7.1.5 Computational learning theory

Historically, SVM have largely been motivated and analyzed using a theoretical framework known as computational learning theory(=statistical learning theory).

This has its origins with PAC(probably approximately correct, a learning framework) creator.

Goal of PAC : how large a data set needs to be in order to give good generalization.

It also gives bounds for the computational cost of learning.

Suppose that a data set D of size N is drawn from some joint distribution $p(x, t)$,

x : input variable , t : class label, and that

Attention to 'noise free' situations in which the class labels are determined by deterministic function $t = g(x)$.

In PAC learning, we say that a function $f(x; D)$,

drawn from a space F of such functions on the basis of the training set D, has good generalization

if its expected error rate is below some pre-specified threshold ϵ so that $E_{x,t}[I(f(x; D) \neq t)] < \epsilon$

$I(\cdot)$: indicator function, and the expectation is with respect to the distribution $p(x, t)$.

The quantity on the left-hand side is a random variable, because it depends on the training set D,

and the PAC framework requires that $E_{x,t}$ holds, with probability greater than $1 - \delta$,

for a data set D drawn randomly from $p(x, t)$.

Here δ is another pre-specified parameter, and

the terminology 'probably approximately correct' comes from the requirement that with high probability(greater than $1 - \delta$), the error rate be small (less than ϵ).

For a given choice of model space F, given parameters ϵ and δ,

PAC learning aims to provide bounds on the minimum size N of data set needed to meet this criterion.

A key quantity in PAC learning is the VC dimension, which provides a measure of the complexity of a space of functions, and which allows the PAC framework to be extended to spaces containing an infinite number of function.

The bounds derived within the PAC framework are often described as worst case, because they apply to any choice for the distribution $p(x, t)$, so long as both the training and the test examples are drawn (independently) from the same distribution, and for any choice for the function $f(x)$ so long as it belongs to F

in real-world applications of machine learning, we deal with distributions that have significant regularity, for example in which large regions of input space carry the same class label.

As a consequence of the lack of any assumptions about the form of the distribution, the PAC bounds are very conservative, i.e., they strongly over-estimate the data size required to achieve a given generalization performance.

For this reason, PAC bounds have found few, if any, practical applications.

One attempt to improve the tightness of the PAC bounds is the PAC-Bayesian framework, which considers a distribution over the space F of functions, somewhat analogous to the prior in a Bayesian treatment.

This still considers any possible choice for $p(x, t)$, and so although the bounds are tighter, they are still very conservative.