7. Learning via Uniform Convergence

Tuesday, January 18, 2022 11:28 AM

§4.1[SS] Uniform Convergence

A sufficient condition for PAC learning
(necessary in some cases, cf. Thm. 6.7, though modern perspective is that uniform convergence is often asking for too much)

Def A training set S is an ε-representative sample (with respect to set $Z = X \times Y$, set $\mathcal{H}$, loss ℓ, distr. D) if

$$(\forall h \in \mathcal{H}) \quad |\hat{L}_S(h) - L_D(h)| \le \varepsilon$$

hence "uniform" flavor, eg. $\|\hat{L}_S - L_D\|_\infty \le \varepsilon$

Such a set is quite nice:

Lemma 4.2 If S is $\varepsilon/2$-representative then if $h_S \in ERM_{\mathcal{H}}(S)$ i.e. $\in \operatorname{argmin}_{h \in \mathcal{H}} \hat{L}_S(h)$ then

$$L_D(h_S) \le \left(\min_{h \in \mathcal{H}} L_D(h) \right) + \varepsilon$$

exactly what we want for PAC learning

proof:

let $h \in \operatorname{argmin}_{h \in \mathcal{H}} L_D(h)$

$$L_D(h_S) \le \hat{L}_S(h_S) + \varepsilon/2 \le \hat{L}_S(h) + \varepsilon/2 \quad [\text{since } h_S \text{ minimizes } \hat{L}_S]$$

($L_D(h_S) \le \hat{L}_S(h_S) + \varepsilon/2$ by $\varepsilon/2$-repr.)

$$\le L_D(h) + \varepsilon/2 + \varepsilon/2 \quad [\text{by } \varepsilon/2\text{-repr. again}]$$

$$= \left(\min_{h \in \mathcal{H}} L_D(h) \right) + \varepsilon \qquad \square$$

Building on this concept...

Def A hypothesis class $\mathcal{H}$ has the uniform convergence property (w/ respect to Z, ℓ) if $\exists\, m_{\mathcal{H}}^{UC} : (0,1)^2 \to \mathbb{N}$ s.t.
$\forall\, \varepsilon, \delta \in (0,1)$, $\forall$ prob. distr. D over $Z = X \times Y$,
if S has $m \ge m_{\mathcal{H}}^{UC}(\varepsilon, \delta)$ i.i.d samples, then

S is ε-representative with probability $\ge 1 - \delta$

Uniform over $h \in \mathcal{H}$ and all possible D

Corollary If $\mathcal{H}$ has the uniform convergence property with $m_{\mathcal{H}}^{UC}$, then ① $\mathcal{H}$ is agnostic PAC learnable

② $m_{\mathcal{H}}(\varepsilon,\delta) \le m_{\mathcal{H}}^{UC}(\varepsilon,\delta)$

③ $ERM_{\mathcal{H}}$ is a valid algorithm to learn $\mathcal{H}$

(proof is immediate)

§4.2 Finite classes ($|\mathcal{H}| < \infty$) have the uniform conv. property

Fix ε, δ. We want S to be ε-representative w/ p. $\ge 1-\delta$,

i.e. $D^m(\{S : \forall h \in \mathcal{H}, \ |\hat{L}_S(h) - L_D(h)| \le \varepsilon\}) \ge 1-\delta$

or $D^m(\underbrace{\{S : \exists h \in \mathcal{H} \text{ st. } " - - - - " > \varepsilon\}}_{\text{event } E}) < \delta$

$$E = \bigcup_{h \in \mathcal{H}} \underbrace{\{S : |\hat{L}_S(h) - L_D(h)| > \varepsilon\}}_{\text{event } E_h}$$

Same trick we did last lecture

So via union bound, $D^m(E) \le \sum_{h \in \mathcal{H}} D^m(E_h)$

Rmk: union bound aka Boole's Ineq.

