CMPUT 654 Fa 23: Theoretical Foundations of Machine Learning **Fall 2023**

Lecture 5: ERM and learning the AND class (Sept 19)

Lecturer: Csaba Szepesvári *Scribes: Zixin Zhong*

Note: *LaTeX template courtesy of UC Berkeley EECS dept. (link to directory)*
Disclaimer: *These notes have* ***not*** *been subjected to the usual scrutiny reserved for formal publications. They may be distributed outside this class only with the permission of the Instructor.*

5.1 Review

5.1.1 Recall: Concentration inequalities

Theorem 5.1 (Additive Chernoff's Inequality)**.** *Let* $X_1, \dots, X_n \in [0,1]$ *be i.i.d. random variables,* $\bar{X}_n = \frac{1}{n}(X_1 + \dots + X_n)$, $\mu = \mathbb{E}X_1$. *We have*

(a) $\forall \delta \in (0,1)$, *with probability* $1-\delta$,

$$\bar{X}_n \le \mu + \sqrt{\frac{\log(1/\delta)}{2n}};$$

(b) $\forall \delta \in (0,1)$, *with probability* $1-\delta$,

$$\bar{X}_n \ge \mu - \sqrt{\frac{\log(1/\delta)}{2n}}.$$

Theorem 5.2 (Multiplicative Chernoff's Inequality)**.** *Let* $X_1, \dots, X_n \in [0,1]$ *be i.i.d. random variables,* $\bar{X}_n = \frac{1}{n}(X_1 + \dots + X_n)$, $\mu = \mathbb{E}X_1$. *We have*

(a) $\forall \delta \in (0,1)$, *with probability* $1-\delta$,

$$\bar{X}_n \le \mu + \sqrt{\frac{2\mu\log(1/\delta)}{n}} + \frac{1}{3n};$$

(b) $\forall \delta \in (0,1)$, *with probability* $1-\delta$,

$$\bar{X}_n \ge \mu - \sqrt{\frac{2\mu\log(1/\delta)}{n}}. \qquad (*)$$

5.1.2 Recall: PAC-learning

Let function $f_* : \{0,1\}^d \to \{0,1\}$, $X_1, X_2, \dots, X_n \in \{0,1\}^d := \underline{2}^d$ be i.i.d. random variables drawn from distribution P_X, data set $D_n = \{(X_1, f_*(X_1)), \dots, (X_n, f_*(X_n))\}$.

Let $f_* \in \mathcal{F} \subset \underline{2}^{\underline{2}^d}$ and $f \in \underline{2}^{\underline{2}^d}$. In other words, $\mathcal{X} = \underline{2}^d$, $\mathcal{Y} = \underline{2}$, $f_* \in \mathcal{Y}^{\mathcal{X}}$, $f : \mathcal{X} \to \mathcal{Y}$.

Let $P_X^{f_*} := P(X_1, f_* X_1)$, and

$$\begin{aligned}
&L(f) = \mathbb{P}\left(f(X) \ne f_*(X)\right) = L(P_X^{f_*}, f),\\
&l : \underline{2} \times \underline{2} \to \underline{2}, \quad l(y, y') = \mathbf{1}(y \ne y'),\\
&L(P_X^{f_*}, f) = \int P(\mathrm{d}x, \mathrm{d}y)\, l(f(x), y).
\end{aligned}$$

Definition 5.3 (PAC-Learning)**.** Fix $\mathcal{C} = (\mathcal{C}_d)_{d\geq 1}$, where $\mathcal{C}_d \subset \underline{2}^{\underline{2}^d}$. $\mathcal{C}$ is **PAC-learnable (Probably Approximately Correctly)** if $\exists$ polynomial $p \in \mathbb{R}[x,y,z]$ (**computed with polynomial cost**) and $\mathcal{A} = (\mathcal{A}_{n,d})_{n\geq 1, d\geq 1}$ where $\mathcal{A}_{n,d} : \left(\underline{2}^d \times \underline{2}\right)^n \to \underline{2}^{\underline{2}^d}$

