FOSML: Lecture 7

The bound of the growth function in terms of VC dimension immediately yields a gen. bound in terms of VC dimension:

Corollary (VC dimension generalization bounds): Let $\mathcal{H}$ be a family of functions taking values in $\{-1, 1\}$, with VC dimension d. Then $\forall \delta > 0$, with probability exceeding $1-\delta$,

$$\forall h \in \mathcal{H}, \quad R(h) \leq \hat{R}_h(S) + \sqrt{\frac{2d \log\left(\frac{em}{d}\right)}{m}} + \sqrt{\frac{\log(1/\delta)}{2m}}.$$

$$= \hat{R}_h(S) + O\left(\sqrt{\frac{\log(m/d)}{m/d}}\right).$$

- In particular, as $d \to \infty$, our generalization bound blows up. This suggests a "principle of parsimony": all else being equal, we ought to prefer the simplest explanation, e.g. hypothesis class of smaller VC dimension.

- A VC-analysis also yields lower bounds on the generalization error, which may be interpreted as a measure of fundamental hardness.

- Recall that $R_D(h_S, f)$ is the generalization error for a function $f \in \mathcal{H}$ with respect to the predictor h_S for a sample $S \sim D$.

Theorem (lower bound, realizable case): Let $\mathcal{H}$ be a hypothesis class with VC-dimension $d > 1$. Then $\forall m \geq 1$, and any learning algorithm A, there exists a distribution D over X and a target function $f \in \mathcal{H}$ such that $\mathbb{P}_{S \sim D}\left(R_D(h_S, f) > \frac{d-1}{32m}\right) \geq \frac{1}{100}$.

Proof: Since $VCdim(\mathcal{H}) = d$, $\exists$ a set $\overline{X} = \{x_0, x_1, \dots, x_{d-1}\}$ that is shattered by $\mathcal{H}$. For any $\varepsilon > 0$, we choose D such that its support is reduced to the shattered set $\overline{X}$ and so that one point (WLOG, let it be x_0) is sampled with probability $1 - 8\varepsilon$, and all other points $(x_1, \dots, x_{d-1})$ with probability $\frac{8\varepsilon}{d-1}$. That is, D is s.t.

(a.) $\mathbb{P}_{x \sim D}(x = x_0) = 1 - 8\varepsilon$,

(b.) $\mathbb{P}_{x \sim D}(x = x_i, i = 1, \dots, d-1) = \frac{8\varepsilon}{d-1}$.

For this choice of D, most samples would contain x_0, and $\mathcal{A}$ cannot do better than flip a coin on the points $\{x_i\}_{i=1}^{d-1}$ which do not appear in the training set. We now proceed to show that sufficiently many of these points $\{x_i\}_{i=1}^{d-1}$ are not in the sample, and thus that the generalization error is lower bounded in terms of d.

Suppose WLOG that $\mathcal{A}$ makes no error on x_0. Let $S \sim D$ be a sample, and let $\overline{S} = S \cap \{x_1, \dots, x_{d-1}\}$. Let $\mathcal{S}$ be the set of samples S of size m s.t. $|\overline{S}| \leq \frac{d-1}{2}$.

Now, fix a sample $S \in \mathcal{S}$ and let V be the uniform distribution over all labelings $f : \overline{X} \to \{0,1\}$, which are all in $\mathcal{H}$ since $\overline{X}$ is shattered by $\mathcal{H}$. We estimate:

$$\begin{aligned}
\mathop{\mathbb{E}}_{f \sim V}\left(R_D(h_S, f)\right) &= \sum_f R_D(h_S, f)\, \mathbb{P}_V(f) \\
&= \sum_f \sum_{x \in \overline{X}} \mathbb{1}_{\{h_S(x) \neq f(x)\}}\, \mathbb{P}_D(x)\, \mathbb{P}_V(f) \\
&\geq \sum_f \sum_{x \notin \overline{S}} \mathbb{1}_{\{h_S(x) \neq f(x)\}}\, \mathbb{P}_D(x)\, \mathbb{P}_V(f) \\
&= \sum_{x \notin \overline{S}} \left(\sum_f \mathbb{1}_{\{h_S(x) \neq f(x)\}}\, \mathbb{P}_V(f) \right) \mathbb{P}_D(x)
\end{aligned}$$

$$= \sum_{x \notin \bar{S}} \left(\frac{1}{2}\right) \mathbb{P}_D(x)$$

since all possible labelings for f are considered uniformly by U

$$= \frac{1}{2} \sum_{x \notin \bar{S}} \mathbb{P}_D(x)$$

$$\geq \frac{1}{2} \cdot \underbrace{\left(\frac{d-1}{2}\right)}_{|\bar{S}|} \cdot \underbrace{\left(\frac{8\varepsilon}{d-1}\right)}_{\mathbb{P}_D(x=x_i)} = 2\varepsilon.$$

upper bound on

So, we have shown that $\forall S \in \mathcal{S}$, $\mathop{\mathbb{E}}_{f \sim U}\left(R_D(h_S, f)\right) \geq 2\varepsilon$

$$\Rightarrow \mathop{\mathbb{E}}_{S \sim \mathcal{S}}\left(\mathop{\mathbb{E}}_{f \sim U}\left(R_D(h_S, f)\right)\right) \geq 2\varepsilon.$$

