Comp 585 Scribe 3/21

Joshua Yaffee

Spring 2024

1 Review

First, let's recall the necessary preliminaries from what we discussed in previous lectures.

1. The VC-dimension, $d = VCdim(C)$ of a concept class C is the maximal size of a set S that can be shattered by C.
2. A particular concept class has Uniform Convergence iff $\forall\ \epsilon,\ \delta \in (0,1) \exists m$ s.t. $Pr[\sup_{c \in C} |err(c) - e\hat{r}r(c)| < \epsilon] \geq 1 - \delta$
3. If C has finite VC-dimension, C has uniform convergence.

And an important lemma followed:

Lemma 1. *For class C with growth function $\tau_C(m)$ (or briefly stated, $\tau(m)$) for all distributions, D and parameters ϵ, δ and sample sets S of size m:*

$$Pr[\forall c \in C | e\hat{r}r_S(c) - err(c)| < \epsilon] > 1 - \delta$$

is satisfied when

$$\epsilon = \frac{4 + \sqrt{\log(\tau(2m))}}{\delta * \sqrt{2m}}$$

And this gives us the sample complexity for m. Last class, we began the proof and arrived at the following:

$$\mathbb{E}[\sup_{c \in C} |err(c) - e\hat{r}r_S(c)|] \leq \mathbb{E}_{S,S'} \mathbb{E}_{\sigma \in u\{1,-1\}} [\sup_{z \in C_{S \cup S'}} \frac{1}{m} | \sum_{i=1}^{m} \sigma_i (\mathbb{1}\{z_i \neq y_i\} - \mathbb{1}\{z_i' \neq y_i'\})|]$$

where $C_{S \cup S'}$ is the set of (**z**, **z'**) such that there exists a concept $c \in C$ such that $c(\mathbf{x}) = \mathbf{z}, c(\mathbf{x'}) = \mathbf{z'}$

2 Continuation of Proof

Let $A(\sigma_i) = \sigma_i(\mathbb{1}\{z_i \neq y_i\} - \mathbb{1}\{z_i' \neq y_i'\})$. Note that $\mathbb{E}[A(\sigma_i] = 0$. So,

$$\begin{aligned} Pr\left[|\sum_{i=1}^{m} A(\sigma_i)| > m\alpha \right] &= Pr\left[\frac{1}{m} |\sum_{i=1}^{m} A(\sigma_i)| - \mathbb{E}[A(\sigma_i)] > \alpha \right] \\ &\leq 2e^{-2m\alpha^2} \end{aligned}$$

Thus, we can now use union bound:

$$\forall z \in C_{S \cup S'} Pr\left[\exists z : \frac{1}{m} |\sum_{i=1}^{m} A(\sigma_i)| - \mathbb{E}[A(\sigma_i)] > \alpha \right] \leq 2\tau(2m) e^{-2m\alpha^2}$$

And we use the lemma found in the next section with $a = \frac{1}{sqrt2m}$, $b = \tau(2m)$ to bound the probability above by:

$$\frac{2 + \sqrt{\log(\tau(2m))}}{\sqrt{2m}}$$

and by bounding this expression above by ϵ, our Lemma is proved. Next, let's take a look at the lemma used to make this final conclusion.

Lemma 2. *X is a random variable and x' is a scalar. Suppose $\exists a > 0$, $b > e$ s.t. $\forall t\ Pr[|X - x'| > t] \leq 2be^{-t^2/a^2}$ then,*

$$\mathbb{E}[|X - x'|] \leq a(2 + \sqrt{\log b}$$

The proof goes as follows:

$$\begin{aligned}
\mathbb{E}[|X - x'|] &= \int_t Pr[|X - x'| > t]dt \\
&\leq \sum_{i=0}^{\infty} f(t_{i-1})(t_i - t_{i-1}) \\
&\leq \sum_{i=0}^{\infty} f(t_{i-1})(t_i) \\
&\leq t_0 * 1 + \sum_{i=1}^{\infty} f(t_{i-1})(t_i) \\
&\leq t_0 + \sum_{i=i}^{\infty} Pr[|X - x'| > t_{i-1}](t_i) \\
\text{Choosing } t_i = a(i + \sqrt{\log b}) &\iff \frac{t_i^2}{a^2} = -i^2 - 2i\sqrt{\log b} - \log b \\
&\leq a\sqrt{\log b} + 2ab\sum_{i=1}^{\infty}(i + \sqrt{\log b})e^{-(i-1)} \\
&\leq a\sqrt{\log b} + 2ab\int_{i+\sqrt{\log b}}^{\infty} xe^{-(x-1)^2}dx \\
\text{Substituting} x - 1 = y& \\
&\leq a\sqrt{\log b} + 4ab\int_{\sqrt{\log b}}^{\infty} y - e^{-y^2}dy \\
&\leq a\sqrt{\log b} + \left[\frac{-e^{-y^2}}{2}\right]_{\sqrt{\log b}}^{\infty} \\
&\leq a\sqrt{\log b} + \frac{2ab}{b} \\
&\leq a(2 + \sqrt{\log b} \text{ as desired.}
\end{aligned}$$

And so as this lemma is confirmed, Lemma 1 must hold.

3 Learning in the Presence of Noise

An important observation to make at this point is aside from the realizable case, we never assumed deterministic data to prove any of our results so far. Now we introduce a new setting where we still have pairs $(x, f(x))$ but instead of observing $f(x)$ we observe $l(x)$ as defined here:

$$l(x) = \begin{cases} f(x) \text{ with prob. } 1 - \eta \\ 1 - f(x) \text{ with prob. } \eta \end{cases}$$

Using this, we can redefine PAC learnability for this type of noise.

Definition 1. *We say c is PAC learnable in the presence of noise iff $\exists$ an algorithm such that for all ϵ, δ, η the algorithm outputs $\hat{c}$ such that:*

$$err(\hat{c}) \leq min_{c \in C} err(c) + \epsilon$$

We end class with the following two observations. First, if we denote $p := Pr_{(x,y)\ D}[c(x) \neq f(x)]$ then by the law of total probability, $err(c) = (1 - p)\eta + p(1 - \eta) = p + \eta - 2p\eta$. Second, if $\eta < 1/2$ minimizing $err(c)$ w.r.t. $f(x)$ is equivalent to minimizing $err(c)$ w.r.t. $l(x)$. This concluded the 3/21 lecture.