CSC 588: Machine learning theory **Spring 2022**

Lecture 6: ERM Analysis; PAC Learning Infinite Hypothesis Classes

Lecturer: Chicheng Zhang *Scribe: Ryan Murphy*

1 Analysis of ERM

Theorem: Empirical Risk Minimization (ERM) with hypothesis class $\mathcal{H}$ with m training examples drawn i.i.d. from $\mathcal{D}$ such that $m \geq f(\varepsilon, \delta) = \frac{2}{\varepsilon^2}\left(\ln|\mathcal{H}| + \ln\frac{2}{\delta}\right)$ outputs $\hat{h} \in \mathcal{H}$ such that with probability $1-\delta$

$$\operatorname{err}\left(\hat{h}, \mathcal{D}\right) \leq \min_{h \in \mathcal{H}} \operatorname{err}(h, \mathcal{D}) + \varepsilon. \tag{1}$$

Proof: We construct a favorable event E such that $E = \bigcap_{h\in\mathcal{H}} \left\{|\operatorname{err}(h,\mathcal{S}) - \operatorname{err}(h,\mathcal{D})| \leq \frac{\varepsilon}{2}\right\}$. Event E represents the case where the true error of every hypothesis in $\mathcal{H}$ differs from the empirical error by at most $\frac{\varepsilon}{2}$. In the case of event E, by the "key observation" from Lecture 4 with $\mu = \frac{\epsilon}{2}$,

$$\operatorname{err}(\hat{h}, \mathcal{D}) \leq \min_{h \in \mathcal{H}} \operatorname{err}(h, \mathcal{D}) + \varepsilon. \tag{2}$$

To show $\mathbb{P}(E) \geq 1-\delta$, it suffices to show $\mathbb{P}(\bar{E}) \leq \delta$. We construct another event B_h such that

$$B_h = \left\{|\operatorname{err}(h,\mathcal{S}) - \operatorname{err}(h,\mathcal{D})| > \frac{\varepsilon}{2}\right\} \tag{3}$$

$$\overline{E} = \bigcup_{h\in\mathcal{H}} B_h \tag{4}$$

Where $\forall h \in \mathcal{H}$ empirical error deviates from generalization error by at most $\frac{\varepsilon}{2}$. What can we say about chance of B_h? If we can bound B_h we can use union bound.

$$D(B_h) \leq 2\exp\left(-2 \cdot \frac{m \cdot (\frac{\varepsilon}{2})^2}{(1-0)^2}\right) = 2\exp\left(-\frac{m\varepsilon^2}{2}\right) \leq \frac{\delta}{|\mathcal{H}|} \tag{5}$$

View training error as an average of i.i.d. Bernoulli random variables. The union bound implies

$$\mathbb{P}(\overline{E}) \leq \sum_{h\in\mathcal{H}} \mathbb{P}(B_h) \leq |\mathcal{H}| \cdot \frac{\delta}{|\mathcal{H}|} = \delta. \ \square \tag{6}$$

So we see the union bound trick does not "blow up" sample complexity by too much – the sample complexity's dependence on $|\mathcal{H}|$ is only logarithmic.

Exercise: If we use Chebyshev's inequality to bound B_h what sample complexity guarantees can we show for ERM? (Hint: It will still give a valid upper bound but the bound will be worse.)

If we instead fix the sample budget what guarantee can we make about ε?

Corollary: Set ε such that $\frac{2}{\varepsilon^2}\ln\frac{2|\mathcal{H}|}{\delta} = m$

$$\varepsilon = \sqrt{\frac{2\ln|\mathcal{H}| + 2\ln\frac{2}{\delta}}{m}} \tag{7}$$

Therefore, ERM with $\mathcal{H}$ with a fixed budget of m i.i.d. training examples from $\mathcal{D}$, outputs classifier $\hat{h} \in \mathcal{H}$ such that with probability $1-\delta$,

$$\text{err}(\hat{h}, \mathcal{D}) \le \min_{h\in\mathcal{H}} \text{err}(h, \mathcal{D}) + \sqrt{\frac{2\ln|\mathcal{H}| + 2\ln\frac{2}{\delta}}{m}}. \tag{8}$$

We see the error bound is monotonically decreasing with sample size, and also depends on the log of the size of the hypothesis class; this is called the "Occam's Razor" bound (Occam's Razor $\implies$ a short explanation tends to be more valid than a long explanation.)

Question: In the context of Occam's Razor, what does $\mathcal{H}$ actually mean? What if we double the possible $\mathcal{H}$?

We can think of D as some natural phenomenon, and we would like to pick a good explanation h for it (i.e., $\text{err}(h, D)$ is small). $\mathcal{H}$ is a set of candidate explanations. The cardinality of hypothesis class $|\mathcal{H}|$ is complexity of explanation; the error $\text{err}(\hat{h}, \mathcal{D})$ is power or validity of explanation $\hat{h}$.