Fubini Theorem: $\Rightarrow \mathop{\mathbb{E}}_{f \sim U}\left(\mathop{\mathbb{E}}_{S \sim \mathcal{S}}\left(R_D(h_S, f)\right)\right) \geq 2\varepsilon.$

$\int\int = \int\int$ under certain conditions

In particular, $\mathop{\mathbb{E}}_{S \sim \mathcal{S}}\left(R_D(h_S, f_0)\right) \geq 2\varepsilon$ for some $f_0 \in H$. Note that $R_D(h_S, f_0) \leq \mathbb{P}_D\left(\bar{X} - \{x_0\}\right)$, so that we may analyze this expected value as

$$\mathop{\mathbb{E}}_{S \in \mathcal{S}}\left(R_D(h_S, f_0)\right) = \sum_{S : R_D(h_S, f_0) \geq \varepsilon} R_D(h_S, f_0)\, \mathbb{P}\left(R_D(h_S, f_0)\right)$$

$$+ \sum_{S : R_D(h_S, f_0) < \varepsilon} \quad " \qquad "$$

$$\leq \mathbb{P}_D\left(\bar{X}\setminus\{x_0\}\right)\mathbb{P}_{S\in\mathcal{S}}\left(R_D(h_S,f_0)\geq\varepsilon\right)+\varepsilon\,\mathbb{P}_{S\in\mathcal{S}}\left[R_D(h_S,f_0)<\varepsilon\right]$$

$$\leq 8\varepsilon\,\mathbb{P}_{S\in\mathcal{S}}\left(R_D(h_S,f_0)\geq\varepsilon\right)+\varepsilon\left(1-\mathbb{P}_{S\in\mathcal{S}}\left(R_D(h_S,f_0)\geq\varepsilon\right)\right)$$

$$=7\varepsilon\,\mathbb{P}_{S\in\mathcal{S}}\left(R_D(h_S,f_0)\geq\varepsilon\right)+\varepsilon$$

We have thus shown that $2\varepsilon \leq 7\varepsilon\,\mathbb{P}_{S\in\mathcal{S}}\left(R_D(h_S,f_0)\geq\varepsilon\right)+\varepsilon$

$$\Rightarrow \frac{1}{7}\leq \mathbb{P}_{S\in\mathcal{S}}\left(R_D(h_S,f_0)\geq\varepsilon\right)$$

Thus, $\mathbb{P}_S\left(R_D(h_S,f_0)\geq\varepsilon\right)\geq \mathbb{P}_{S\in\mathcal{S}}\left(R_D(h_S,f_0)\geq\varepsilon\right)\mathbb{P}(S\in\mathcal{S})$

$$\geq \tfrac{1}{7}\,\mathbb{P}(S\in\mathcal{S}).$$

We will conclude by lowering bounding $\mathbb{P}(\mathcal{S})=\mathbb{P}(S\in\mathcal{S})$. Recall the multiplicative Chernoff bound: "for any $X_1,\dots,X_m$ independent r.v. distributed according to D with mean p and support $[0,1]$, $\forall\,\gamma\in[0,\frac{1}{p}-1]$, the following hold for $\hat{p}=\frac{1}{m}\sum_{i=1}^{m}X_i$: $\mathbb{P}(\hat{p}\geq(1+\gamma)p)\leq\exp(-mp\gamma^2/3)$; $\mathbb{P}(\hat{p}\leq(1-\gamma)p)\leq\exp(-mp\gamma^2/2)$." Applying this, we see: $1-\mathbb{P}(\mathcal{S})=\mathbb{P}(S_m\geq 8\varepsilon(1+\gamma))\leq\exp(-8\varepsilon m\gamma^2/3)$. Setting $\varepsilon=(d-1)/32m$ and $\gamma=1$ yields

$$\mathbb{P}\left(S_m\geq\frac{d-1}{2}\right)\leq\exp\left(-(d-1)/12\right)\leq\exp(1/12)\leq 1-7\delta \text{ for } \delta\leq 1/100.$$ We conclude $\mathbb{P}(\mathcal{S})\leq 7\delta \Rightarrow \mathbb{P}_S\left(R_D(h_S,f_0)\geq\varepsilon\right)\geq\delta$, as desired. ∎