Complexity Measures

Vladimir Feinberg

July 9, 2017

Complexity measures evaluate the expressiveness of a hypothesis class; they are useful to the extent with which they relate sample and generalization error.

1 Setup

We suppose that our data comes in the form of ordered pairs from $\mathcal{X} \times \mathcal{Y}$. Samples follow a particular distribution $(x, y) \sim D$. A hypothesis class $\mathcal{H}$ is set of functions $\mathcal{X} \to \mathcal{Y}$.

A common approach to supervised learning is ERM, where m iid samples from D, S, are used to find the $h \in \mathcal{H}$ minimizing a specified loss $\ell : \mathcal{Y}^2 \to \mathbb{R}$ over this set. Complexity measures then let us quantify exactly how much loss we can expect when sampling from D again.

We seek to quantify the generalization gap with the help of our notions of complexity. For a fixed $h \in \mathcal{H}$:

$$\varepsilon = \mathbb{E}\left[\ell\left(h(x), y\right)\right)|(x, y) \sim D] - \mathbb{E}\left[\ell\left(h(x), y\right)\right)|(x, y) \sim \text{Uniform}(S)]$$

Analysis of Rademacher complexity is agnostic to h, ℓ; the hypothesis class might as well consist of functions $g : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ yielding their composition. VC dimension analysis, however, requires $\mathcal{Y} = \{0, 1\}$ and $\ell(a, b) = \mathbb{1}\{a = b\}$. VC dimension is still useful for regression problems, by thresholding hypotheses $h \mapsto \mathbb{1}h > \beta$ for fixed β.[1]

Thus, it is useful to find bounds on ε, the difference between the generalization loss $\mathbb{E}\left[\ell\left(h(x), y\right)\right)]$, where $(x, y) \sim D$, and sample loss, where the loss is the expectation before taken for (x, y) is uniform over S.

Let the gap between the generalization and sample error be ε.

2 Complexity Measures

The empirical Rademacher complexity $\hat{R}_S$ assumes a fixed sample S from D^m. It relates complexity of a function class $\mathcal{G}$ containing vectorized functions $g \in \mathcal{G}$ which take elements $z_i = (x_i, y_i)$ in S and return costs through the correlation of $\mathcal{G}$ with noise. Let $\boldsymbol{\sigma} \sim \text{Uniform}\left(\pm 1\right)^m$. Rademacher complexity is then the average empirical one.

$$\hat{R}_S(\mathcal{G}) = \mathbb{E}_{\boldsymbol{\sigma}} \sup_g \frac{1}{m} \sum_{i=1}^m g(z)\sigma_i, \quad R_m(\mathcal{G}) = \mathbb{E}_S\, \hat{R}_S(\mathcal{G})$$

VC dimension accomplishes a similar task for binary classification by rating the complexity of a hypothesis class $\mathcal{H}$. Let hypotheses $\mathcal{H} \ni h : \mathcal{X} \to \mathcal{Y} = \{\pm 1\}$ be applied elementwise over a vector of inputs $\mathbf{x}$. First we define the growth function $\Pi_{\mathcal{H}} : \mathbb{N} \to \mathbb{N}$, which defines the maximum number of distinctions a hypothesis class can make over all sets of points in the input space:

$$\Pi_{\mathcal{H}}(m) = \max_{\mathbf{x} \in \mathcal{X}^m} |\{h(\mathbf{x}) \,|\, h \in \mathcal{H}\}|$$

Then the VC dimension of $\mathcal{H}$ is then $\max\{m \in \mathbb{N} \,|\, \Pi_{\mathcal{H}}(m) = 2^m\}$.

[1] https://stats.stackexchange.com/questions/140430

3 Overview of Results

Proofs can be found in a cogent write-up by Prof. Beckage from the University of Kansas.

3.1 VC Generalization Bounds

Upper bound. If d is the VC-dimension of $\mathcal{H}$, then for any D wp $1-\delta$:

$$\varepsilon \leq \tilde{O}\left(\sqrt{\frac{d-\log\delta}{m}}\right)$$

The above inequality is random since it depends on S, the D^m-valued rv. TODO. find source removing tilde?

Agnostic lower bound. We may find a D such that with a fixed nonzero probability (a non-negligible set of candidate samples S), the following holds:

$$\varepsilon \geq \Omega\left(\sqrt{\frac{d}{m}}\right)$$

The above implies that in the common case of agnostic hypothesis learning, where we do not know distribution D, VC-dimension is, *up to logarithmic factors, asymptotically efficient* in quantifying the generalization gap.

Realizability. Suppose D is realizable wrt $\mathcal{H}$, so that there exists an $f \in \mathcal{H}$ such that for almost any (x, y) sampled from D, $f(x) = y$. Then all statements above hold but with $\sqrt{\varepsilon}$ instead of ε.

3.2 Growth Function Bounds

Sauer's Lemma implies that VC dimension d bounds the growth function: in a graph of the logarithm of the growth function vs m, growth is linear since $\Pi_{\mathcal{H}}(n) = n$ for $n \leq d$. Then for $n > d$, growth is at most logarithmic, i.e., $\log\Pi_{\mathcal{H}} = O(\log m)$. With Massart's Lemma we have wp $1-\delta$:

$$\varepsilon \leq O\left(\sqrt{\frac{\log\Pi_{\mathcal{H}}(m)-\log\delta}{m}}\right)$$

Since the above would be large if $\log\Pi_{\mathcal{H}}(m) \simeq m$, it is clear why Sauer's Lemma enables the essential relationship between learnability and complexity.

3.3 Rademacher bounds

With R_m either the empirical or expected Rademacher complexity over the sample for a given h, ℓ we have again wp $1-\delta$:

$$\varepsilon \leq 2R_m + O\left(\frac{\log {}^1\!/\!_\delta}{m}\right)$$

R_m may be NP-hard to compute, depending on $\mathcal{H}$. This tells us Rademacher complexity could only be a useful improvement over VC-bounds, asymptotically, if we have an efficient approximation for the empirical Rademacher complexity or some knowledge of D as required to compote the true Rademacher complexity.

4 Hardness of Learning

Rademacher and Gaussian Complexities: Risk Bounds and Structural Results by Bartlett and Mendelson.