Caveats of using Hoeffding's Inequality: an example

Consider data drawn from a uniform distribution D such that $\mathcal{X} \sim \text{uniform}([0,1])$. There is a threshold at $\frac{1}{2}$ and all samples less than $\frac{1}{2}$ are negative, and all greater than $\frac{1}{2}$ are positive.

Algorithm: (Memorization)

Given training set $\mathcal{S}$, return a classifier that predicts perfectly on the training set. If sample is not in training set always return +1:

$$\hat{h}(x) = \begin{cases} y_i & x = x_i \quad \text{for some } i \\ +1 & \text{otherwise} \end{cases} \tag{9}$$

Questions:

1. What is $\hat{h}$'s training error rate? A direct calculation yields that, $\text{err}(\hat{h}, \mathcal{S}) = \frac{1}{m}\sum_{i=1}^{m} I(\hat{h}(x_i) \ne y_i) = 0$
2. Is it true that $\forall\delta > 0$

$$\mathbb{P}\left(\left|\text{err}(\hat{h}, \mathcal{S}) - \text{err}(\hat{h}, \mathcal{D})\right| \le \sqrt{\frac{\ln\frac{2}{\delta}}{m}}\right) \ge 1-\delta? \tag{10}$$

3. What is $\hat{h}$'s generalization error rate? A direct calculation yields that, $\text{err}(\hat{h}, \mathcal{D}) = \frac{1}{2}$ because $\mathbb{P}\left(\hat{h}(x) = +1\right) = 1$

We now come back to answer question 2. To use Hoeffding's inequality to analyze

$$\text{err}(\hat{h}, \mathcal{S}) = \frac{1}{m}\sum_{i=1}^{n} I\left(\hat{h}(x_i) \ne y_i\right), \tag{11}$$

it must be the case that $I\left(\hat{h}(x_i) \ne y_i\right)$ are being drawn i.i.d. from Bernoulli($\text{err}(\hat{h}, \mathcal{D})$), but since the mean parameter of this Bernoulli distribution is $\text{err}(\hat{h}, \mathcal{D}) = \frac{1}{2} \ne 0$, there is some contradiction. The problem is that, to apply Hoeffding's Inequality, we need $\hat{h}$ to be chosen before seeing the sample set $\mathcal{S}$. In the case of the memorization example $\hat{h}$ depends on S.

2 Infinite hypothesis classes: PAC learning variants

We have seen $|\mathcal{H}| \leq \infty$ implies $\mathcal{H}$ is (agnostic) PAC learnable. What if $|\mathcal{H}| = \infty$?

Example: a PAC learnable hypothesis class with infinite cardinality.

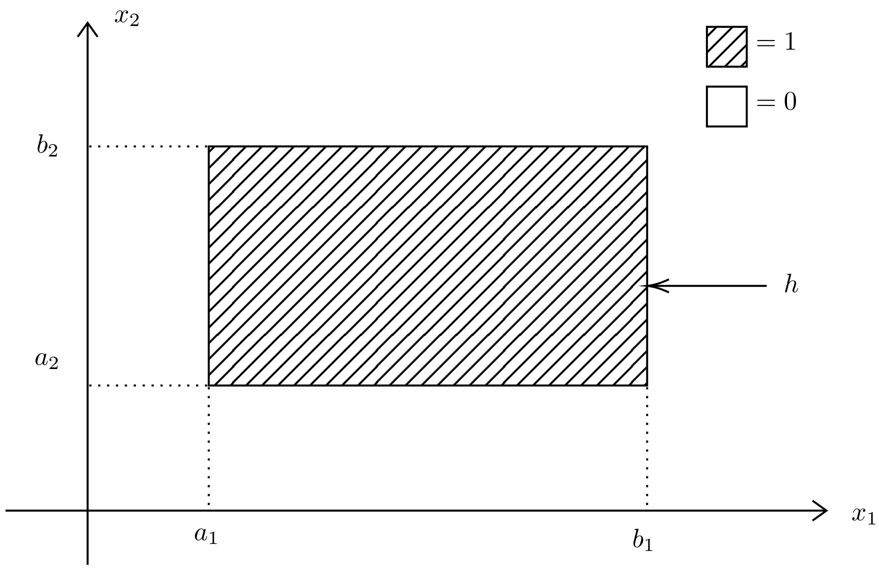


Figure 1: Example of hypothesis class of 2d rectangles which we will show is PAC learnable despite having an infinite cardinality.

