Can Neural Network and Statistical Learning Theory be Formulated in terms of Continuous Complexity Theory?

1. Neural Networks and Statistical Learning Theory

Many areas of mathematics, statistics, and computer science deal with extrapolation of functions from partial information or examples.

E.G., data mining, prediction

Function approximation problem: How *to* best estimate the function f from partial information (examples)

$$y = Nf = (f(x_1) + \epsilon_1, \ldots, f(x_k) + \epsilon_k)$$

i.e., values of f at a finite number of points $x_1, \ldots, x_k$, with possible error ϵ_k?

More generally:

Given normed linear space F and an unknown $f \in F,$ how to best estimate f (in the norm of F) given information

$$Nf = (L_1 f, \ldots, L_k f),$$

where L_i are (linear or nonlinear) functionals (henceforth implicitly assume possible error terms ϵ_k in the components of N).

IBC is a complete theory of information and algorithmic complexity developed for study of function approximation problem and its generalizations.

Other work in function approximation is closely related to continuous complexity theory. We wish to classify them within the framework of complexity analysis established within this theory.

This other work includes:

- Statistical learning theory [Vapnik, Poggio].

- Learning in neural network theory [Rumelhart, Hinton, Williams],

- Computational learning theory [Kearns, Vazirani],

- Regularization theory [Poggio, Girosi]

- Regression theory in statistics

- Maximum entropy method [Jaynes],

- Theory of V-C dimension and approximation [Vapnik, Chervenenkis, Poggio, Girosi, etc.],

- Adaptive resonance theory [Carpenter, Grossberg]

- Approximation theory

- Decision tree methodologies (AI)

Goal 1: identify ϵ-complexities of algorithms in these areas (so far, there are almost no results)

Goal 2: classify such methods in context of continuous complexity analyses

Goal 3: precisely define optimality within such classes of approaches, and identify optimal algorithms from among differing approaches above.

Resulting Goal: Form a normative index of such methods according to their (now comparable) optimality properties.

Some currently used methods seem outside the domain of the information-based

continuous complexity model; we wish to show that this model includes these existing approaches.

Desired outcome: Use continuous complexity formulation to move closer to a more inclusive theory of continuous optimal algorithms, into which most current approaches for function extrapolation would fit.

Important philosophical point:

For comparison of various methods of data extrapolation from partial information, need universal separation of information into

- *a priori* information = prior information about f

- *a posteriori* information = data $Nf = (f(x_1), \ldots, f(x_n))$

Examples:

1. *Interpolatory approach (worst case approach)*

A priori information:

$$f \in F_1 = \text{balanced convex function set}$$

A posteriori information: $Nf = y$, *i.e.*

$$f \in N^{-1}y$$

Optimal algorithms approximate center of set $F_1 \cap N^{-1}y$ through algorithm ϕ :

$$\phi(Nf) \approx f.$$

2. *Average case approach*

A priori information: Unknown function f has an a priori probability distribution μ on a Banach space F.

A posteriori information:

$$Nf = (L_1 f, \ldots, L_n f)$$

where L_i are linear functionals.

Choose estimate of f to be the average of μ conditioned on $Nf = y =$ given data.

Optimal algorithm for Gaussian measure: spline algorithm

$$\phi(Nf) = \sum_j L_j(f) C_\mu L_j$$

where C_μ is the covariance operator, defined by

$$L_1(C_\mu L_2) = \int_F L_1(f) L_2(f) \mu(df).$$

3. Maximum likelihood approaches

Choose estimate of f which is consistent with Nf and whose probability is the largest, for the measure μ restricted to $\{f : Nf = y\}$.

4. *Regularization approach*

Ex: Maximum likelihood methods of Bayesian statistics:

Given: a priori probability distribution μ on F; assume f chosen according to μ.

Algorithm: maximize density of μ subject to the *a posteriori* information $Nf = y$.

More generally, this involves minimization of a weighted combination

$$H_\lambda(f) = \|Nf - y\|^2 + \lambda\Lambda(f).$$

In fact, we feel that the union of the interpolatory and regularization approaches constitutes a very comprehensive set of algorithms $\phi(Nf)$ which explicitly separate the two types of information.

5. Vapnik-Chervenenkis (VC) approach

Given nested family $\{V_\lambda\}$ of candidate *a priori* spaces increasing in size (and complexity) with λ.

Approach:
λ small
$\Rightarrow$ candidate set $N^{-1}y \cap V_\lambda$ small
$\Rightarrow$ selection of approximation

$$\tilde{f} \in N^{-1}y \cap V_\lambda$$

is from reasonable sized set.
Approach: choose *smallest* λ such that $N^{-1}y \cap V_\lambda$ is non-empty.

6. *Neural Network Algorithms*

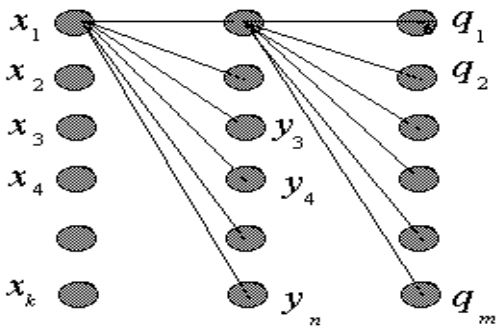
- Use examples Nf of an unknown i-o function f.

