)} \end{aligned}$$

Unwinding the recursion, we get

$$w^* \cdot w_{T+1} \geq w^* \cdot w_1 + \gamma M_T = \gamma M_T \ . \tag{1}$$

Now, we use Cauchy-Schwarz inequality to get the upper bound,

$$w^* \cdot w_{T+1} \leq \|w^*\| \cdot \|w_{T+1}\| \ . \tag{2}$$

Moreover,

$$\begin{aligned}\|w_{t+1}\|^2 &= \|w_t + y_t x_t m_t\|^2 \\ &= \|w_t\|^2 + 2y_t(w_t \cdot x_t)m_t + \|x_t\|^2 m_t \\ &\le \|w_t\|^2 + 0 + \|x_{1:T}\|^2 m_t \ ,\end{aligned}$$

where the last step follows because $y_t(w_t \cdot x_t) < 0$ when a mistake is made and $\|x_t\| \le \|x_{1:T}\|$. Unwinding the recursion once again, we get,

$$\|w_{T+1}\|^2 \le \|w_1\|^2 + \|x_{1:T}\|^2 M_T = \|x_{1:T}\|^2 M_T \ . \tag{3}$$

Combining (1), (2) and (3) gives,

$$\gamma M_T \le w^* \cdot w_{T+1} \le \|w^*\| \cdot \|w_{T+1}\| \le \|w^*\| \cdot \|x_{1:T}\| \sqrt{M_T} \ .$$

This implies that $M_T \le \|w^*\|^2 \cdot \|x_{1:T}\|^2/\gamma^2$. □

2 Lower Bound

Theorem 2.1. *Suppose $\mathcal{X} = \left\{x \in \mathbb{R}^d \,\middle|\, \|x\| \le 1\right\}$ and $\frac{1}{\gamma^2} \le d$. Then for any deterministic algorithm, there exists a data set which is separable by a margin of γ on which the algorithm makes at least $\lfloor \frac{1}{\gamma^2} \rfloor$ mistakes.*

Proof. Let $n = \lfloor \frac{1}{\gamma^2} \rfloor$. Note that $n \le d$ and $\gamma^2 n \le 1$. Let $\mathbf{e}_i$ be the unit vector with a 1 in the ith coordinate and zeroes in others. Consider $\mathbf{e}_1, \ldots, \mathbf{e}_n$. We now claim that, for any $b \in \{-1, +1\}^n$, there is a w with $\|w\| \le 1$ such that

$$\forall i \in [n], \ b_i(w_i \cdot \mathbf{e}_i) = \gamma \ .$$

To see this, simply choose $w_i = \gamma b_i$. Then the above equality is true. Moreover, $\|w\|^2 = \gamma^2 \sum_{i=1}^n b_i^2 = \gamma^2 n \le 1$.

Now given an algorithm $\mathcal{A}$, define the data set $\{(x_i, y_i)\}_{i=1}^n$ as follows. Let $x_i = \mathbf{e}_i$ for all i and $y_1 = -\mathcal{A}(x_1)$. Define y_i for $i > 1$ recursively as

$$y_i = -\mathcal{A}(x_1, y_1, \ldots, x_{i-1}, y_{i-1}, x_i) \ .$$

It is clear that the algorithm makes n mistakes when run on this data set. By the above claim, no matter what y_i's turn out to be, the data set is separable by a margin of γ. □

3 The Winnow Algorithm

Algorithm 2 WINNOW

Input parameter: $\eta > 0$ (learning rate)

$w_1 \leftarrow \frac{1}{d}\mathbf{1}$
for $t = 1$ to T **do**
  Receive $x_t \in \mathbb{R}^d$
  Predict $\mathrm{sgn}(w_t \cdot x_t)$
  Receive $y_t \in \{-1, +1\}$
  if $\mathrm{sgn}(w_t \cdot x_t) \neq y_t$ **then**
    $\forall i \in [d]$, $w_{t+1,i} \leftarrow \frac{w_{t,i} \exp(\eta y_t x_{t,i})}{Z_t}$ where $Z_t = \sum_{i=1}^d w_{t,i} \exp(\eta y_t x_{t,i})$
  else
    $w_{t+1} \leftarrow w_t$
  end if
end for