So look at $D^m(E_h)$

(Before, in realizable case, $\hat{L}_S(h) = 0$ and we could break $D^m(E_h)$ into $\prod_{i=1}^m D(\cdots)$ via independence ...

this trick no longer applies)

Since $\hat{L}_S(h) = \frac{1}{m}\sum_i \ell(h, z_i)$ and $z_i \sim D$ iid,

we have $\mathbb{E}\underbrace{\hat{L}_S(h)}_{\text{r.v.}} = \underbrace{L_D(h)}_{\text{its mean}}$

So today's lecture is statistics: how likely is a r.v. to deviate significantly from its mean?

= Statistics Interlude =

1) as $m \to \infty$, the law of large numbers says $\hat{L}_{S(m)}(h) \to L_D(h)$, but there's no rate

2) Central Limit Thm says $\sqrt{m}\,\hat{L}_S(h) \sim$ Normal as $m \to \infty$

(basic CLT has no rate, but Berry-Esseen CLT does)

Could we use facts about normal distr.?

No, not good bounds: see Vershynin book

3) What about Chebyshev's inequality?

Recall Markov's Ineq: if $Z \ge 0$ is a r.v. then

$\mathbb{P}[Z \ge a] \le \frac{1}{a} \cdot \mathbb{E}[Z]$

Let θ be a r.v. w/ variance σ^2, then Chebyshev's Ineq. says

$$\forall\ \varepsilon \geq 0, \quad \mathbb{P}[\ |\theta - \mathbb{E}[\theta]| > \varepsilon\ ] \leq \sigma^2/\varepsilon^2$$

Let's see if we can apply to our situation

Define $\theta_i = \ell(h, z_i)$, a r.v. w/ mean $L_D(h)$

So $\theta = \hat{L}_S(h) = \frac{1}{m}\sum_i \theta_i$

What is $\mathrm{Var}(\theta)$? Suppose $\mathrm{Var}(\theta_i) = \sigma^2$

$$\mathrm{Var}\left(\tfrac{1}{m}\textstyle\sum_i \theta_i\right) = \tfrac{1}{m^2}\mathrm{Var}\left(\textstyle\sum_i \theta_i\right)$$
$$= \tfrac{1}{m^2}\cdot\textstyle\sum_i \mathrm{var}(\theta_i) \quad \text{by independence}$$
$$= \tfrac{1}{m}\cdot\sigma^2$$

So via Chebyshev, $\mathbb{P}[\ |\hat{L}_S(h) - L_D(h)| > \varepsilon\ ] \leq \frac{\sigma^2}{m\varepsilon^2}$

That's for event E_h, so now union bound

$$D^m(E) \leq \sum_{h\in H} D^m(E_h) \leq |H| \cdot \frac{\sigma^2}{m\cdot\varepsilon^2} =: \delta$$

So our sample complexity m_H^{UC} would be

$$m_H^{UC}(\varepsilon, \delta) = \frac{1}{\varepsilon^2}\cdot\frac{1}{\delta}\cdot\sigma^2\cdot|H|$$

- $\frac{1}{\varepsilon^2}$: not great but workable
- $\frac{1}{\delta}$: not good
- σ^2: vague but acceptable
- $|H|$: terrible: we want/need a $\log(|H|)$ or similar

Ex: $H = \{$ 1D model, parameter is double precision number $\}$

so $|H| = 2^{64}$

then $|H|$ vs $\log(|H|)$ is a BIG DEAL

Union bound + Chebyshev = Bad Idea

4) Chernoff-style bounds: Hoeffding, Chernoff, Bernstein etc "Concentration Ineq." Exploit $\sum_i$ iid structure

Many variants: see Handout.

