Open Problem: Information Complexity of VC Learning

Grégoire Fournier

MSCS, MCS 548

March 3, 2025

Motivation

- How much information does the output of a learner reveal about its input? A natural example of a quantity that can be used to quantify the revealed information is the mutual information $I(A(Z); Z)$ between the output of an algorithm A and its input Z (i.i.d samples).
- Measuring the information complexity (IC) of a learning algorithm can be very informative, as it is related to several properties or guarantees that we might wish to establish. In particular, **low information complexity entails generalization guarantees**[1] [4][2].

Plan

1. VC-dimension

2. MI and CMI

Growth function

Definition (Growth function (def 3.3))

The growth function $\Pi_H : \mathbb{N} \rightarrow \mathbb{N}$ for a hypothesis set H is defined as:

$$\forall m \in \mathbb{N}, \Pi_H = max_{\{x_1,...,x_m\}} |\{(h(x_1), ..., h(x_m) : h \in H\}|$$

$\Pi_H(m)$ is the maximum number of distinct ways in which m points can be classified using hypotheses in H.

This provides a combinatorial measure of the richness of the hypothesis set H.

Link with Rademacher complexity

Theorem (Massart (thm. 3.3))

If H takes values in $\{-1, 1\}$,

$$\mathcal{R}_m(H) \leq \sqrt{\frac{2log(\Pi_H(m))}{m}}$$

$\forall\delta > 0$, $wp \geq 1 - \delta$, $\forall h \in H$,

$$\mathcal{R}_m(H) \leq \hat{\mathcal{R}}_m(H) + \sqrt{\frac{2log(\Pi_H(m))}{m}} + \sqrt{\frac{log(1/\delta)}{2m}}$$

VC-dimension

Definition (VC-dimension (def 3.4))

The VC-dimension of H is the size of the largest set S for which all dichotomies can be realized by functions in H, i.e $\Pi_H(m) = 2^m$.

We say S is fully shattered by H.

$$VCdim(H) = \{m \in \mathbb{N} = 2^m\}|$$

$VCdim(H) = d$ if there exists a set of size d that can be fully shattered.

VC-dimension

Hypothesis classes with bounded Vapnik-Chervonenkis (VC) dimension exhibit strong uniform convergence, which implies sample-efficient PAC learning.

Theorem (Sauer (thm. 3.5))

If H takes values in $\{-1, 1\}$ with VC-dimension d, then $\forall \delta > 0$, wp $\geq 1 - \delta$, $\forall h \in H$:

$$\mathcal{R}_m(H) \leq \hat{\mathcal{R}}_m(H) + \sqrt{\frac{2d \ \log(\frac{em}{d})}{m}} + \sqrt{\frac{\log(1/\delta)}{2m}}$$

KL divergence

Definition (Kullback–Leibler divergence)

The KL divergence, D_{KL} (also called relative entropy), is a measure of how one probability distribution is different from a second, reference probability distribution.
On a probability space $\mathcal{X}$ for two distribution P, Q:

$$D_{KL}(P||Q) = \sum_{x \in \mathcal{X}} P(x) log(\frac{P(x)}{Q(x)})$$

$$D_{KL}(P||Q) \geq 0$$

Mutual information

Definition (Mutual information)

The MI of two random variables is a measure of the mutual dependence between the variables. More specifically, it quantifies the "amount of information" obtained about one random variable by observing the other random variable.

For two random variable with values in $\mathcal{X}$ and $\mathcal{Y}$:

$$I(X;Y) = D_{KL}(\mathbb{P}_{(X,Y)}||\mathbb{P}_X\mathbb{P}_Y)$$

$$I(X;Y) = \sum_{y\in\mathcal{Y}}\sum_{x\in\mathcal{X}} p_{(X,Y}(x,y) log(\frac{p_{(X,Y}(x,y)}{p_X(x)p_Y(y)})$$

Reasoning about generalization via MI

- Let $A : \mathcal{Z}^n \to \mathcal{W}$ be a randomized or deterministic algorithm.
- Let $\mathcal{D}$ be a probability distribution on $\mathcal{Z}$.

Theorem (MI of an algorithm [1])

$$|\mathbb{E}_{Z \leftarrow \mathcal{D}^n, A}[\ell(A(Z), Z) - \ell(A(Z), \mathcal{D})]| \leq \sqrt{\frac{2\, I(A(Z); Z)}{n}}$$

Reasoning about generalization via CMI

- Let $A : \mathcal{Z}^n \to \mathcal{W}$ be a randomized or deterministic algorithm, $\mathcal{D}$ be a probability distribution on $\mathcal{Z}$.
- Let $\tilde{Z} \in \mathcal{Z}^{n\times 2}$ consisting of $2n$ samples drawn independently from $\mathcal{D}$.
- Let $S \in \{0,1\}^n$ be uniformly random.
- Define $\widetilde{Z}_S \in \mathcal{Z}^n$ by $(\widetilde{Z}_S)_i = \widetilde{Z}_{i,S_i+1}$.

Then define $CMI_{\mathcal{D}}(A) := I(A(\widetilde{Z}_S); S|\widetilde{Z})$

Theorem (CMI of an algorithm [3])

$$|\mathbb{E}_{Z\leftarrow\mathcal{D}^n,A}[\ell(A(Z),Z) - \ell(A(Z),\mathcal{D})]| \leq \sqrt{2CMI_{\mathcal{D}}(A)}$$

Conclusion

Recall ERM is $argmin_{w\in\mathcal{W}}\ \ell(w,Z)$.

Theorem ([3])

Let $\mathcal{Z}=\mathcal{X}\times\{0,1\}$, $\mathcal{W}=\{:\mathcal{X}\to\{0,1\}\}$ an hypothesis class with VC dimension d. Then there exists an empirical risk minimizer $A:\mathcal{Z}^n\to\mathcal{W}$ st. $CMI(A)\leq d\ log(n)+2$.

Do all classes with bounded VC dimension admit a learner with low information complexity?

References

[1] James Zou Daniel Russo. *Controlling Bias in Adaptive Data Analysis Using Information Theory*.

[2] Cynthia Dwork et al. Preserving statistical validity in adaptive data analysis.

[3] Lydia Zakynthinou Thomas Steinke. Reasoning about generalization via conditional mutual information.

[4] Aolin Xu and Maxim Raginsky. Information-theoretic analysis of generalization capability of learning algorithms.