- Apply algorithm ϕ to get approximation

$$\tilde{f} = \phi(Nf)$$

$\tilde{f}$ is the function computed by the network after training with examples Nf; chosen from parameterized class P of network-computable functions

$$f(\mathbf{x}) = \mathbf{q},$$



with

$$\mathbf{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_k \end{bmatrix}; \ \mathbf{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}; \ \mathbf{q} = \begin{bmatrix} q_1 \\ \vdots \\ q_m \end{bmatrix}.$$

Function $f(x)$ defined by:

$$y_i(\mathbf{x}) = \sum_j w_{ij} \, G_j(\mathbf{x})$$

$$q_i(\mathbf{y}) = \sum_j v_{ij} y_j$$

$=$ linear operation.

Class $P = P_n$ (measured by size n of middle layer) = set of functions computable by network with middle layer of size n;

seek closest match to f in $\tilde{f} \in P_n$.

Same arguments as for the V-C method $\Rightarrow$:

We want P_n sufficiently diverse to approximate f, but not so diverse that there are too many solutions (and solution problem is ill-posed). Thus make n as small as possible, so long as $N^{-1}y \cap P_n$ is non-empty.

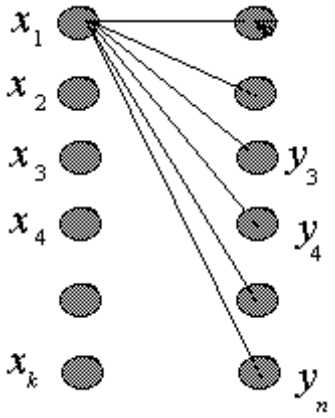
Effectively, the a priori class P_n is similar to a class F_1 consisting of a ball in a Sobolev space, i.e. an F optimizing for smoothness.

Indeed, if P_n is specialized to compute radial basis functions (RBF's), we know optimal approximations from P explicitly minimize for Sobolev norm.

$\exists$ many variations of approximation of i-o functions f with a priori assumptions (e.g., $f \in P_n$) essentially assuming smoothness of f.

7. Adaptive resonance theory (ART):

Algorithm: dynamic neural network weight allocation procedure for classification of input vectors **x.**



For different classes C_i of input vectors **x**, different output vectors y_i respond.

This network is similar (in its simplest from) to first two layers of feedforward radial basis function network

Difference: second competitive processing stage where the hidden neuron $y_i = G_i(x)$ with highest activation suppresses all other neurons $y_j \neq y_i$.

Effectively, ART network computes the function

$$f^*(x) = (0, \ldots, 0, G_j(x), 0, \ldots, 0), \qquad (1)$$

where the right hand side represents the choice of the $G_j(x) = y_j$ with the maximum value.

Note: $G_i(x) =$ maximum over all j indicates that input was in class C_i.

A posteriori information: the data vector **x**
a priori information: fact that the i-o function is approximable in the form (1).

Smoothness of $G_i(x)$ is effectively an a priori assumption on smoothness of the separators of classes C_i above

Claim: It is possible to compare the various approaches to function extrapolation using

well-defined comparisons, based on identification in each case of *a priori* and *a posteriori* information.

Illustration: Comparison of optimization and average case approach of complexity theory.

Let H be a Hilbert space, and let $f \in H$ be unknown, with information

$$Nf = (L_1 f, \ldots, L_n f).$$

A priori information: f has small norm with respect to some operator, e.g.,

$$\|Af\| = \text{small}$$

(for example, $Af = f - \Delta f$ on a compact manifold would mean f has small Sobolev norm, i.e., is smooth).

If information is inexact, optimization approach gives

$$f = \arg\min\{\|Nf - y\|^2 + \lambda\|Af\|^2\}$$

where $y = Nf +$ error $=$ information.

If information is exact, then optimization approach gives the $\lambda \to 0$ case:

$$f = \arg\min\{\|Af\|^2 : Nf = y\}.$$

This approach regarding a priori information implies a

Bayesean viewpoint: there exists an a priori measure μ on H whose density at

$f \in H$ is a "function of" $\|Af\|$ and decreases monotonically with $\|Af\|$.

First guess at a priori density:

$$d\mu(f) = h(\|Af\|)\, df$$

this does not exist in infinite dimension unless A is invertible and A^{-1} is trace class; in that case we have a Gaussian measure completely consistent with this:

$$d\mu = \lim_{\dim\to\infty} \frac{1}{(2\pi)^{\dim/2}\det(A)} e^{-\|Af\|^2} df$$

$=$ Gaussian with covariance operator A^{-2}

We want a measure μ which "depends" only on $\|Af\|$; however, a priori assumptions do not assume it is Gaussian.

However, a general such a priori measure μ can be constructed as follows:

Define sets $H_c = \{f \in H : \|Af\| = c\}$.
We want measures μ which are “constant” on sets H_c.

Define μ_c = conditional measure of μ on H_c.

Note: μ supported on $\cup_c H_c$.

Also: measure theoretically

$$\cup_c H_c = H_1 \times \mathbb{R}^+.$$

Thus (extending definition from finite dimension), any measure $d\nu(f)$ which

depends only on $\|Af\|$ has the form of a measure on

$$H_1 \times \mathbb{R}^+ = \cup_c H_c,$$

with conditional measures $\nu_c = \mu_c$ on H_c, and some marginal measure ν_m on $\mathbb{R}^+$.

Henceforth assume the measure ν_m has finite mean.

With this as the general form of a measure consistent with our a priori assumptions, we have:

Theorem: *The ϵ-complexity of the regularization approach for any problem (under the accompanying Bayesian assumption) is equal to the ϵ-complexity of the average case complexity assuming a Gaussian measure with covariance operator A^{-2}.*

Conclusion: Complexities are entirely independent of the choice of specific a priori assumptions. All that is necessary is the regularization assumption that "$||Af||$ should be small"; all complexities consistent with this assumption can be computed from the average case setting.

Proposition: Under the same measure μ, maximum likelihood also gives the same $\epsilon-$ complexity.