COMP 585 **Spring 2024**

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

March 5th, 2024 *Scriber: Jason S. Lenderman*

Lecture 15

1 Uniform convergence

Recall that a class $\mathcal{C}$ has the *uniform convergence property* if $\forall \epsilon, \delta \in (0,1)$ and any distribution D over $\mathcal{X} \times \{0,1\}$ there exists an $m \in \mathbb{N}$ (depending on ϵ and δ, but not D) such that if $(X_1, Y_1), \ldots (X_m, Y_m)$ are m i.i.d. samples with distribution$D$, then

$$\mathbf{Pr}\big[|\widehat{\mathrm{err}}(c) - \mathrm{err}(c)| \leq \epsilon \ \forall c \in \mathcal{C}\big] > 1 - \delta,$$

where

$$\widehat{\mathrm{err}}(c) = \frac{|\{i \in [m] : c(X_i) \neq Y_i\}|}{m}$$

and

$$\mathrm{err}(c) = \mathbf{Pr}_{(X,Y)\sim D}\big[c(X) \neq Y\big]$$

are the empirical and expected error of $c \in \mathcal{C}$, respectively.

Theorem 1. *Suppose that class $\mathcal{C}$ over instance space $\mathcal{X}$ has the uniform convergence (UC) property. Then $\mathcal{C}$ is agnostic-PAC learnable via the ERM algorithm. More precisely, given $\epsilon, \delta \in (0,1)$ there exists an $m \in \mathbb{N}$ such that for any distribution D on $\mathcal{X} \times \{0,1\}$*

$$\mathbf{Pr}_{(X_1,Y_1),\ldots(X_m,Y_m)\sim D^m}\left[err\left(c_{ERM}\right) \leq + \min_{c\in\mathcal{C}} err(c) + \epsilon\right] \geq 1 - \delta.$$

for any $c_{ERM} \in \mathcal{C}$ satisfying $\widehat{err}\left(c_{ERM}\right) = \min_{c\in\mathcal{C}} \widehat{err}(c)$

Proof. Since $\mathcal{C}$ has the UC property there exists an $m \in \mathbb{N}$ such that

$$\mathbf{Pr}_{(X_1,Y_1),\ldots,(X_m,Y_m)\sim D^m}\left[|\widehat{\mathrm{err}}(c) - \mathrm{err}(c)| \leq \frac{\epsilon}{2} \ \forall c \in \mathcal{C}\right] > 1 - \delta. \tag{1}$$

Now let c_{ERM} be any member of $\mathcal{C}$ minimizing the empirical risk, i.e.

$$\widehat{\mathrm{err}}\left(c_{\mathrm{ERM}}\right) = \min_{c\in\mathcal{C}} \widehat{\mathrm{err}}(c)$$

and let c^* be any member of $\mathcal{C}$ minimizing the true risk, i.e.

$$\text{err}\left(c^*\right) = \min_{c \in \mathcal{C}} \text{err}(c).$$

Then we have

$$\text{err}\left(c_{\text{ERM}}\right) - \text{err}\left(c^*\right) = \text{err}\left(c_{\text{ERM}}\right) - \widehat{\text{err}}\left(c_{\text{ERM}}\right) + \widehat{\text{err}}\left(c_{\text{ERM}}\right) - \widehat{\text{err}}\left(c^*\right) + \widehat{\text{err}}\left(c^*\right) - \text{err}\left(c^*\right).$$

But $\widehat{\text{err}}\left(c_{\text{ERM}}\right) - \widehat{\text{err}}\left(c^*\right) < 0$ since c_{ERM} is a minimizer of the empirical risk. Therefore it follows

$$\begin{aligned}
&\mathbf{Pr}_{(X_1,Y_1),\ldots,(X_n,Y_n)\sim D^n}\big[\widehat{\text{err}}\left(c_{\text{ERM}}\right) - \text{err}\left(c^*\right) \leq \epsilon\big] \\
&\quad \geq \mathbf{Pr}_{(X_1,Y_1),\ldots,(X_n,Y_n)\sim D^n}\big[\text{err}\left(c_{\text{ERM}}\right) - \widehat{\text{err}}\left(c_{\text{ERM}}\right) + \widehat{\text{err}}\left(c^*\right) - \text{err}\left(c^*\right) \leq \epsilon\big] \\
&\quad \geq \mathbf{Pr}_{(X_1,Y_1),\ldots,(X_n,Y_n)\sim D^n}\left[\text{err}\left(c_{\text{ERM}}\right) - \widehat{\text{err}}\left(c_{\text{ERM}}\right) \leq \frac{\epsilon}{2} \;\text{ and }\; \widehat{\text{err}}\left(c^*\right) - \text{err}\left(c^*\right) \leq \frac{\epsilon}{2}\right] \\
&\quad \geq 1 - \delta
\end{aligned}$$