Theorem 3.1. *Suppose Assumption M holds. Further assume that* $w^* \geq \mathbf{0}$. *Let*

$$M_T := \sum_{t=1}^{T} \mathbf{1}\left[\mathrm{sgn}(w_t \cdot x_t) \neq y_t\right]$$

denote the number of mistakes the WINNOW *algorithm makes. Then, for a suitable choice of* η, *we have,*

$$M_T \leq \frac{2\|x_{1:T}\|_\infty^2 \cdot \|w^*\|_1^2}{\gamma^2} \ln d \, .$$

Proof. Let $u^* = w^*/\|w^*\|$. Since we assume $w^* \geq \mathbf{0}$, u^* is a probability distribution. At all times, the weight vector w_t maintained by WINNOW is also a probability distribution. Let us measure the progress of the algorithm by analyzing the *relative entropy* between these two distributions at time t. Accordingly, define

$$\Phi_t := \sum_{i=1}^{d} u_i^* \ln \frac{u_i^*}{w_{t,i}} \, .$$

When there is no mistake $\Phi_{t+1} = \Phi_t$. On a round when a mistake occurs, we have

$$\begin{aligned}
\Phi_{t+1} - \Phi_t &= \sum_{i=1}^{d} u_i^* \ln \frac{w_{t,i}}{w_{t+1,i}} \\
&= \sum_{i=1}^{d} u_i^* \ln \frac{Z_t}{\exp(\eta y_t x_{t,i})} \\
&= \ln(Z_t) \sum_{i=t}^{d} u_i^* - \eta y_t \sum_{i=1}^{d} u_i^* x_{t,i} \\
&= \ln(Z_t) - \eta y_t (u^* \cdot x_t) \\
&\leq \ln(Z_t) - \eta\gamma/\|w^*\|_1 \, , \qquad (4)
\end{aligned}$$

where the last inequality follows from the definition of u^* and Assumption M. Let $L = \|x_{1:T}\|_\infty$. Then $y_t x_{t,i} \in [-L, L]$ for all t, i. Then we can bound

$$Z_t = \sum_{i=1}^{d} w_{t,i} e^{\eta y_t x_{t,i}}$$

using the convexity of the function $t \mapsto e^{\eta t}$ on the interval $[-L, L]$ as follows.

$$\begin{aligned}
Z_t &\leq \sum_{i=1}^{d} \frac{1 + y_t x_{t,i}/L}{2} e^{\eta L} + \frac{1 - y_t x_{t,i}/L}{2} e^{-\eta L} \\
&= \frac{e^{\eta L} + e^{-\eta L}}{2} \sum_{i=1}^{d} w_{t,i} + \frac{e^{\eta L} - e^{-\eta L}}{2} \left(y_t \sum_{i=1}^{d} w_{t,i} x_{t,i} \right) \\
&= \frac{e^{\eta L} + e^{-\eta L}}{2} + \frac{e^{\eta L} - e^{-\eta L}}{2} y_t (w_t \cdot x_t) \\
&\leq \frac{e^{\eta L} + e^{-\eta L}}{2}
\end{aligned}$$

because having a mistake implies $y_t(w_t \cdot x_t) \leq 0$ and $e^{\eta L} - e^{-\eta L} > 0$. So we have proved

$$\ln(Z_t) \leq \ln\left(\frac{e^{\eta L} + e^{-\eta L}}{2} \right) \, . \qquad (5)$$

Define

$$C(\eta) := \eta\gamma/\|w^*\|_1 - \ln\left(\frac{e^{\eta L} + e^{-\eta L}}{2}\right) \ .$$

Combining (4) and (5) then gives us

$$\Phi_{t+1} - \Phi_t \leq -C(\eta)\mathbf{1}\left[y_t \neq \operatorname{sgn}(w_t \cdot x_t)\right] \ .$$

Unwinding the recursion gives,

$$\Phi_{T+1} \leq \Phi_1 - C(\eta)M_T \ .$$

Since relative entropy is always non-negative $\Phi_{T+1} \geq 0$. Further,

$$\Phi_1 = \sum_{i=1}^{d} u_i^* \ln(du_i^*) \leq \sum_{i=1}^{d} u_i^* \ln d = \ln d$$

which gives us

$$0 \leq \ln d - C(\eta)M_T$$

and therefore $M_T \leq \frac{\ln d}{C(\eta)}$. Setting

$$\eta = \frac{1}{2L}\ln\left(\frac{L + \gamma/\|w^*\|_1}{L - \gamma/\|w^*\|_1}\right)$$

to maximize the denominator $C(\eta)$ gives

$$M_T \leq \frac{\ln d}{g\left(\frac{\gamma}{L\|w^*\|_1}\right)}$$

where $g(\epsilon) := \frac{1+\epsilon}{2}\ln(1+\epsilon) + \frac{1-\epsilon}{2}\ln(1-\epsilon)$. Finally, noting that $g(\epsilon) \geq \epsilon^2/2$ proves the theorem. $\square$