We'll do Hoeffding

First, some tools

Def A Rademacher (or symmetric Bernoulli) r.v. X means $\{\pm 1\}$ output

$$X = \begin{cases} 1 & \text{w.p. } 1/2 \\ -1 & \text{w.p. } 1/2 \end{cases}$$

usually $\{0,1\}$ output

and a Rademacher vector means each component is an independent Rademacher

Lemma "Exponentiated Markov"

If X is a r.v., and $\lambda \in \mathbb{R}$ a constant, $\lambda > 0$ aka Moment Generating Funct'n $MGF(\lambda)$

$$\mathbb{P}[X \geq \varepsilon] \leq \exp(-\lambda\varepsilon) \cdot \mathbb{E}[e^{\lambda X}]$$

proof

$$\mathbb{P}[X \geq \varepsilon] = \mathbb{P}[e^{\lambda X} \geq e^{\lambda\varepsilon}]$$ since $X \mapsto e^{\lambda X}$ is monotonically increasing $(\lambda > 0)$

$$\leq e^{-\lambda\varepsilon}\, \mathbb{E}[e^{\lambda X}]$$ via Markov's $(e^{m} \geq 0)$ □

Lemma "Hoeffding's Lemma"

If X is Rademacher, $\mathbb{E}[e^{\lambda X}] \leq e^{\lambda^2/2}$

proof $\mathbb{E}[e^{\lambda X}] = \frac{1}{2}e^{\lambda} + \frac{1}{2}e^{-\lambda} = \cosh(\lambda) \leq e^{\lambda^2/2}$ via Taylor Remainder Thm. or similar. □

now...

Thm (Hoeffding, simple, one-sided)

Let X_i be i.i.d Rademacher, $\vec{a} \in \mathbb{R}^n$, then $\forall\, \varepsilon > 0$,

$$\mathbb{P}\left[\sum_{i=1}^{n} a_i X_i \geq \varepsilon\right] \leq \exp\left(\frac{-\varepsilon^2}{2\,\|a\|_2^2}\right)$$

Common cases:
$a_i = 1 \Rightarrow \|a\|^2 = n$
$a_i = 1/n \Rightarrow \|a\|^2 = 1/n$

Thm (Hoeffding, more general)

Let X_i be independent and bounded $X_i \in [m_i, M_i]$, then $\forall\, \varepsilon > 0$,

$$\mathbb{P}\left[\left|\sum_{i=1}^{n} (X_i - \mathbb{E}[X_i])\right| \geq \varepsilon\right] \leq 2 \cdot \exp\left(\frac{-2\varepsilon^2}{\sum_{i=1}^{n} (M_i - m_i)^2}\right)$$

proof:

Let $\lambda > 0$ (TBD). wlog let $\|\vec{a}\|_2 = 1$

By exponentiated Markov,

$$\mathbb{P}\left[\Sigma' a_i X_i > \varepsilon\right] \leq \exp(-\lambda\varepsilon)\, \mathbb{E}\left[\exp\left(\lambda \textstyle\sum_i' a_i X_i\right)\right]$$

$$= \exp(-\lambda\varepsilon)\, \mathbb{E}\left[\textstyle\prod_i' \exp(\lambda a_i X_i)\right]$$ how exp works

$$= \exp(-\lambda\varepsilon) \textstyle\prod_i' \mathbb{E}[\exp(\lambda a_i X_i)]$$ independence

$$\leq \exp(-\lambda\varepsilon) \textstyle\prod_i' \exp\left((\lambda a_i)^2/2\right)$$ Hoeffding lemma

$$= \exp(-\lambda\varepsilon) \exp\left(\lambda^2/2 \textstyle\sum_i' a_i^2\right)$$ how exp works; $\sum_i' a_i^2 = 1$ since $\|a\|_2^2 = 1$

$$= \exp\left(-\lambda\varepsilon + \lambda^2/2\right)$$ polynomial

now choose λ (to minimize the bound)

i.e., $\lambda = \varepsilon$, so ...