as desired, where the final inequality follows from the the inequality at (1). □

2 Overfitting

Although we are guaranteed to have agnostic-PAC learnability when $\mathcal{C}$ has the UC property, it is possible that if $\mathcal{C}$ is very "rich" then we might *overfit* the data leading to a situation where one or more hypotheses in $\mathcal{C}$ that are minimizers of the empirical error, nevertheless have a true which is error significantly larger than $\min_{c\in\mathcal{C}} \text{err}(c)$. For example, if $\mathcal{C}$ is set of indicator functions of all measurable subsets of $[0, 1]$ and D is taken to be the joint distribution of (X, Y) where $X \sim U[0, 1]$ and $Y = 1$ w.p. 1. Then clearly $\min_{c\in\mathcal{C}} \text{err}(c) = 0$. But given "training sample" $(X_1, Y_1), \ldots, (X_m, Y_m)$, the hypothesis $\hat{c}$ satisfying $\hat{c}(X_1) = \hat{c}(X_2) = \cdots \hat{c}(X_m) = 1$ and $\hat{c}(x) = 0$ for all $x \in [0, 1] \setminus \{X_1, \ldots, X_m\}$ minimizes the empirical risk but has

$$\begin{aligned}
\text{err}(\hat{c}) &= \mathbf{Pr}_{(X,Y),(X_1,Y_1),\ldots,(X_m,Y_m)\sim D^{m+1}}\big[\hat{c}(X) \neq Y\big] \\
&= \mathbf{Pr}_{(X,Y),(X_1,Y_1),\ldots,(X_m,Y_m)\sim D^{m+1}}\big[X \notin \{X_1, \ldots, X_m\}\big] = 1,
\end{aligned}$$

where the final equality follows by the fact that the marginal distribution of D on the first coordinate is continuous (more precisely uniform on $[0, 1]$.)

3 PAC-learnability of finite classes

The following theorem establishes that ERM "works", i.e., choosing any minimizer of the empirical risk is a PAC learning algorithm for concept class $\mathcal{C}$, in the realizable case when $\mathcal{C}$ is finite.

Theorem 2. *Let $\mathcal{C}$ be a finite concept class over instance space $\mathcal{X}$, then $\mathcal{C}$ is agnostic-PAC learnable via ERM.*

Proof. Given $\epsilon, \delta \in (0,1)$, take m to be an integer no less than $\log(|\mathcal{C}|/\delta)/\epsilon$ and let $(X_1, Y_1), \ldots, (X_m, Y_m)$ be an i.i.d. sample with some distribution D. Since we're in the realizable case we assume there is some $c^* \in \mathcal{C}$ with $\mathrm{err}\,(c^*) = 0$. Our goal is to show the probability that ERM fails is small. More precisely, we want to show that any member of $\mathcal{C}$ which is minimizer of the empirical risk, say $\hat{c}$, satisfies $\mathbf{Pr}[\mathrm{err}\,(\hat{c}) < \epsilon] \geq 1 - \delta$. Now let $\mathcal{C}_b = \{c \in \mathcal{C} : \mathrm{err}(c) > \epsilon\}$ denote the collection of "bad" hypotheses. We want to show that the probability of any member of $\mathcal{C}_b$ being a minimizer of the emprical risk is small. To see this take any $c \in \mathcal{C}_b$ and note that

$$\mathbf{Pr}[\widehat{\mathrm{err}}\,(c) = 0] \leq (1-\epsilon)^m \leq e^{-m\epsilon}.$$