Consider the case where samples are from $\mathbb{R}^2$, $\mathcal{X} = \mathbb{R}^2$, and have binary labeling such that $y = \{0, 1\}$. We consider classifiers which are defined by axis-aligned rectangular regions.

$$\mathcal{H} = \{\text{rectangles}\} = \{h_{a_1,b_1,a_2,b_2} : a_1 \leq b_1 \ \&\ a_2 \leq b_2\} \tag{12}$$

$$h_{a_1,b_1,a_2,b_2}(x) = \begin{cases} 1 & x_1 \in [a_1, b_1] \ \&\ x_2 \in [a_2, b_2] \\ 0 & \text{otherwise} \end{cases} \tag{13}$$

Also, consider $\mathcal{D}$ realizable by $\mathcal{H}$ so data can be separated by a rectangle correctly.

Finally we consider algorithm $\mathcal{A}$. Given training set $\mathcal{S}$ return classifier $\hat{h}$ as the smallest rectangle enclosing all positive examples in the training set. This is the "closure" algorithm which attempts to output a minimum covering rectangle.

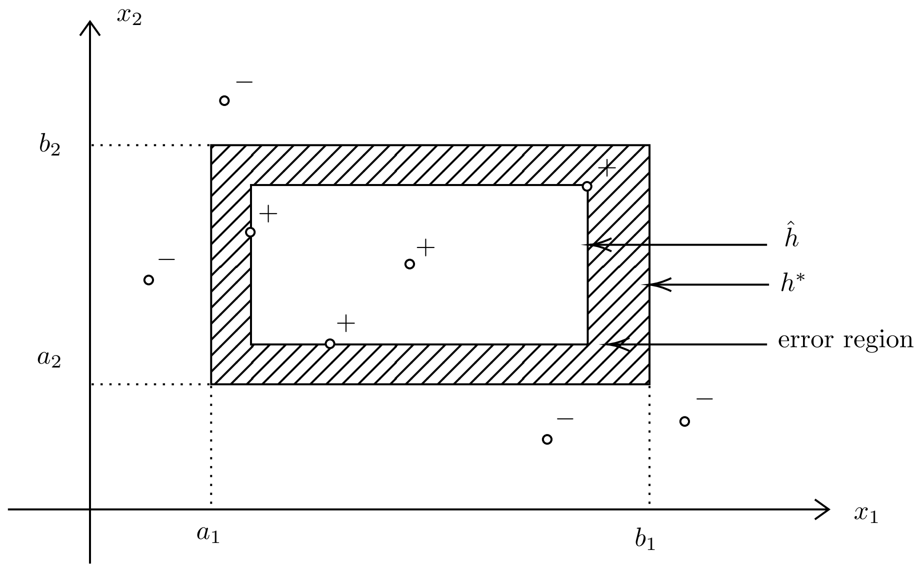


Figure 2: Example of estimator $\hat{h}$ based on a sample dataset $\mathcal{S}$. Note that the optimal classifier h^* contains $\hat{h}$ with an error region.

We can analyze this algorithm's performance in the PAC framework.

Claim: If $\mathcal{A}$ receives a training set $\mathcal{S}$ of size $m \geq \frac{4}{\varepsilon} \ln \frac{4}{\delta}$ i.i.d. from $\mathcal{D}$ then with probability $1-\delta$, $\text{err}(\hat{h}, \mathcal{D}) \leq \varepsilon$. (PAC is guaranteed)

Proof:

1. For $h \in \mathcal{H}$, define $R(h) =$ (rectangle associated with h). As shown in figure 2, $R(\hat{h}) \subseteq R(h^*)$. This implies that h cannot make false negatives, only false positives.