$$\ldots = \exp(-\varepsilon^2/2)$$ □

Ex $X_i \overset{iid}{\sim}$ Rademacher, $i \in [m]$, so $\mathbb{E}[X_i] = 0$

$\mathbb{E}[X_i^2] = 1$ hence $\sigma^2 = 1$

1) Hoeffding $\mathbb{P}\left[\left| \frac{1}{m}\sum_{i=1}^m X_i \right| \geq \varepsilon \right] \leq 2\cdot\exp(-m\varepsilon^2/2)$ i.e. $\vec{a} \in \mathbb{R}^m$, $a_i = \frac{1}{m}$, $\|a_i\|^2 = \frac{1}{m}$

2) Chebyshev $\mathbb{P}\left[\left| \frac{1}{m}\sum_{i=1}^m X_i \right| \geq \varepsilon \right] \leq \frac{\sigma^2}{m}\cdot\frac{1}{\varepsilon^2} = \frac{1}{m\varepsilon^2}$ since $\mathrm{Var}\left(\frac{1}{m}\sum X_i\right) = \sigma^2/m$

3) Exact: by symmetry, $\mathbb{P}\left[\left|\frac{1}{m}\sum X_i\right| \geq \varepsilon \right] = 2\cdot\mathbb{P}\left[\frac{1}{m}\sum X_i \geq \varepsilon\right]$

transform: $X_i = \begin{cases} 1 \\ -1 \end{cases} \Rightarrow \tilde{X}_i = \begin{cases} 1 \\ 0 \end{cases}$ so $\tilde{X}_i = \frac{1}{2}(X_i + 1)$, $X_i = 2\tilde{X}_i - 1$

So $= 2\cdot\mathbb{P}\left[\sum_{i=1}^m (2\tilde{X}_i - 1) \geq m\cdot\varepsilon \right]$

$= 2\cdot\mathbb{P}\left[\sum \tilde{X}_i \geq \frac{m}{2}(1+\varepsilon) \right]$

$= 2\cdot\left(1 - \mathbb{P}\left[\sum \tilde{X}_i < m/2\,(1+\varepsilon) \right]\right)$

$= 2\left(1 - F(m/2\,(1+\varepsilon))\right)$ where $F(\alpha)$ is CDF of a $(1/2, m)$ binomial r.v.

Ex	True	Hoeffding	Chebyshev
$m = 2000$, $\varepsilon = 0.1$	$6.8\cdot10^{-6}$	$9.1\cdot10^{-5}$	.05
$m = 100{,}000$ $\varepsilon = .02$	$2.5\cdot10^{-10}$	$4.1\cdot10^{-9}$	.025

... going back to our uniform convergence analysis ...

So via ~~Chebyshev~~ Hoeffding, $\mathbb{P}\left[|\hat{L}_S(h) - L_D(h)| > \varepsilon \right] \leq \cancel{\frac{\sigma^2}{m\varepsilon^2}}$ $2\cdot\exp(-m\varepsilon^2/2)$

That's for event E_h, so now union bound

$$D^m(E) \leq \sum_{h\in\mathcal{H}} D^m(E_h) \leq |\mathcal{H}|\cdot 2\cdot\exp(-m\varepsilon^2/2) =: \delta$$

i.e., solving for m, $-m\varepsilon^2/2 = \log(\delta/2|\mathcal{H}|)$

$$m = 2/\varepsilon^2 \log\left(\frac{2|\mathcal{H}|}{\delta}\right)$$

Summary:

Corollary 4.6 U.C. for finite classes (binary classification, agnostic)

Let $|\mathcal{H}| < \infty$, ℓ be the 0-1 loss function, then $\mathcal{H}$ has the uniform convergence property with

$$m_{\mathcal{H}}^{UC}(\varepsilon,\delta) \leq \left\lceil \frac{1}{2\varepsilon^2}\log\left(\frac{2|\mathcal{H}|}{\delta}\right) \right\rceil$$

... and hence agnostic PAC learnable with

$$m_{\mathcal{H}}(\varepsilon,\delta) \leq m_{\mathcal{H}}^{UC}(\varepsilon/2, \delta)$$

my derivation above differs by a factor of 2, but not a big deal (don't trust me... but don't trust book either)