Using the union bound it now follows that

$$\begin{aligned} \mathbf{Pr}[\exists c \in \mathcal{C} \text{ with } \mathrm{err}(c) > \epsilon \text{ and } \widehat{\mathrm{err}}(c) = 0] &= \mathbf{Pr}[\exists c \in \mathcal{C}_b \text{ with } \widehat{\mathrm{err}}(c) = 0] \\ &\leq |\mathcal{C}_b| e^{-m\epsilon} \\ &\leq |\mathcal{C}| e^{-m\epsilon} \leq \delta. \end{aligned}$$

where first inequality is a consequnece of the union bound and the final inequality follows from $m \geq \log(|\mathcal{C}|/\delta)/\epsilon$. Since we're in the realizable case we know that the any empirical risk minimizer $\hat{c} \in \mathcal{C}$ has $\widehat{\mathrm{err}}(\hat{c})) = 0$, therefore it follows that ERM will produce a hypothesis having true error at most ϵ with probability at least $1 - \delta$. ☐

In the agnostic case it is also possible to show that ERM is a PAC learning algorithm for $\mathcal{C}$ when $\mathcal{C}$ is finite. To show this one first establishes that any finite concept class $\mathcal{C}$ has the UC property via the result of Problem 1 (below) and then applies Theorem 1.

Problem 1. Suppose $\mathcal{C}$ is a finite class and

$$m = O\left(\frac{\log |\mathcal{C}|/\delta}{\epsilon^2}\right).$$

Then for all $c \in \mathcal{C}$ we have $|\widehat{\mathrm{err}}(c) - \mathrm{err}(c)| < \epsilon/2$ with probability at least $1 - \delta$.

4 No free lunch theorem

Let $\mathcal{X}$ be some instance space and for some $m \in \mathbb{N}$ let $x_1, \ldots, x_{2m}$ be distinct points on $\mathcal{X}$. Let $\mathcal{C}$ be concept class consisting of all possible labellings of $x_1, \ldots, x_{2m}$. Note that $|\mathcal{C}| = 2^{2m}$. Now fix some concept $c^* \in \mathcal{C}$ and let D be the joint distribution of $(X, c^*(X))$ where X is taken to be a random variable with uniform distribution on $\{x_1, \ldots, x_{2m}\}$.

Take $T = \{(X_i, Y_i) : i \in [m]\}$ to be m i.i.d. random variables with distribution D (WLOG assume distinct X_i's are distinct) and let

$$\mathcal{P} = \{c \in \mathcal{C} : \widehat{\text{err}}(c) = 0\}$$

denote the set of "promising" concepts. Note that $|\mathcal{P}| = 2^m$ since each $c \in \mathcal{P}$ is determined on the set $\{X_1, \ldots, X_m\}$ by the condition that $\widehat{\text{err}}(c) = 0$. But how many concepts in $\mathcal{P}$ have true error less than ϵ? Let

$$\mathcal{M} = \{c \in \mathcal{C} : \text{err}(c) > \epsilon \text{ and } \widehat{\text{err}}(c) = 0\}$$

denote the set of "misleading" concepts. We want to show that the fraction of promising concepts that are misleading is large. Let C be a uniform random concept in $\mathcal{P}$. We have

$$\begin{aligned}
\mathbf{E}[|\mathcal{M}|] &= \mathbf{E}\left[|\mathcal{P}|\frac{|\mathcal{M}|}{|\mathcal{P}|}\right] \\
&= 2^m \mathbf{Pr}[\text{err}(C) > \epsilon] \\
&= 2^m \mathbf{Pr}\left[\sum_{i=1}^{2m} \mathbb{1}_{\{C(x_i) \neq c^*(x_i)\}} > 2m\epsilon\right] \\
&= 2^m \left(1 - \mathbf{Pr}\left[\sum_{i=1}^{2m} \mathbb{1}_{\{C(x_i) \neq c^*(x_i)\}} \leq 2m\left\{\frac{1}{2} - \left(\frac{1}{2} - \epsilon\right)\right\}\right]\right) \\
&\geq 2^m \left[1 - \exp\left(-\frac{8m^2\left(\frac{1}{2} - \epsilon\right)^2}{2m}\right)\right] \\
&= 2^m \left[1 - \exp\left(-4m\left(\frac{1}{2} - \epsilon\right)^2\right)\right]
\end{aligned}$$

where the final inequality follows by Hoeffding's bound. Now taking $\epsilon \leq \frac{1}{4}$ and $m \geq 20$ it follows that $\mathbf{E}[|\mathcal{M}|] \geq (0.99)2^m$, i.e. on average greater than 99% of the promising concepts are misleading.