**CSC 588: Machine learning theory** **Spring 2021**

# Lecture 5: Agnostic PAC learning and VC theory

*Lecturer: Chicheng Zhang* *Scribe: Alonso Granados*

## 1 Analysis of ERM in nonrealizable settings

In the previous lectures, we learned that the consistency algorithm PAC learns hypothesis $\mathcal{H}$ where $\mathcal{H}$ is finite and the iid samples are $D$ realizable with respect to $\mathcal{H}$. Now we will remove the realizability assumption as in many practical problems this assumption does not hold.

Again we consider the ERM for the training set, $\hat{h} = \text{argmin}_{h\in\mathcal{H}}\, err(h, S)$, but now that we don't assume that $D$ is realizable with respect to $\mathcal{H}$. Because of noiseness of the training example It is possible that the true optimal, $h^* = \text{argmin}_{h\in\mathcal{H}}\, err(h, D)$, is not returned by ERM. In figure 1, we present an example where ERM would pick the hypothesis $h_3$ despite it is the worst hypothesis.

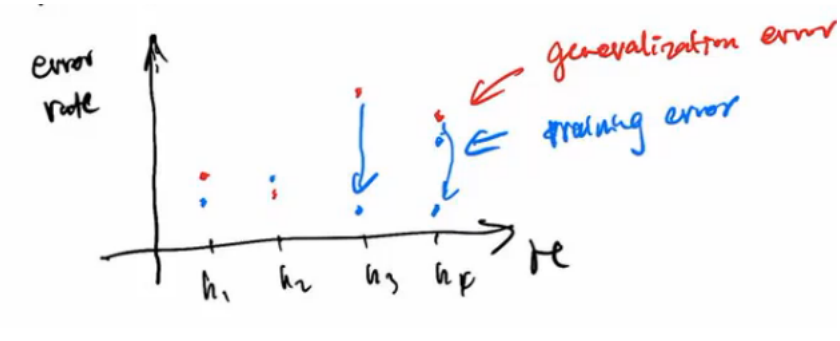


Figure 1: An example from training and generalization error.

To address this problem, we propose to bound the interval for which the training error is located with respect to the generalization error. We consider the event $A$,

$$\forall h \in \mathcal{H} : |\,\text{err}(h, S) - \text{err}(h, D)| \leq \mu$$

Assume that $A$ happens. In this setting, we would like to provide an upper bound on $\text{err}(\hat{h}, D)$. We define $\gamma = \min_{h\in\mathcal{H}} \text{err}(h, D)$ and using the optimality of ERM

$$\text{err}(\hat{h}, S) \leq \text{err}(h^*, S)$$

Using our bound assumption,

$$\text{err}(h^*, S) \leq \gamma + \mu$$

and

$$\text{err}(\hat{h}, D) \leq \text{err}(\hat{h}, S) + \mu \leq \gamma + 2\mu$$

Now we want to find $\mu$ such that event $A$ has high probability $1 - \delta$. Let's check that

$$A = \cap_{h\in\mathcal{H}}\{|\,\text{err}(h, S) - \text{err}(h, D)| \leq \mu\}$$

We consider the compliment to use the Union Bound inequality,

$$P(A^c) = P(\cup_{h\in\mathcal{H}}\{|\operatorname{err}(h,S) - \operatorname{err}(h,D)| > \mu\}) \leq \sum_{h\in\mathcal{H}} P(\{|\operatorname{err}(h,S) - \operatorname{err}(h,D)| > \mu\})$$

Using the Hoeffding's inequality,

$$P(A^c) \leq \sum_{h\in\mathcal{H}} 2e^{-2m\mu^2} \leq |\mathcal{H}|2e^{-2m\mu^2}$$

We choose $\delta = |\mathcal{H}|2e^{-2m\mu^2}$.

Therefore, with probability $1-\delta$, $A$ happens and $\operatorname{err}(\hat{h},D) \leq \min_{h\in\mathcal{H}} \operatorname{err}(h,D) + 2\sqrt{\frac{\ln|\mathcal{H}|+\ln\frac{2}{\delta}}{2m}}$. This gives the following theorem.

