F. SML: Lecture 3

ex: Consider $X = \{0,1\}^n$, the set of Boolean vectors of length n. Let $\mathcal{U}_n$ be the concept class formed as the power set of X, i.e. the set of all subsets of X. $\mathcal{U}_n$ is quite large: $|X| = 2^n$

$$\Rightarrow |\mathcal{U}_n| = 2^{2^n}$$ But, still finite.

How does our Theorem play with this concept class $\mathcal{U}_n$? Well, to guarantee a consistent hypothesis, the concept class must be contained in the hypothesis class, i.e.

$$\mathcal{U}_n \subset \mathcal{H} \Rightarrow |\mathcal{U}|_n \leq |\mathcal{H}|.$$

($\mathcal{U}_n$: things we are trying to learn; $\mathcal{H}$: possible hypotheses)

This puts a lower bound on the size of any consistent hypothesis class: $2^{2^n} \leq |\mathcal{H}|$. The sample bound in our Theorem is then quite bad:

$$m \geq \frac{1}{\varepsilon}\left(\log|\mathcal{H}| + \log\left(\frac{1}{\delta}\right)\right)$$

$$\geq \frac{1}{\varepsilon}\left(\log\left(2^{2^n}\right) + \log\left(\frac{1}{\delta}\right)\right)$$

$$= \frac{1}{\varepsilon}\left(\log(2)\cdot 2^n + \log\left(\frac{1}{\delta}\right)\right).$$

This does not imply PAC learning (it doesn't imply not PAC learning either, though in

fact $\mathcal{U}_n$ is not PAC learnable.) Indeed, n is the cost of encoding an element of X, so the sample complexity guaranteed by our theorem is exponential in the size of an element of $X \Rightarrow$ PAC learning is not assured.

- Part of the problem here is requiring consistency. Indeed, this forces $\mathcal{H}$ to be quite large in the above example (size $\sim 2^{2^n}$). This suggests relaxing to inconsistent algorithms.

- To do so, recall Hoeffding's inequality:

Theorem (Hoeffding's Inequality): Let $\{X_i\}_{i=1}^m$ be independent random variables with X_i taking values in $[a_i, b_i] \subset \mathbb{R}$. Then $\forall \varepsilon > 0$, the sum $S_m = \sum_{i=1}^m X_i$ satisfies:

(1.) $$\mathbb{P}\left((S_m - \mathbb{E}(S_m)) \geq \varepsilon \right) \leq \exp\left(-2\varepsilon^2 / \sum_{i=1}^m (b_i - a_i)^2 \right)$$

(2.) $$\mathbb{P}\left((S_m - \mathbb{E}(S_m)) \leq \varepsilon \right) \leq \exp\left(-2\varepsilon^2 / \sum_{i=1}^m (b_i - a_i)^2 \right).$$

Remark: Hoeffding's Inequality is an example of a concentration inequality, which describes how a random variable (in this case S_m) concentrates around its mean (in this case $\mathbb{E}(S_m)$). We will find use for such results over and over again. They are major tools in statistical and machine learning.

- From Hoeffding, we immediately get:

Corollary: Fix $\varepsilon > 0$. For any hypothesis $h: X \to \{0,1\}$, the following inequalities hold:

(1.) $\underset{S \sim D}{\mathbb{P}}\left(\hat{R}_S(h) - R(h) \geq \varepsilon \right) \leq \exp\left(-2\varepsilon^2 m\right)$

(2.) $\mathbb{P}\left(\hat{R}_S(h) - R(h) \leq \varepsilon \right) \leq \exp\left(-2\varepsilon^2 m\right)$.

Remark: Taking a union yields

$$\mathbb{P}\left(\left| \hat{R}_S(h) - R(h) \right| \geq \varepsilon \right) \leq 2\exp\left(-2\varepsilon^2 m\right).$$

In other words, the empirical error of a hypothesis h concentrates around its (true) generalization error. Clearly we expect the discrepancy to decay with the sample size m. Hoeffding establishes this decay is exponential.

- To rephrase:

Theorem (Generalization bound, $|\mathcal{H}|=1$): Fix a hypothesis $h: X \to \{0,1\}$. Then for any $\delta > 0$, $\mathbb{P}\left(R(h) \leq \hat{R}_S(h) + \sqrt{\log\left(\frac{2}{\delta}\right)/2m} \right) \geq 1 - \delta$.

- The question, then, is how to generalize this to account for $|\mathcal{H}| \geq 2$. As in the case where $\mathcal{A}$ is consistent, we shall require a uniform bound over $h \in \mathcal{H}$.

Theorem (Learning bound, $|\mathcal{H}|$ finite, inconsistent): Let $|\mathcal{H}| < \infty$ be a finite hypothesis class. Then $\forall \delta > 0$,

$$\mathbb{P}\left(R(h) \leq \hat{R}_S(h) + \sqrt{\frac{\log|\mathcal{H}| + \log\frac{2}{\delta}}{2m}}, \ \forall h \in \mathcal{H} \right) \geq 1 - \delta.$$

Proof: Let $\{h_i\}_{i=1}^{|\mathcal{H}|}$ be the (finitely many) elements of $\mathcal{H}$. Then:

$$\mathbb{P}(\exists h \in \mathcal{H} \mid |\hat{R}_S(h) - R(h)| > \varepsilon)$$

$$= \mathbb{P}\left(\bigcup_{i=1}^{|\mathcal{H}|} |R_S(h_i) - R(h_i)| > \varepsilon \right)$$

$$\leq \sum_{i=1}^{|\mathcal{H}|} \mathbb{P}\left(|R_S(h_i) - R(h_i)| > \varepsilon \right)$$

$$\leq 2 \cdot |\mathcal{H}| \cdot \exp(-2n\varepsilon^2) \quad \text{by our corollary to Hoeffding.}$$

Setting $2 \cdot |\mathcal{H}| \cdot \exp(-2n\varepsilon^2)$ to δ and solving for ε gives the desired result. ∎