Learning Club Review -

Reasoning About Generalization via Conditional Mutual Information

T. Steinke & L. Zakynthinou, COLT 2020
Presented by Ela Fallik

A fundamental question in the field of machine learning is how to ensure that a learning algorithm produces output that generalizes the underlying distribution, rather than over-fitting the data. Concretely, a **standard learning scheme** includes a distribution $\mathcal{D}$ over a population $\mathcal{Z}$, a hypothesis class $\mathcal{W}$, a loss function $l : \mathcal{W} \times \mathcal{Z} \to \mathbb{R}$ and a learning algorithm $\mathcal{A}$ that given n i.i.d samples $Z \in \mathcal{Z}^n$ drawn from $\mathcal{D}$, returns an hypothesis $\mathcal{A}(Z) = w$. Our goal is to find an hypothesis $w \in \mathcal{W}$ that minimizes the true loss,

$$l(w, \mathcal{D}) = \mathbb{E}_{z' \sim \mathcal{D}} [l(w, z')]$$

However, we cannot evaluate the true loss since the distribution $\mathcal{D}$ is unknown. We instead compute an approximation based on the given samples - the empirical loss,

$$l(w, Z) = \frac{1}{n} \sum_{i=1}^{n} l(w, Z_i)$$

Thus the question of generalization becomes: How can we ensure that the empirical loss is a good indicator for the real loss? That is, that the generalization error is low?

Many methods and frameworks exists to prove generalization. The main one is the classical theory of uniform convergence, which studies the hypothesis class to set generalization bounds. However, many other methods exist which consider generalization to be a property of the algorithm, as an algorithm might generalize better than suggested by uniform convergence bounds. Some examples are compression schemes, uniform stability and differential privacy. These methods are mostly incompatible with each other, and unifying them into a single framework is the main goal of the article.

The solution proposed in the article is based on information theory. A central notion of these approaches is the **mutual information (MI)** $\mathcal{I}(\mathcal{A}(Z); Z)$: How much information the output of the algorithm holds on its input. An important property of this measurement is that low MI implies generalization, and there are a few methods to bound MI.

However, there are significant shortcomings to this method. The most critical one is that when the domain is infinite, the MI is unbounded, and thus it can be incompatible with uniform convergence and many other methods. This is a common problem, as unbounded MI is a property of many hypothesis classes.

Conditional Mutual Information

Conditional mutual information (CMI), however, is always bounded. It can be viewed as a normalization of the mutual information s.t. each data point can expose up to one bit of information. This measurement tells us how well we can recognize the input given the output of the algorithm. Specifically, we are given $2n$ random samples $\tilde{Z}$ (sampled from $\mathcal{D}$), from which n were randomly chosen to be included in the input Z. We set S to be their indices, and set $\tilde{Z}_S := Z$. The CMI $\text{CMI}_{\mathcal{D}}(\mathcal{A}) = \mathcal{I}\left(\mathcal{A}\left(\tilde{Z}_S\right); S|\tilde{Z}\right)$ measures how well we can distinguish between the actual input $\tilde{Z}_S$ and the rest of $\tilde{Z}$, given the output $\mathcal{A}\left(\tilde{Z}_S\right)$.
Investigating the basic properties of this measure, we first note that it depends on the algorithm $\mathcal{A}$ and the distribution $\mathcal{D}$, but is independent of the loss function l. Additionally, we can see that it is non negative (if $\mathcal{A}$ is independent from it's output), and bounded from above by $n \cdot \log 2$ (if $\mathcal{A}$ reveals all of its input, allowing arbitrary overfitting). It's also bounded from above by the entropy of the output $\mathcal{H}(\mathcal{A}(Z))$.
How can we use the CMI to bound the generalization error? The article first shows how the generalization error can be bounded by the CMI, and then demonstrates the unifying nature of this framework by using many of the existing methods and frameworks in order to bound the CMI.

CMI and Generalization Error

The bound on the generalization error depends on the loss function. We'll focus on the bound of the standard generalization error for bounded linear losses (theorem 2.1) for intuition, but other versions for these losses exists, and also for unbounded (in terms of range) and non-linear losses. We only note that there's a wide spectrum of bounds that can be obtained, which demonstrate the strength of this framework.

Theorem 1. *Let $\mathcal{A} : \mathcal{Z}^n \to \mathcal{W}$ be a randomized algorithm, and $l : \mathcal{W} \times \mathcal{Z} \to [0,1]$. Let $\mathcal{D}$ be a distribution on $\mathcal{Z}$. Then*

$$\left|\mathbb{E}_{Z\sim\mathcal{D}^n}\left[l\left(\mathcal{A}\left(Z\right), Z\right) - l\left(\mathcal{A}\left(Z\right), \mathcal{D}\right)\right]\right| \leq \sqrt{\frac{2}{n} CMI_{\mathcal{D}}\left(\mathcal{A}\right)}$$

This theorem relates the expected empirical loss to the expected true loss. It's proof relies heavily on the following lemma:

Lemma 1. *Let X, Y be random variables on Ω, and $f : \Omega \to \mathbb{R}$ a measurable function. Then*

$$\mathbb{E}\left[f\left(X\right)\right] \leq D\left(X||Y\right) + \log \mathbb{E}\left[e^{f(Y)}\right]$$

One step further, we get for each $t > 0$

$$\mathbb{E}\left[f\left(X\right)\right] \leq \frac{\mathrm{D}\left(X||Y\right) + \log\left(\mathbb{E}\left[e^{t\cdot f(Y)}\right]\right)}{t}$$