**Theorem 1.** *Suppose $\mathcal{H}$ is finite. If the ERM algorithm is given $m$ iid examples from $D$, then with probability $1-\delta$, its output $\hat{h}$ is such that*

$$\operatorname{err}(\hat{h},D) \leq \min_{h\in\mathcal{H}} \operatorname{err}(h,D) + 2\sqrt{\frac{\ln|\mathcal{H}| + \ln\frac{2}{\delta}}{2m}}. \tag{1}$$

*In other words, it Agnostic PAC learns hypothesis class $\mathcal{H}$ with sample complexity $m(\epsilon,\delta) = \frac{2}{\epsilon^2}(\ln|\mathcal{H}| + \ln\frac{2}{\delta})$.*

## 2 Inifinite classes can be PAC learnable

In general, if $\mathcal{H}$ is infinite, Theorem 2 will give vacuous guarantees as the error bound is infinite. Can we develop general tools for analyzing ERM for infinite hypothesis classes? First, let's see an example showing that there may be hope achieving this.

We will consider a problem where $|\mathcal{H}|$ is infinite:

$$\mathcal{X} = [0,1]$$
$$y = \{1,-1\}$$
$$\mathcal{H} = \{h_t : 2\mathcal{I}(x > t) - 1; t \in [0,1]\}$$

and $D$ realizable by $h_{t^*}$ with example $x \sim \text{uniform}([0,1])$.

Given $\epsilon$, define $t_L$ and $t_R$ such that $\epsilon$ equals the probability of interval $[t_L, t^*]$ and$[t^*, t_R]$ (Figure 2). If we have a training sample inside both $[t_L, t^*]$ and $[t^*, t_R]$, then $\operatorname{err}(\hat{h},D) \leq \epsilon$. Define $E_L$ be the event that a training sample is inside $[t_L, t^*]$; Define $E_R$ be the event that a training sample is inside $[t^*, t_R]$. We would like to set the sample size $m$ such that $P(E_L \cap E_R) \geq 1-\delta$.

To this end, again we take the complement and apply Union Bound,

$$P(E_L^c \cup E_R^c) \leq P(E_L^c) + P(E_R^c) = 2(1-\epsilon)^m \leq 2e^{-m\epsilon}.$$

Note: The final result is using the fact that we are dealing with iid uniform samples and $1 - x \leq e^{-x}$.

Therefore, if the sample size $m \geq \frac{1}{\epsilon}\ln\frac{2}{\delta}$, the above probability is at most $\delta$, which implies that $P(E_L \cap E_R) \geq 1-\delta$. This gives the following theorem.

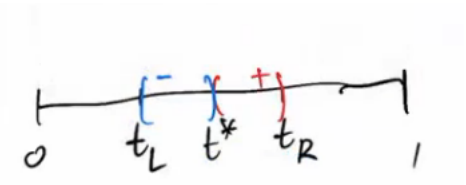


Figure 2: $\epsilon$ range around optimal solution.

**Theorem 2.** *For the previous problem the consistency algorithm learns hypothesis class $\mathcal{H}$ with sample complexity $m(\epsilon, \delta) = \frac{1}{\epsilon} \ln \frac{2}{\delta}$.*

# 3 VC Theory

VC dimension defines a complexity measure that can be used even for hypothesis classes with infinite cardinality.

**Definition 3.** *For a hypothesis class $\mathcal{H}$ (such that $\mathcal{X} = \{1, -1\}$) and sequence of examples $S = (x_1 \dots x_n)$, we define the projection of $\mathcal{H}$ on $S$ as*

$$\Pi_{\mathcal{H}}(S) = \{(h(x_1) \dots h(x_n)) : h \in \mathcal{H}\}$$

**Definition 4.** *$\mathcal{H}$ shatters $S$ if $|\Pi_{\mathcal{H}}(S)| = 2^n$.*

**Definition 5.** *The VC dimenstion of $\mathcal{H}$ ($VC(\mathcal{H})$) is $max\{n \in \mathbf{N} : \mathcal{H}$, can shatters $n$ points$\}$.*