**COMP 585** **Spring 2024**

*Probabilistic Toolkit for Learning and Computing* *Instructor: Maryam Aliakbarpour*

*April 2, 2024* *Scriber: Yu-Ching Shen*

# Lecture 21

## 1 Weak learning

We discuss weak learning in this lecture. In weak learning, it is sufficient to output a solution whose error is smaller than 1/2.

**Definition 1.1** (Weakly learnable)**.** *We say a concept class $\mathcal{C}$ is weakly learnable if there is an algorithm $\mathcal{A}$ and a parameter $\gamma \in (0, 1/2)$ such that for all distribution $\mathcal{D}$ and $\delta \in (0, 1)$, the algorithm $\mathcal{A}$ receives $m(\delta, \mathcal{C})$ samples and outputs a concept $\hat{c}$ for which we have $\mathbf{Pr}\left[err_{\mathcal{D}}(\hat{c}) \geq \frac{1}{2} - \gamma\right] \leq \delta$.*

*(The function $err_{\mathcal{D}}(\cdot)$ is defined by $err_{\mathcal{D}}(c) := \mathbf{Pr}_{(x,y)\sim\mathcal{D}}[c(x) \neq y]$.)*

Recall that in the previous lectures, the algorithm is required to achieve a "small" error.

**Definition 1.2** (Strongly learnable)**.** *We say a concept class $\mathcal{C}$ is strong learnable if there is an algorithm $\mathcal{B}$ such that for all distribution $\mathcal{D}$ and $\delta \in (0, 1)$, the algorithm $\mathcal{B}$ receives $m(\delta, \mathcal{C})$ samples and outputs a concept $\hat{c}$ for which we have $\mathbf{Pr}[err_{\mathcal{D}}(\hat{c}) \geq \varepsilon] \leq \delta$.*

**Remark 1.** *If an algorithm outputs a concept $\hat{c} \in \mathcal{C}$, we call it proper learning. On the other hand, if an algorithm outputs a concept which may or may not be in the concept class $\mathcal{C}$, we call it improper learning. We will see an improper learning algorithm in this lecture.*

We can see strong learnability trivially implies weak learnability by choosing $\varepsilon < 1/2 - \gamma$. We are going to show that the opposite direction is also correct. That is, if there is an algorithm $\mathcal{A}$ that learns $\mathcal{C}$ with an error less than $1/2-\gamma$, then there is an algorithm $\mathcal{B}$ that uses $\mathcal{A}$ to learn $\mathcal{C}$ with an error less than $\varepsilon$. The algorithm $\mathcal{B}$ may use more samples than that are used in $\mathcal{A}$.

**Theorem 2.** *If a concept class $\mathcal{C}$ is weakly learnable, then $\mathcal{C}$ is also strongly learnable.*

## 2 Weak learning $\Rightarrow$ strong learning: algorithm

We present the strong learning algorithm $\mathcal{B}$ that uses a weak learning algorithm $\mathcal{A}$ as follows. We consider the concepts whose range is $\pm 1$.

**AdaBoost algorithm:**

1. Let $S = \{(x_i, y_i)\}_{i=1}^m$ be the training set, and let $\mathcal{D}_1(i) = 1/m$.
2. For $t = 1, 2, \ldots, T$:

   2.1 Run $\mathcal{A}$ on $\mathcal{D}_t(x, y)$. Get output $\hat{c}_t$.
   The distribution $\mathcal{D}_t(x, y)$ is defined by

   $$\mathcal{D}_t(x,y) = \begin{cases} 0 & \text{if } (x,y) \notin S, \\ \mathcal{D}_t(i) & \text{if } (x,y) = (x_i, y_i). \end{cases}$$

   2.2 Compute $\varepsilon_t := err_{\mathcal{D}_t}(\hat{c}_t) = 1/2 - \gamma_t$.

   2.3 Compute $\alpha_t := \frac{1}{2}\ln(\frac{1-\varepsilon_t}{\varepsilon_t}) > 0$.

   2.4 Define $\mathcal{D}_{t+1}(i) := \frac{\mathcal{D}_t(i)e^{-\alpha_t \hat{c}_t(x_i) y_i}}{Z_t}$, where $Z_t = \sum_{i=1}^m \mathcal{D}_t(i) e^{-\alpha_t \hat{c}_t(x_i) y_i}$.

3. Output $\hat{c}(\cdot) := \text{sign}(\sum_{t=1}^T \alpha_t \hat{c}_t(\cdot))$.
   The function $\text{sign}(\cdot)$ denotes the sign function.

The number of samples $m$ and the number of iterations $T$ have not been determined yet. In Step 2b, we can calculate $\varepsilon$ because $\mathcal{D}_t$ is known. Also, we have that $\gamma_t \geq \gamma$ for all $t$.

It is the first time in this course that we have seen an algorithm output a concept that may not be in the concept class.

# 3 The empirical error of the output concept

In the remaining lecture, we are going to compute the empirical error of $\hat{c}$ associated with the sample set $S$. We have the following result.

**Lemma 3.1.** *Let $\hat{err}_S(\hat{c})$ be the empirical error of the output concept $\hat{c}$ defined by $\frac{1}{|S|}\sum_{(x_i,y_i)\in S} \mathbb{1}_{\{\hat{c}(x_i)\neq y_i\}}$, where $\mathbb{1}$ denotes the indicator. It holds that $err_S(\hat{c}) \leq e^{-2T\gamma^2}$.*

Before proving Lemma 3.1, we first give the following identities and inequalities related to $\mathcal{D}_t$ and $Z_t$.

