

Complexity Measures

Vladimir Feinberg

May 30, 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 $\mathcal{X} \rightarrow \mathcal{Y}$.

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 $\ell : \mathcal{Y}^2 \rightarrow \mathbb{R}$ 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 $h \in \mathcal{H}$:

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

Analysis of Rademacher complexity is agnostic to h, ℓ ; the hypothesis class might as well consist of functions $g : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}$ yielding their composition. VC dimension analysis, however, requires $\mathcal{Y} = \{0, 1\}$ and $\ell(a, b) = \mathbb{1}\{a \neq b\}$. 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 $\mathbb{E} [\ell(h(x), y)]$, 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

Rademacher complexity. TODO. relates complexity through noise correlation. equivalently, gives a consistent view of “error” through a “uniform” noise model.

VC-dimension. TODO.

3 Overview of Results

Proofs can be found in a cogent write-up by Prof. Beckage from the University of Kansas, copied into this repository as `complexity-measures-beckage.pdf`.²

¹<https://stats.stackexchange.com/questions/140430>

²http://ittc.ku.edu/~beckage/ml800/VC_dim.pdf

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

3.2 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 compute the true Rademacher complexity.

TODO. NP-hardness? More computational learning theory.

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

TODO. Growth function as rademacher/VC bridge. Graph of (log Growth func) vs m linear then logarithmic past d . Note no nuance in number of shatterable sets, just that one exists [missing direction for analysis?]