Compression Schemes

Alireza Kazemipour

CMPUT 654: Theoretical Foundations of Machine Learning

December 7th, 2023

Flow

What? Why? How?

- Domain set $\mathcal{X}$.
- Probability distribution over $\mathcal{X}$. $\mathcal{D}$: $\mathcal{X} \to [0, 1]$
- Label set $\mathcal{Y}$. $\mathcal{Y} = \{0, 1\}$
- True labeling function f. f: $\mathcal{X} \to \mathcal{Y}$
- Training data $\mathcal{S}^m$. $\mathcal{S}^m = ((x_1, y_1), \cdots, (x_m, y_m))$
- Concept class $\mathcal{H}$. $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$
- The learner's output $h \in \mathcal{H}$. h: $\mathcal{X} \to \mathcal{Y}$
- Measure of success
  $L_{\mathcal{D},f}(h) \coloneqq \mathbb{P}_{x\sim\mathcal{D}}[h(x) \neq f(x)] \coloneqq \mathcal{D}(\{x : h(x) \neq f(x)\})$

What? PAC Learnability $\longrightarrow$ Finite Concept Classes

- Objective:

$$\mathcal{D}^m(\{\mathcal{S}|_x : L_{D,f}(h_{\mathcal{S}}) > \epsilon\}) \leq \delta$$

- So:

$$m \geq \frac{\log(\frac{|\mathcal{H}|}{\delta})}{\epsilon}$$

- δ: Confidence Parameter
- ϵ: Accuracy Parameter

What? PAC Learnability $\longrightarrow$ Finite VC Dimension

- Objective:

$$\mathcal{D}^m(\{\mathcal{S}|_x : L_{D,f}(h_{\mathcal{S}}) > \epsilon\}) \leq \delta$$

- So:

$$m \geq \max(\frac{32d}{\epsilon}\log(\frac{16e}{\epsilon}), \frac{16}{\epsilon}\log(\frac{2}{\delta})) \tag{1}$$

$$m \geq \max(\frac{8d}{\epsilon}\log(\frac{8d}{\epsilon}), \frac{4}{\epsilon}\log(\frac{2}{\delta})) \tag{2}$$

- d: $VCdim(\mathcal{H})$
- e: Euler's number

Claim

Compression Schemes give weaker conditions for PAC learnability $\iff$ *A <u>lower</u> lower bound on the Sample Complexity*

Why? PAC Learnability $\longrightarrow$ Compression Schemes

"***Can we do better?***"

- Kernel κ. $\kappa : \bigcup_{m=k}^{\infty} \mathcal{S}^m \to \mathcal{S}^k$ (Compressor)
- Reconstructor ρ. $\rho : \mathcal{S}^k \times \mathcal{X} \to \mathcal{Y} = \{0, 1\}$ (Decompressor)

- $\forall m \geq k$, $\mathcal{S}^k$ is a subsequence of length k of $\mathcal{S}^m$
- $\forall m, x_i \in \mathcal{S}^m|_x \implies \rho(\kappa(\mathcal{S}^m), x_i) = f(x_i)$

How? Compression Schemes

- (Previously) Objective:

$$\mathcal{D}^m(\{\mathcal{S}|_x : L_{D,f}(h_{\mathcal{S}}) > \epsilon\}) \leq \delta$$

- (Now) Objective:

$$\mathcal{D}^m(\{\mathcal{S}|_x : L_{D,f}(\rho(\kappa(\mathcal{S}^m), x)) > \epsilon\}) \leq \delta$$

Let T be the collection of all k-element subsequences of the sequence $(1, 2, ..., m)$. For any $\bar{t} = (t_1, ..., t_k) \in T$:

$$A_{\bar{t}} = \{\mathcal{S}^m : \kappa(\mathcal{S}^m) = \mathcal{S}^k\}$$

$$E_{\bar{t}} = \{\mathcal{S} \in A_{\bar{t}} : P(\{x : \rho(\kappa(\mathcal{S}), x) = f(x)\}) < 1 - \epsilon\}$$

$$U_{\bar{t}} = \{\mathcal{S}^m : P(\{x : \rho(\mathcal{S}^k, x) = f(x)\}) < 1 - \epsilon\}$$

$$B_{\bar{t}} = \{\mathcal{S}^m : \text{mark } \rho(\mathcal{S}^k, x_i) = f(x), \forall x_i \text{ s.t. } i \notin t)$$

$$E_{\bar{t}} = U_{\bar{t}} \cap A_{\bar{t}} \xRightarrow{A_{\bar{t}} \subseteq B_{\bar{t}}} P(E_{\bar{t}}) \leq P(U_{\bar{t}} \cap A_{\bar{t}}) \leq \binom{m}{k}(1 - \epsilon)^{m-k}$$

- Previously:

$$m \geq \max(\frac{8d}{\epsilon}\log(\frac{8d}{\epsilon}), \frac{4}{\epsilon}\log(\frac{2}{\delta}))$$

- Now:

$$m \geq \max(\frac{4k}{\epsilon}\log(\frac{4k}{\epsilon}) + 2k, \frac{2}{\epsilon}\log(\frac{1}{\delta}))$$

Theorem

If a Concept Class is PAC-Learnable then it has a Compression Scheme of size k[1]*.*

[1] S. Moran and A. Yehudayoff. Sample compression schemes for vc classes. Journal of the ACM (JACM), 63(3):1–10, 2016.

Questions

1. Why are there different bounds based on $VCdim$?
2. Why is the Accuracy Parameter ϵ turned into Confidence-like Parameter ?
3. Why/When is $k \leq d$?
4. Is there an *Information Theoric* approach for the same purpose?

Thank You! *:)*