17. Fundamental Thm of ML

Sunday, February 11, 2024 8:22 AM

(for binary classification)

Sauer Lemma (aka Sauer-Shelah-Perles) (Lemma 6.10 [SS] or Thm 3.17 / Cor. 3.18 Mohri)

let $d = \text{VCdim}(\mathcal{H})$, then $\forall m \in \mathbb{N}$, $\tau_{\mathcal{H}}(m) \le \sum_{i=0}^{d} \binom{m}{i}$ (define $\binom{k}{l} = 0$ if $k < l$)

Furthermore, if $m \le d$ (or $d = \infty$) this bound is vacuous (i.e. it's 2^m)

but if $m > d$,

$$\tau_{\mathcal{H}}(m) \le \sum_{i=0}^{d} \binom{m}{i} \le \left(\frac{em}{d}\right)^d,$$ a polynomial in m (vs. 2^m)

e = Euler's constant

proof sketch

note binomial thm, $(1+x)^m = \sum_{i=0}^{m} \binom{m}{i} x^i$ so $(x=1)$ $2^m = \sum_{i=0}^{m} \binom{m}{i}$

Mohri (Cor. 3.18) uses binomial thm plus $(1-x) \le e^{-x}$

[SS] uses Stirling's approx. for $n!$ and induction, see Lemma A.5 or Lemma 6.10

Relating back ...

Thm 3.3 Mohri / Thm 26.5 [SS], $Y \subseteq [0,1]$ $\forall \delta > 0$, w.p. $\ge 1-\delta$ (over m iid samples in S)

$$\forall f \in \mathcal{F}, \quad \mathbb{E}_{z \sim \mathcal{D}} f(z) - \frac{1}{m}\sum_{i=1}^{m} f(z_i) \le \begin{cases} 2 R_m(\mathcal{F}) + \sqrt{\log(\delta^{-1})/2m} \\ 2 \hat{R}_S(\mathcal{F}) + 3\sqrt{\log(2\delta^{-1})/2m} \end{cases}$$

based on $\text{Rep}_{\mathcal{D}}(\mathcal{F}, S) = \sup_{f \in \mathcal{F}} \mathbb{E} f(z) - \hat{\mathbb{E}}_S f(z)$

so w/ $\mathcal{F} = \ell \circ \mathcal{H}$ for binary loss ℓ

Thm 3.5 Mohri $Y = \{\pm 1\}$, $\mathcal{D}$ anything, $\forall \delta > 0$, w.p. $\ge 1-\delta$ over S (m iid samples)

$$\forall h \in \mathcal{H}, \quad L_{\mathcal{D}}(h) \le \hat{L}_S(h) + \begin{cases} R_m(\mathcal{H}) + \sqrt{\log(\delta^{-1})/2m} \\ \hat{R}_S(\mathcal{H}) + 3\sqrt{\log(2\delta^{-1})/2m} \end{cases}$$

and, implication of **Massart's Lemma**

$$R_m(\mathcal{H}) \le \sqrt{2\log(\tau_{\mathcal{H}}(m))/m}$$

(or... use uniform convergence, Thm 6.11 [SS]:

$$\forall \mathcal{D}, \forall \delta > 0, \text{ wp} \ge 1-\delta, \ \forall h \in \mathcal{H} \quad |L_{\mathcal{D}}(h) - \hat{L}_S(h)| \le \frac{4 + \sqrt{\log(\tau_{\mathcal{H}}(2m))}}{\delta \cdot \sqrt{2m}}$$)

So just need to bound Rademacher Complexity:

So combine Massart with Sauer to get (letting $d = \text{VCdim}(\mathcal{H})$)

$$R_m(\mathcal{H}) \le \sqrt{2\log\left(\left(\frac{em}{d}\right)^d\right)/m} = \sqrt{\frac{2d\log(em/d)}{m}} \le \sqrt{\frac{2\,\text{VCdim}(\mathcal{H})\log(e \cdot m)}{m}}$$

(for binary classification)

17a. Fundamental Thm of ML (p. 2)

Sunday, February 11, 2024 8:47 AM

§6.4 [SS]

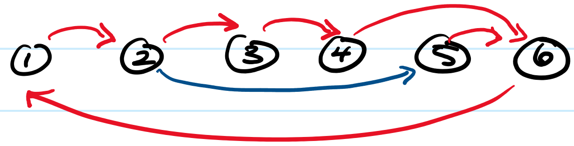
Qualitative Version

Thm 6.7 "Fundamental Thm. of Statistical (or PAC) learning" (for binary classification)

For $Y = \{0,1\}$ and the 0-1 loss function, the following are equivalent:

1. $\mathcal{H}$ has the uniform convergence property
2. Any ERM rule is a succesful (agnostic) PAC learner
3. $\mathcal{H}$ is agnostic PAC learnable
4. $\mathcal{H}$ is PAC learnable
5. Any ERM rule is a succesful PAC learner for $\mathcal{H}$
6. $\mathcal{H}$ has finite VCdimension

proof outline



Remarks

- for general learning (any loss function), uniform convergence ⇒ agnostic PAC learner
  For binary classif. (0-1 loss), vice-versa is true also!
- Some variants apply to regression (ℓ^1 or ℓ^2 loss) but not all learning tasks have such theorems
- See Thm 6.8 [SS] for a quantitative version
  i.e. agnostic PAC learnable w/ $m_{\mathcal{H}}(\varepsilon,\delta) \le C \cdot \dfrac{\text{VCdim}(\mathcal{H}) + \log(1/\delta)}{\varepsilon^2}$
  and this is tight up to a constant.
- For binary classif., $\text{VCdim} < \infty$ iff PAC learnable ... pretty neat!

Proofs

-- mostly follow from our previous results

6 ⇒ 1 via Massart's Lemma + Sauer lemma to bound Rademacher complexity, and this bounds representativeness which is basically what's needed for uniform convergence. See Thm 6.11 [SS], use Markov's Ineq. to

History Vapnik + Chervonenkis '71

As necessary condition for PAC, see Blumer, Andrzej Ehrenfeucht, David Haussler & Manfred Warmuth '89

student of

CU CS faculty (emeritus)
Founding member of CS dept.