Bounding generalization error by a sample-conditioned count of classifiers

Yoram Gat

June 18, 2015

Abstract

This talk will present a result useful for bounding the generalization error of certain classification algorithms. Like VC theory, the bound relies on symmetrization. However, the bound applies to algorithms with high VC-dimension ranges.

One application of the bound is for showing the good generalization behavior of the “support point machine” - a variant of the support vector machine which does not require an inner-product space structure.

- $\mathscr{X}$ - a **feature space**, $\mathscr{Y}$ - a **label space**. Their product is the **labeled feature space** $\mathscr{Z} = \mathscr{X} \times \mathscr{Y}$.
- A **classifier** is a map $f : \mathscr{X} \to \mathscr{Y}$. The **space of classifiers** is $\mathscr{F}$.
- A **classification algorithm** or **learning algorithm** is a map from **a training set** to a classifier $L : \mathscr{Z}^n \to \mathscr{F}$.

- An **error function** corresponding to a classifier f is a map $e_f : \mathscr{Z} \to \{0, 1\}$:

$$e_f(z) = e_f(x, y) = 1(f(x) \neq y).$$

  The **space of error functions** is $\mathscr{E}$.

- Shorthand for the average error of a sample:

$$e_f^n(\mathbf{Z}) = e_f^n(z_1, \ldots, z_n) = \frac{1}{n} \sum_1^n e_f(z_i).$$

The **generalization error of a classification algorithm**:

$$g(L) = \mathbb{E}_{\mathbf{Z}} \mathbb{E}_{Z'} e_{L(\mathbf{Z})}(Z') - e^n_{L(\mathbf{Z})}(\mathbf{Z}).$$

Symmetrization

$$
\begin{aligned}
g(L) &= \mathbb{E}_{\mathbf{Z}}\mathbb{E}_{Z'}\left[e_{L(\mathbf{Z})}(Z') - e^{n}_{L(\mathbf{Z})}(\mathbf{Z})\right] \\
&= \mathbb{E}_{\mathbf{Z}}\mathbb{E}_{\mathbf{Z}'}\left[e^{n'}_{L(\mathbf{Z})}(\mathbf{Z}') - e^{n}_{L(\mathbf{Z})}(\mathbf{Z})\right] \\
&= \mathbb{E}_{\mathbf{Z},\mathbf{Z}'}\mathbb{E}_{\text{Partition}}\left[e^{n'}_{L(\bar{\mathbf{Z}})}(\bar{\mathbf{Z}}') - e^{n}_{L(\bar{\mathbf{Z}})}(\bar{\mathbf{Z}})\right].
\end{aligned}
$$

Let the pair $\mathbf{z}, \mathbf{z}'$ be the **train-test set** and define the **partition generalization error**:

$$
g(L, \mathbf{z} \cup \mathbf{z}') = \mathbb{E}_{\text{Partition}}\left[e^{n'}_{L(\bar{\mathbf{z}})}(\bar{\mathbf{z}}') - e^{n}_{L(\bar{\mathbf{z}})}(\bar{\mathbf{z}})\right],
$$

giving

$$
g(L) = \mathbb{E}_{\mathbf{Z},\mathbf{Z}'}\, g(L, \mathbf{Z} \cup \mathbf{Z}').
$$

VC theory

- Given $\mathbf{z}, \mathbf{z}'$, define $\mathscr{E} \downarrow \mathbf{z} \cup \mathbf{z}'$ - the **sample-domain space of error functions**. This is $\mathscr{E}$ where the domain of the error functions has been limited to $\mathbf{z} \cup \mathbf{z}'$.
- The partition generalization error can be bounded in terms of the size of the sample-domain space of error functions:

$$g(L, \mathbf{z} \cup \mathbf{z}') \leq \left|\mathscr{E} \downarrow \mathbf{z} \cup \mathbf{z}'\right| \epsilon(n, n').$$

- If $\mathscr{F}$ has low VC dimension then the size of $\mathscr{E} \downarrow \mathbf{z} \cup \mathbf{z}'$ is small and therefore $g(L, \mathbf{z} \cup \mathbf{z}')$ is small for any L with range $\mathscr{F}$.
- This works well if $\mathscr{F}$ is, say, the space of planes, or balls, or boxes in d dimensions.