$$\forall x \text{ if } \hat{h}(x) = 1 \implies h^*(x) = 1 \tag{14}$$

$$h^* = h_{a_1^*, b_1^*, a_2^*, b_2^*} \tag{15}$$

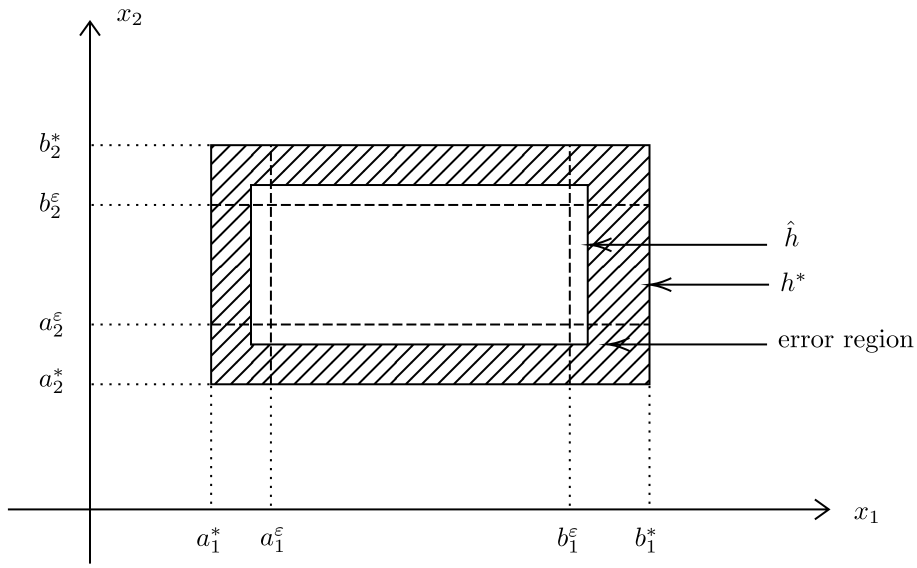


Figure 3: Example of thresholds which bound the error of each edge of the rectangle.

2. We can define a threshold very close to a_1^*, a_1^ε such that

$$\mathbb{P}\left(x \in [a_1^*, a_1^\varepsilon] \times [a_2^*, b_2^*]\right) = \frac{\varepsilon}{4} \tag{16}$$

Similarly define b_1^ε, a_2^ε, and b_2^ε; also denote by R_1, R_2, R_3, and R_4 the associated regions, e.g. $R_1 = [a_1^*, a_1^\varepsilon] \times [a_2^*, b_2^*]$, etc.

Observation: If $\mathcal{S}$ contains examples in all of R_1, R_2, R_3, R_4 then

$$\begin{aligned} \text{err}(\hat{h}, \mathcal{D}) &= \mathbb{P}\left(\{\hat{h}(x) = 0, h^*(x) = 1\}\right) & (17) \\ &\leq \mathbb{P}\left(R_1 \cup R_2 \cup R_3 \cup R_4\right) & (18) \\ &\leq \sum_{j=1}^{4} \mathbb{P}(R_j) = 4 \cdot \frac{\varepsilon}{4} = \varepsilon & (19) \end{aligned}$$

3. Define an event E such that $E = \{\forall j = 1, ..., 4, \mathcal{S} \text{ contains example in } R_j\}$.

Rest of proof is left as an exercise:

Exercise:
Write E as an intersection

$$E = \bigcap_{j=1}^{4} \{\mathcal{S} \text{ contains example in} R_j\}$$

Using DeMorgan's law and union bound, show $\mathbb{P}(E) \geq 1 - \delta$.

3 General Characterization of Infinite Hypothesis Classes

We will describe the VC dimension (VC coming from authors names Vapnik and Chervonenkis). This will give us a more general tool to characterize the complexity of a hypothesis class that goes beyond hypothesis class sizes (we have seen that size fails for characterizing the complexity of infinite hypothesis classes.)

Definition:

Given hypothesis class $\mathcal{H} \subseteq (\mathcal{X} \to \{\pm 1\})$ and a sequence of unlabeled examples $\mathcal{S} = (x_1, ..., x_n)$ we define the projection of $\mathcal{H}$ on $\mathcal{S}$ as

$$\Pi_{\mathcal{H}}(\mathcal{S}) = \{(h(x_1), ..., h(x_n)) : h \in \mathcal{H}\} \tag{20}$$

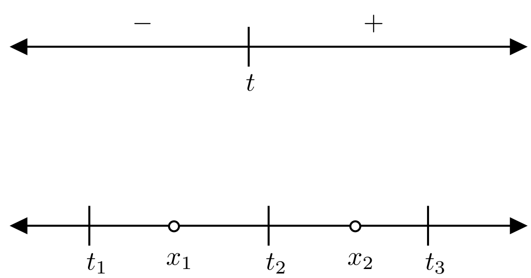
The size of this set will be the combination of possible labellings, which can be trivially bounded by:

$$|\Pi_{\mathcal{H}}(\mathcal{S})| \leq 2^n. \tag{21}$$

Example:

We consider the case where the data is a set of values from $\mathbb{R}$, each with a label in ± 1. We consider the hypothesis class a threshold value which splits the real numbers.

$$\mathcal{H} = \{\text{thresholds}\} = \{h_t : t \in \mathbb{R}\} \tag{22}$$



Possible labels:

$$\begin{aligned} t_1 &= ++ \\ t_2 &= -+ \\ t_3 &= -- \end{aligned}$$

Figure 4: Example of threshold hypothesis function and possible labellings for a dataset of size 2.

$$\Pi_{\mathcal{H}}(\mathcal{S}) = \{(-1, -1), (+1, +1), (-1, +1)\} \tag{23}$$

If $|\Pi_{\mathcal{H}}(\mathcal{S})| = 2^n$ then $\mathcal{H}$ "shatters" $\mathcal{S}$. The example above shows that $\mathcal{H}$ does not shatter (x_1, x_2) but does shatter (x_1).