Insights from Statistical Learning Theory

One issue in learning is whether a simple or complex function should be used for classification. We don’t want to underfit (too simple) or overfit (too complex).

The complexity of a type of function can be characterized by its VC dimension ($VC = Vapnik$ and Chervonenkis).

The VC dimension roughly corresponds to the number of weights/basis functions.

For good generalization, the number of examples should be many times larger than the VC dimension.

For SVMs, a smaller distance between the $-1$ and $+1$ boundaries implies a higher VC dimension. Also, more support vectors implies a higher VC dimension.

Figures on the following pages are from B. Scholkopf and A. Smola, *Learning with Kernels*, MIT Press, 2002.
This Function is too Simple
A simple function might miss too many examples.

This Function is too Complex
A complex function might rely too much on outliers.
This Function is Just Right

A function with the right amount of complexity only makes mistakes on outliers.

Background

We assume the training examples are generated independently from a probability distribution $P(x, y)$.

The goal is to find a function $f_a$ that will correctly classify $x$ generated from $P(x, y)$.

If there are no restrictions on $f_a$, then two different functions can agree on training examples, but disagree otherwise.
By itself, minimizing training error does not imply small test error.

**VC Dimension**

One way to restrict $f_a$ is to limit the VC dimension of what can be learned. VC dimension $d$ implies some set of $d$ examples and $2^d$ functions that split the examples all possible ways.
VC Dimension Bound

With probability $1 - \delta$, the test error will exceed training error by at most:

$$\sqrt{\frac{1}{m} \left( d \left( 1 + \ln \frac{2m}{d} \right) + \ln \frac{4}{\delta} \right)}$$

where $m$ is the number of training examples and $d$ is the VC dimension.

The key term is $\sqrt{d(\ln m)/m}$. To make this small, $m$ must be many times larger than $d$.

SVM Bounds

The VC dimension of a SVM is at most

$$1 + 4R^2/M^2$$

$R$ is the radius of the examples. $M$ is the margin, the distance between the $+1$ and $-1$ boundaries of the SVM.

Also the expected test error of an SVM is at most:

$$\frac{\text{number of support vectors}}{m}$$

where $m$ is the number of training examples.