To use this lemma, we first note that, since the rest of $\tilde{Z}$ which is not in the input are n i.i.d samples independently sampled from $\mathcal{D}$, evaluating the required generalization error is the same as evaluating

$$\text{gen-err}_{\mathcal{D}}\left(\mathcal{A}\right) = \mathbb{E}_{\mathcal{A},\tilde{Z},S}\left[l\left(\mathcal{A}\left(\tilde{Z}_S\right),\tilde{Z}_S\right) - l\left(\mathcal{A}\left(\tilde{Z}_S\right),\tilde{Z}\backslash\tilde{Z}_S\right)\right] =: \mathbb{E}_{\tilde{Z}}\left[\mathbb{E}_{\mathcal{A},S}\left[f_{\tilde{Z}}\left(\mathcal{A}\left(\tilde{Z}_S\right),S\right)\right]\right]$$

Therefore, we can bound the generalization error by applying the lemma on $f_{\tilde{Z}}$. The algorithm will be trained on $\tilde{Z}_S$ ($W = \mathcal{A}\left(\tilde{Z}_S\right)$), and we compare $X = (W, S)$ - the case in which S partitions $\tilde{Z}$ for the evaluation of the generalization error, and $Y = (W, S')$ - some other S' (independent from S) partitions $\tilde{Z}$. Additionally, by definition of conditional mutual information, we have that

$$\text{CMI}_{\mathcal{D}}\left(\mathcal{A}\right) = \mathcal{I}\left(\mathcal{A}\left(\tilde{Z}_S\right);S|\tilde{Z}\right) = \mathbb{E}_{\tilde{Z}}\left[\mathcal{I}\left(\mathcal{A}\left(\tilde{Z}_S\right);S\right)\right] = \mathbb{E}_{\tilde{Z}}\left[\mathrm{D}\left(\mathcal{A}\left(\tilde{Z}_S\right),S||\mathcal{A}\left(\tilde{Z}_S\right),S'\right)\right]$$

Therefore, we get from the lemma, for each $t > 0$

$$\text{gen-err}_{\mathcal{D}}\left(\mathcal{A}\right) \leq \frac{\text{CMI}_{\mathcal{D}}\left(\mathcal{A}\right) + \mathbb{E}_{\tilde{Z}}\left[\log\mathbb{E}_{\mathcal{A},S,S'}\left[e^{tf_{\tilde{Z}}\left(\mathcal{A}\left(\tilde{Z}_S\right),S'\right)}\right]\right]}{t} = \frac{\text{CMI}_{\mathcal{D}}\left(\mathcal{A}\right) + \mathbb{E}_{\tilde{Z}}\left[\log\mathbb{E}_{W,S'}\left[e^{tf_{\tilde{Z}}\left(W,S'\right)}\right]\right]}{t}$$

Optimizing over $t > 0$, we get the bound

$$\text{gen-err}_{\mathcal{D}}\left(\mathcal{A}\right) \leq \inf_{t>0}\frac{\text{CMI}_{\mathcal{D}}\left(\mathcal{A}\right) + \mathbb{E}_{\tilde{Z}}\left[\log\mathbb{E}_{W,S'}\left[e^{tf_{\tilde{Z}}\left(W,S'\right)}\right]\right]}{t}$$

That is, we get a bound of the generalization error in terms of the CMI and the moment generating function, which we can bound using the independence assumption on S', Hoeffding's lemma and the bound on the range of l. Over all we get

$$\text{gen-err}_{\mathcal{D}}\left(\mathcal{A}\right) \leq \inf_{t>0}\frac{\text{CMI}_{\mathcal{D}}\left(\mathcal{A}\right) + \mathbb{E}_{\tilde{Z}}\left[\log\mathbb{E}_{W,S'}\left[e^{tf_{\tilde{Z}}\left(W,S'\right)}\right]\right]}{t} \leq \inf_{t>0}\frac{\text{CMI}_{\mathcal{D}}\left(\mathcal{A}\right) + \frac{t^2}{2n}}{t} = \sqrt{\frac{2}{n}\text{CMI}_{\mathcal{D}}\left(\mathcal{A}\right)}$$

as required.

Bounding the CMI

The next part of the article demonstrates how known bounds on generalization from different methods imply a bound on the CMI, and thus fit into the CMI framework. Combined with the last section, the analysis via CMI implies essentially the same bounds on generalization as the direct analysis of the other methods. This is true for bounds from compression schemes, approximate differential privacy, distributional stability and more.

Additionally, we see a connection to the uniform convergence framework, which we mention above. Specifically, for a 0-1 loss over space $\mathcal{X} \times \{0, 1\}$, bounded VC dimension is a sufficient condition for generalization. However, this is a property of the output space, while the CMI depends on the algorithm, therefore the two methods are incompatible. Nonetheless, there is a connection between them: For any hypothesis class of bounded VC dimension, there always **exists** an ERM algorithm with bounded CMI. And yet, this bound is not tight: for VCdim $= d$, we get $\text{CMI}_{\mathcal{D}}\left(\mathcal{A}\right) \leq O\left(d\log n\right)$. For some specific cases the gap is can be eliminated, and this is believed to be possible in general, given more information on the structure of the problem.

Another setback is the fact that the CMI bounds are all bounds on an expectation. Potentially, they can be converted into probability bounds via Markov's inequality, but this still won't yield "high probability" bounds (the failure probability decays polynomially with the desired error bound, rather than exponentially). However, there is a suggestion for a possible direction using an extension of the CMI framework.

Other extensions are also suggested (and some implemented) to improve the existing CMI bound and show relationship with other methods (e.g. PAC-bayes and uniform stability).

Summary

Overall, this article outlines a new framework for reasoning about generalization which unifies existing methods to provide a variety of generalization bounds. It doesn't depend on the loss function but on the distribution and the algorithm in a manner that, in my opinion, is very intuitive. It is shown to retrieve known bounds from other conditions through CMI and there are many possibilities for extensions (some explored in the article or other recent works).