A low VC dimension for $\mathscr{F}$ is also a necessary condition for low generalization error in the sense that if $\mathscr{F}$ has a high VC dimension then there exists a learning algorithm L (namely ERM) which will have high generalization error.

Learning in high VC dimensions

- By abandoning ERM, good generalizability can be guaranteed in high VC dimensions.
- E.g., SVM: in a high or infinite dimensional linear space, only select separating planes that have a large margin.

Define the **sample-conditioned space of classifiers**:

$$\mathscr{F}_L(\mathbf{z} \cup \mathbf{z}') = \left\{ L(\bar{\mathbf{z}}) : \bar{\mathbf{z}} \subset \mathbf{z} \cup \mathbf{z}' \right\}.$$

If $\mathscr{F}_L(\mathbf{z} \cup \mathbf{z}')$ is small then $g(L, \mathbf{z} \cup \mathbf{z}')$ is small:

$$g(L, \mathbf{z} \cup \mathbf{z}') \leq \left| \mathscr{F}_L(\mathbf{z} \cup \mathbf{z}') \right| \epsilon(n, n').$$

A learning algorithm L is d-**determined** if there exists some function $L_0 : \mathscr{Z}^d \to \mathscr{F}$ such that

$$L(\mathbf{z}) = L_0(z_{i_1(\mathbf{z})}, \ldots, z_{i_d(\mathbf{z})}).$$

When L is d-determined,

$$\left|\mathscr{F}_L(\mathbf{z} \cup \mathbf{z}')\right| = \binom{n+n'}{d} \ll \binom{n+n'}{n}.$$

Applications

Edited nearest neighbors

Of n training points, pick d (in any way the learner chooses) and use those for [k-]nearest neighbor classification.

Let $s(x_1, x_2)$ be an arbitrary similarity function mapping pairs in $\mathscr{X}$ to $[-1, 1]$.

Let

$$L(\mathbf{z})(x) = \operatorname{sgn} \sum_{i=1}^{n} y_i \alpha_i s(x_i, x),$$

where $\sum_i \alpha_i = 1, \alpha_i > 0$.

Define a d-determined L' as follows:

$$L'(\mathbf{z})(x) = \operatorname{sgn} \frac{1}{d} \sum_{j=1}^{d} y_{K_j} s(x_{K_j}, x),$$

where $K_1, \ldots, K_d$ are IID sampled from $\{1, \ldots, n\}$ with probabilities $\alpha_1, \ldots, \alpha_n$.

When L has a large margin, it is well approximated by L'. The low generalization error of L' then implies low generalization error of L.

When $\mathscr{E}$ is finite,

$$
\begin{aligned}
g(L, \mathbf{z} \cup \mathbf{z}') &\leq \mathbb{E}_{\text{Partition}} \max_{e \in \mathscr{E}} e^{n'}(\bar{\mathbf{z}}') - e^{n}(\bar{\mathbf{z}}) \\
&\leq \sum_{e \in \mathscr{E}} \mathbb{E} \left| e^{n'}(\bar{\mathbf{z}}') - e^{n}(\bar{\mathbf{z}}) \right| \\
&\leq |\mathscr{E}| \, \epsilon(n, n').
\end{aligned}
$$

Combining VC theory with sample-set conditioning

Sample-set conditioning can be combined with VC theory.

Define the **sample-conditioned space of error functions**:

$$\mathscr{E}_L(\mathbf{z}, \mathbf{z}') = \{e_f : f \in \mathscr{F}_L(\mathbf{z} \cup \mathbf{z}')\}.$$

Then the partition generalization error can be bounded in terms of the **sample-domain, sample-conditioned space of error functions**:

If $\mathscr{E}_L(\mathbf{z}, \mathbf{z}') \downarrow \mathbf{z}, \mathbf{z}'$ is small then $g(L, \mathbf{z} \cup \mathbf{z}')$ is small.