**Fact 3.2.**

$$\mathcal{D}_t(i) = D_1(i)\prod_{t=1}^{t-1}\frac{e^{-\alpha_j \hat{c}_j(x_i) y_i}}{Z_j} = \frac{1}{m}\frac{e^{-\sum_{j=1}^{t-1}\alpha_j \hat{c}_j(x_i) y_j}}{\prod_{j=1}^{t-1} Z_j}. \tag{1}$$

**Fact 3.3.** *Let $F(x) := \sum_{j1}^T \alpha_j \hat{c}_j(x)$. Plug it into Eq. (1) and set $i = T+1$. We have*

$$\mathcal{D}_{T+1}(i) = \frac{e^{-y_i F(x_i)}}{m\prod_{j=1}^T Z_j}. \tag{2}$$

Note that $y_iF(x_i) \leq 0$ corresponds to the event that there is a mislabeling for $(x_i, y_i)$.

**Fact 3.4.**

$$\mathbb{1}_{\{y_iF(x_i)\leq 0\}} \leq e^{-y_iF(x_i)}. \tag{3}$$

*Proof.*

- If $y_iF(x_i) > 0$, then $\mathbb{1}_{\{y_iF(x_i)\leq 0\}} = 0$. And we have $0 \leq e^a$ for all $a \in \mathbb{R}$.
- If $y_iF(x_i) \leq 0$, then $\mathbb{1}_{\{y_iF(x_i)\leq 0\}} = 1$. And we have $1 \leq e^a$ for all $a \geq 0$.

□

**Fact 3.5.** *The value of $Z_t$ in terms of error is given by*

$$Z_t = e^{\alpha_t}\varepsilon_t + e^{-\alpha_t}(1-\varepsilon_t) \tag{4}$$

*Proof.*

$$\begin{aligned}
Z_t &= \sum_{i\in S} \mathcal{D}_t(i)e^{-\alpha_t\hat{c}_t(x_i)y_i} \\
&= \sum_{i:y_i\neq\hat{c}_t(x_i)} \mathcal{D}_t(i)e^{\alpha_t} + \sum_{i:y_i=\hat{c}_t(x_i)} \mathcal{D}_t(i)e^{-\alpha_t} \\
&\qquad \text{(Seperate correct and not correct labeling.)} \\
&= err_{\mathcal{D}_t}(\hat{c}_t)e^{\alpha_t} + (1-err_{\mathcal{D}_t}(\hat{c}_t))e^{-\alpha_t} \\
&\qquad \text{(Definition of } err(\cdot)\text{.)} \\
&= e^{\alpha_t}\varepsilon_t + e^{-\alpha_t}(1-\varepsilon_t) \\
&\qquad \text{(Reword error.)}
\end{aligned}$$

□

Combining the above, we are going to prove Lemma 3.1.

*Proof of Lemma 3.1.*

$$
\begin{aligned}
\hat{err}_S(\hat{c}) &= \frac{1}{|S|}\sum_{i\in S}\mathbb{1}_{\{\hat{c}(x_i)\neq y_i\}} && \text{(By definition.)}\\
&= \frac{1}{m}\sum_{i\in S}\mathbb{1}_{\{\text{sign}(\sum_t \alpha_t\hat{c}_t)\neq y_i\}} && (\hat{c}_t := \text{sign}(\sum_t \alpha_t\hat{c}_t).)\\
&= \frac{1}{m}\sum_{i}\mathbb{1}_{\{F(x_i)y_i\leq 0\}} && \text{(Reword mislabeling.)}\\
&\leq \frac{1}{m}\sum_{i} e^{-y_iF(x_i)} && \text{(By Eq. (3).)}\\
&\leq \frac{1}{m}\sum_{i} \mathcal{D}_{T+1}(i)m\prod_{j=1}^{T} Z_j && \text{(By Eq. (2.))}\\
&\leq \prod_{j=1}^{T} Z_j && \text{(Sum over a distribution = 1.)}\\
&= \prod_{t=1}^{T} e^{\alpha_t}\varepsilon_t + e^{-\alpha_t}(1-\varepsilon_t) && \text{(By Eq. (4).)}\\
&= \prod_{t=1}^{T} 2\sqrt{\varepsilon_t(1-\varepsilon_t)} && \text{(By Def. of } \alpha_t \text{ in Step 2c.)}\\
&= \prod_{t=1}^{T} 2\sqrt{(\frac{1}{2}-\gamma_t))(1-\frac{1}{2}+\gamma_t)} && \text{(By Def. of } \varepsilon_t \text{ in Step 2b.)}\\
&= \prod_{t=1}^{T} 2\sqrt{(\frac{1}{4}-\gamma_t^2))} = \prod_{t=1}^{T}\sqrt{(1-4\gamma^2))}\\
&\leq \prod_{t=1}^{T} e^{-4\gamma_t^2/2} && \text{(By } 1-x<e^{-x}.\text{)}\\
&= e^{-2\sum_t \gamma_t^2}\\
&\leq e^{-2T\gamma^2} && \text{(By } \gamma_t\geq\gamma.\text{)}
\end{aligned}
$$

□