$$\begin{aligned}\text{s.t.}\quad & \forall \varepsilon \in (0,1),\ \delta \in (0,1),\ d \geq 1,\ P \in \mathcal{M}_1(\underline{2}^d),\ f_* \in \mathcal{C}_d,\\ & n \geq \lceil p(\underbrace{1/\varepsilon}_{\text{accuracy}}, \underbrace{1/\delta}_{\text{confidence}}, d)\rceil,\\ & X_1, X_2, \ldots, X_n \sim P_X,\\ & f_n = \mathcal{A}_{n,d}\left(\underbrace{(X_1, f_*(X_1)), \ldots, (X_n, f_*(X_n))}_{D_n}\right) \quad \text{i.e. } D_n \overset{\mathcal{A}}{\to} f_n,\end{aligned}$$

we have

$$\mathbb{P}\left(L\left(P_X^{f_*}, f_n\right) \geq \varepsilon\right) \leq \delta.$$

In other words, with probability $1-\delta$, $\mathbb{P}\left(f_n(X) \neq f_*(x) | D_n\right) \leq \varepsilon$.

Remark 5.4 (Example)**.**

$$\begin{aligned}&\mathcal{C}_{\text{AND},d} = \left\{ f : \underline{2}^d \to \underline{2} \mid \exists u \subset [d],\ \forall x \in \underline{2}^d : f(x) = \min_{j\in u} X_j \right\},\\ &\mathcal{C} = \left(\mathcal{C}_{\text{AND},d}\right)_{d\geq 1}.\end{aligned}$$

5.2 ERM: Empirical Risk Minimization

Let

$$\begin{aligned}L_n(f) &= \frac{1}{n}\sum_{i=1}^n \mathbf{1}\left(f(X_i) \neq Y_i\right),\\ f_n &:= \arg\min_{f\in\mathcal{C}_d} L_n(f).\end{aligned}$$

Homework: Show $f_n \arg\min_{f\in\mathcal{C}_d} L_n(f)$ is computationally efficient.

Moreover,

$$f_n := \arg\min_{f\in\mathcal{C}_{\text{AND}}} L_n(f) \qquad \longrightarrow \qquad \text{proper learning.}$$

Method I: Fix $d \geq 1$, P and f_*. Let $D_n \to f_n$. We first decompose $L(f_n)$ as follows:

$$\begin{aligned}L(f_n) &= L(f_n) - L_n(f_n) + L_n(f_n)\\ &= \underbrace{L(f_n) - L_n(f_n)}_{\text{bounded with concentration inequality}} + \underbrace{L_n(f_n) - L_n(f_*)}_{\text{ERM}} + \underbrace{L_n(f_*) - L(f_*)}_{\text{bounded with concentration inequality}} + \underline{L(f_*)}_{=0}.\end{aligned}$$

Next, Hoeffding inequality implies that with probability $1-\delta$,

$$L_n(f_*) - L(f_*) \leq \sqrt{\frac{\log(1/\delta)}{2n}}.$$

Besides, $f_n \in \mathcal{C}_d$ indicates that

$$L(f_n) - L_n(f_n) \leq \max_{f\in\mathcal{C}_d} L(f) - L_n(f_n).$$

Fix $f \in \mathcal{C}_d$. Set

$$\mathcal{U}(f,\delta) = \left\{ L(f) - L_n(f) \le \sqrt{\frac{\log(1/\delta)}{2n}} \right\}.$$

Then

$$\mathbb{P}\left(\mathcal{U}(f,\delta)\right) \ge 1-\delta \quad \Longleftrightarrow \quad \mathbb{P}\left(\mathcal{U}^{\mathrm{c}}(f,\delta)\right) \le \delta.$$

Let $N = |\mathcal{C}_{\text{AND},d}|$ and define 'nice event'

$$\mathcal{U} = \bigcap_{f \in \mathcal{C}_d} \mathcal{U}\left(f, \frac{\delta}{N}\right).$$

Then

$$\mathbb{P}(\mathcal{U}^{\mathrm{c}}) = \mathbb{P}\left(\bigcup_{f\in\mathcal{C}_d} \mathcal{U}^{\mathrm{c}}\left(f,\frac{\delta}{N}\right)\right) \le \sum_{f\in\mathcal{C}_d} \mathbb{P}\left(\mathcal{U}^{\mathrm{c}}\left(f,\frac{\delta}{N}\right)\right) \le \sum_{f\in\mathcal{C}_d}\frac{\delta}{N} = \delta.$$

