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.
Definitions

1. Classifiers and classifications algorithms

- $\mathcal{X}$ - a feature space, $\mathcal{Y}$ - a label space. Their product is the labeled feature space $\mathcal{L} = \mathcal{X} \times \mathcal{Y}$.

- A classifier is a map $f : \mathcal{X} \rightarrow \mathcal{Y}$. The space of classifiers is $\mathcal{F}$.

- A classification algorithm or learning algorithm is a map from a training set to a classifier $L : \mathcal{L}^n \rightarrow \mathcal{F}$. 
An **error function** corresponding to a classifier $f$ is a map $e_f : \mathcal{X} \to \{0, 1\}$:

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

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

- Shorthand for the average error of a sample:

$$e_f^n(Z) = e_f^n(z_1, \ldots, z_n) = \frac{1}{n} \sum_{1}^{n} e_f(z_i).$$
The **generalization error of a classification algorithm**:

\[ g(L) = \mathbb{E}_Z \mathbb{E}_{Z'} e_L(z)(Z') - e^n_{L(Z)}(Z). \]
Symmetrization

\begin{align*}
g(L) &= \mathbb{E}_Z \mathbb{E}_{Z'} \left[ e_L(z)(Z') - e_L^n(z)(Z) \right] \\
&= \mathbb{E}_Z \mathbb{E}_{Z'} \left[ e_L^n(z)(Z') - e_L^n(z)(Z) \right] \\
&= \mathbb{E}_{z,z'} \mathbb{E}_{\text{Partition}} \left[ e_L^n(z)(\bar{Z}) - e_L^n(z)(\bar{Z}) \right].
\end{align*}

Let the pair \( z, z' \) be the \textbf{train-test set} and define the \textbf{partition generalization error}:

\begin{align*}
g(L, z \cup z') &= \mathbb{E}_{\text{Partition}} \left[ e_L^n(z)(\bar{z'}) - e_L^n(z)(\bar{z}) \right],
\end{align*}

giving

\begin{align*}
g(L) &= \mathbb{E}_{z,z'} g(L, z \cup z').
\end{align*}
Given $z, z'$, define $\mathcal{E} \downarrow z \cup z'$ - the **sample-domain space of error functions**. This is $\mathcal{E}$ where the domain of the error functions has been limited to $z \cup z'$.

The partition generalization error can be bounded in terms of the size of the sample-domain space of error functions:

$$g(L, z \cup z') \leq |\mathcal{E} \downarrow z \cup z'| \epsilon(n, n').$$
If $\mathcal{F}$ has low VC dimension then the size of $\mathcal{E} \downarrow z \cup z'$ is small and therefore $g(L, z \cup z')$ is small for any $L$ with range $\mathcal{F}$.

This works well if $\mathcal{F}$ is, say, the space of planes, or balls, or boxes in $d$ dimensions.
A low VC dimension for \( \mathcal{F} \) is also a necessary condition for low generalization error in the sense that if \( \mathcal{F} \) has a high VC dimension then there exists a learning algorithm \( L \) (namely ERM) which will have high generalization error.
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.
Bounding the partition generalization error in high VC dimensions

The sample-conditioned space of classifiers

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

\[
\mathcal{F}_L(z \cup z') = \{ L(\bar{z}) : \bar{z} \subset z \cup z' \}.
\]

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

\[
g(L, z \cup z') \leq |\mathcal{F}_L(z \cup z')| \epsilon(n, n').
\]
A learning algorithm $L$ is \textbf{$d$-determined} if there exists some function $L_0 : \mathcal{X}^d \rightarrow \mathcal{F}$ such that

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

When $L$ is $d$-determined,

$$| \mathcal{F}_L(z \cup z') | = \binom{n + n'}{d} \ll \binom{n + n'}{n}.$$
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 \( \mathcal{X} \) to \([-1, 1]\).

Let

\[
L(z)(x) = \text{sgn} \sum_{i=1}^{n} y_i \alpha_i s(x_i, x),
\]

where \( \sum_i \alpha_i = 1, \alpha_i > 0. \)
Define a \( d \)-determined \( L' \) as follows:

\[
L'(z)(x) = \text{sgn} \left( \frac{1}{d} \sum_{j=1}^{d} y_{K_j}s(x_{K_j}, x) \right),
\]

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 $\mathcal{E}$ is finite,

\[
g(L, z \cup z') \leq \mathbb{E}_{\text{Partition}} \max_{e \in \mathcal{E}} e^n(\bar{z}') - e^n(\bar{z})
\]
\[
\leq \sum_{e \in \mathcal{E}} \mathbb{E} \left| e^n(\bar{z}') - e^n(\bar{z}) \right|
\]
\[
\leq |\mathcal{E}| \epsilon(n, n').
\]
Combining VC theory with sample-set conditioning

Sample-set conditioning can be combined with VC theory.

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

\[ E_L(z, z') = \{ e_f : f \in F_L(z \cup z') \} \]

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

If \( E_L(z, z') \downarrow z, z' \) is small then \( g(L, z \cup z') \) is small.