A Crisis of Faith

Uniform Convergence, the downside

Given: IID samples + classifier space

Find: Number of samples such that the classifier with minimum training error has a true error within ϵ with probability $1-\delta$?

1. Data not IID in practice.

2. Finding classifier with minimum error is computationally infeasible on analyzed spaces.

Why care about classifier with minimum training error? Why not minimal true error?

Uniform Convergence, the other downside

Typical result: If your classifier space has a small {VC Dimension, VC Entropy, Rademacher Complexity} and a reasonable number of samples, then minimum error classifier is good.

1. VC-dimension is difficult to compute for used spaces.

2. VC-Entropy is almost impossible to compute for used spaces.

3. Rademacher Complexity is very difficult to compute for used spaces.

Even when these bounds are computable, they are often unacceptably loose.

PAC learning, the downside

Given: IID samples, y=f(x) for f in some set C={monomials, monotone, xor, dnf, etc...}

Find: Learning algorithm, number of samples s.t. with probability $1-\delta$ classifier chosen with error rate ϵ.

1. Data not IID in practice.
2. S is unknown in practice. (Applied PAC learning = run them all?)
3. Data noisy in practice. (Agnostic PAC results are just about all negative...)

Bit-bounds, the downside

Given: IID samples, “prior” over classifier space

Find: Bound on true error rate.

1. Data not IID in practice.

2. Not, quite, tight enough.

3. Difficult to make work in practice.

Online Learning, the downside

Given: set of classifiers (or "prior" on classifiers)

Find: meta algorithm that does almost as well as best classifier.

1. Typically computationally intractable (?)

If this was succesful, it would be used. It isn't. Why not?

A Mathematicians Viewpoint

I don't care about learning practice, I care about property X of my analysis.

X = "Combinatorics", "Rademacher Complexity", etc...

- This is a disavowal of interest by learning practitioners.

- Assumption: We want the interest of learning practitioners.

Ultimate goal: useful learning algorithms.

The "please employ my friends" viewpoint

1. "First you analyze what happens asymptotically, then you make crude bounds, then you tighten the crude bounds."

... and then you realize the model does not address the problem, the bounds can not be made tight enough, etc...

1. "First you analyze under unrealistic assumptions, then you remove one, and another, and another."

... and then you realize that the first analysis doesn't help later ones.

The “knobs” argument

“A loose analysis is still useful because we can tune the constants as parameters with cross validation.”

1. But nobody does this.

2. Simple ’bias-variance’ tradeoff arguments tell you the same thing. Why do a long and ugly analysis when there are simple ways to get this intuition?

A Good Theory is Feared and Respected by Practitioners

<add graph>

Key Questions

1. How do we affect actions of practitioners?

2. Where are the algorithms?

3. How is one model preferred over another?

4. Are there intuitions and insights driving actions derivable from sample complexity bounds which are NOT derivable from other sources?

Schedule