When $\mathcal{U}$ holds, $L(f) - L_n(f) \le \sqrt{\frac{\log(N/\delta)}{2n}}$ for all $f \in \mathcal{C}_d$, in other words,

$$\max_{f\in\mathcal{C}_d} L(f) - L_n(f) \le \sqrt{\frac{\log(N/\delta)}{2n}}.$$

Theorem 5.5 (Proper learning)**.** *$\mathcal{C}_{AND,d}$ PAC-learnable, f_n minimizing the empirical risk, and proper learning: with propability $1-\delta$,*

$$L(f_n) \le \sqrt{\frac{\log(N+1/\delta)}{2n}} + \sqrt{\frac{\log(N+1/\delta)}{n}}.$$

Remark 5.6. (a) This bound on $L(f_n)$ may **not** be **tight**.

(b) This result shows PAC-learnability:

$$\begin{aligned} 2\sqrt{\frac{\log(N+1/\delta)}{2n}} &\le \varepsilon \\ \Leftrightarrow \frac{n}{2\log(N+1/\delta)} &\ge \frac{1}{\varepsilon^2} \\ \Leftrightarrow n &\ge \frac{2}{\varepsilon^2}\log\left(\frac{N+1}{\delta}\right). \end{aligned}$$

Hence, $p(1/\varepsilon, 1/\delta, d) = 2\log\left((|\mathcal{C}_{\text{AND},d}|+1)/\delta\right)/\varepsilon^2$. Since $|\mathcal{C}_{\text{AND},d}| = 2^d$, we have

$$p\left(\frac{1}{\varepsilon},\frac{1}{\delta},d\right) = \frac{2\log\left(2^d+1\right)+2\log(1/\delta)}{\varepsilon^2} \le \dots$$

Method II: $L(f) - L_n(f) \le ?$

Fix $0 \le \delta \le 1$. By multiplicative Chernoff inequality, with probability $1 - \delta N/(N+1)$,

$$L(f) - L_n(f) \le \sqrt{\frac{2L(f)\log(N+1/\delta)}{n}} \quad \forall f;$$

with probability $1 - \delta/(N+1)$,

$$L_n(f_*) - L(f_*) \leq \sqrt{\frac{2L(f_*)\log(N+1/\delta)}{n}} + \frac{\log((N+1)/\delta)}{3n}.$$

Denote ‘nice event’

$$\mathcal{U} := \left\{ L(f) - L_n(f) \leq \sqrt{\frac{2L(f)\log(N+1/\delta)}{n}} \quad \forall f,\ L_n(f_*) - L(f_*) \leq \sqrt{\frac{2L(f_*)\log(N+1/\delta)}{n}} + \frac{\log((N+1)/\delta)}{3n} \right\}.$$

Then $\mathbb{P}(\mathcal{U}) \geq 1 - \delta$. On $\mathcal{U} \cap \{L(f_n) \neq 0\}$: since $L_n(f_n) \geq 0$ and

$$\frac{L(f_n) - L_n(f_n)}{\sqrt{\frac{2L(f_n)\log\left(\frac{N+1}{\delta}\right)}{n}}} \leq \max_{f \in \mathcal{C}_d, L(f) \neq 0} \frac{L(f) - L_n(f)}{\sqrt{\frac{2L(f)\log\left(\frac{N+1}{\delta}\right)}{n}}} \leq 1,$$

we have

$$L(f_n) \leq \sqrt{\frac{2L(f_n)\log\left(\frac{N+1}{\delta}\right)}{n}}.$$

Furthermore, we have

$$\begin{aligned} L^2(f_n) &\leq \frac{2L(f_n)\log\left(\frac{N+1}{\delta}\right)}{n}, \\ L(f_n) &\leq \frac{2\log\left(\frac{N+1}{\delta}\right)}{n} \leq \varepsilon, \\ \Rightarrow n &\geq \frac{2\log\left(\frac{|\mathcal{C}_d|+1}{\delta}\right)}{\varepsilon}, \\ \Rightarrow p\left(\frac{1}{\varepsilon}, \frac{1}{\delta}, d\right) &= \dots \end{aligned}$$