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.

Notation. TODO move to common place in repo. Big-O notation (and Θ, Ω) is used in the standard sense for single variables: $f = O(g)$ if there exists a c such that $f(x) \le cg(x)$ for all x sufficiently large. In multiple variables, big-O and similarly Θ, Ω require $f(\mathbf{x}) \le cg(\mathbf{x})$ for all $\|x\|_\infty$ sufficiently large. To capture the essence of asymptotics, a tilde will crudely capture asymptotic behavior up to logarithmic factors: $f = \tilde{O}(g)$ if $f = O(g \log^k g)$ for some $k \in \mathbb{N}$.

Abbreviations. TODO move to common place.

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 empirical risk minimization, 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 quanify exactly how much loss we can expect when sampling from D again.

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

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

The growth function. TODO.

VC-dimension. TODO.

3 Overview of Results

Proofs can be found in a cogent write-up by Prof. Beckage from the University of Kansas,[1] copied into this repository.[2] Recall our generalization gap definition. For a fixed $h \in \mathcal{H}$:

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

[1] `http://ittc.ku.edu/~beckage/ml800/VC_dim.pdf`

[2] `https://github.com/vlad17/shallow-ml-notes/raw/7c3db7b7c924fbd2ae2891d4ecfef84e5647dcc8/computational-learning/complexity-measures-beckage.pdf`

3.1 VC Generalization Bounds

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

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

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 nonneglible set of candidate samples S), the following holds:

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

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 $1-\delta$:

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

R_m may be NP-hard to compute, depending on $\mathcal{H}$. This tells us Rademacher complexity could only be a useful imporvement 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.

TODO. NP-hardness? More computational learning theory.