NEURAL NET APPROXIMATION AND ESTIMATION OF FUNCTIONS

Andrew R. Barron

YALE UNIVERSITY
DEPARTMENT OF STATISTICS

Presentation, March 12, 2015

Mathematics Colloquium, University of Osijek

Outline

- The challenge of high-dimensional function estimation
- Univariate & multivariate polynomials, sinusoids, sigmoids
- The failure of rigid approximation models in high dimension
- Flexible approximation
  - by stepwise subset selection
  - by optimization of parameterized basis functions
- Approximation bounds (relating error to number of terms)
- Statistical risk bounds
  - relate error to number of terms and sample size
- Computational challenge
- Summary

Challenging Problem

From observational or experimental data, relate a response variable Y to several explanatory variables $X_1, X_2, \ldots, X_d$

- A fundamental task in academics and industry
- Central to the "Scientific Method"
- Used throughout science and engineering fields

Aspects of this problem are variously called:
Statistical regression, prediction, response surface estimation, analysis of variance, function fitting, function approximation, nonparametric estimation, high-dimensional statistics, data mining, machine learning, computational learning, pattern recognition, informatics, artificial intelligence, cybernetics, artificial neural networks

Core Questions

- Must there be a specific scientific hypothesis about how the best prediction of the response is related to the inputs

$$Y \approx f(X_1, X_2, \ldots, X_d, \underline{\theta})$$

- Or can the relationship be learned from data with a general flexible model?
- Must the form of the relationship be limited: with f a smooth additive function in $X_1, \ldots, X_d$, or linear in the parameter vector $\underline{\theta}$, or restricted to low-order interactions?
- Or can a selection of significant high-order interactions be learned accurately from data?
- What is the relationship between the accuracy of the fit and the number of observational cases n?

Dimensionality

The blessing and the curse of dimensionality

- With increasing number of variables d there is an exponential growth in the number of distinct terms that can be combined in modeling the function
- Larger number of relevant variables d allows in principle for better approximation to the response
- Large d might lead to a need for exponentially large number of observations n or to a need for exponentially large computation time
- Under what conditions can we take advantage of the blessing and overcome the curse.

Three papers

Some available readings from www.stat.yale.edu that illustrate my background in addressing these questions of high dimensionality

- A. R. Barron, R. L. Barron (1988). Statistical learning networks: a unifying view. In *Computing Science and Statistics: Proceedings of the 20th Symposium on the Interface*, American Statistical Association, p.192-203.
- A. R. Barron (1993). Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transactions on Information Theory*, Vol.39, p.930-944.
- A. R. Barron, A. Cohen, W. Dahmen, R. DeVore (2008). Approximation and learning by greedy algorithms. *Annals of Statistics*, Vol.36, p.64-94.

Data

- Data: $(\underline{X}_i, Y_i)$, $i = 1, 2, \ldots, n$
- Inputs: explanatory variable vectors

$$\underline{X}_i = (X_{i,1}, X_{i,2}, \ldots, X_{i,d})$$

- Domain: Either a unit cube in R^d or all of R^d
- Random design: independent $\underline{X}_i \sim P$
- Output: response variable Y_i in R
  - Moment conditions, with Bernstein constant c
- Relationship: $E[Y_i|\underline{X}_i] = f(\underline{X}_i)$ as in:
  - Perfect observation: $Y_i = f(\underline{X}_i)$
  - Noisy observation: $Y_i = f(X_i) + \epsilon_i$ with ϵ_i indep $N(0, \sigma^2)$
  - Classification: $Y \in \{0, 1\}$ with $f(\underline{X}) = P[Y = 1|\underline{X}]$
- Function: $f(\underline{x})$ unknown

Univariate function approximation: $d = 1$

Basis functions for series expansion

$$\phi_0(x), \phi_1(x), \ldots, \phi_K(x), \ldots$$

Polynomial basis (with degree K)

$$1, \quad x, \quad x^2, \ldots, x^K$$

Sinusoidal basis (with period L, and with $K = 2k$),

$$1, \cos(2\pi(1/L)x), \sin(2\pi(1/L)x), \ldots, \cos(2\pi(k/L)x), \sin(2\pi(k/L)x)$$

Piecewise constant on $[0, 1]$

$$1_{\{x \geq 0\}}, 1_{\{x \geq 1/K\}}, 1_{\{x \geq 2/K\}}, \ldots, 1_{\{x \geq 1\}}$$

Other spline bases and wavelet bases are also common

Univariate function approximation: $d = 1$

Standard 1-dim approximation models

Project to the linear span of the basis

- Rigid form (not flexible), with coefficients c_k adjusted to fit the response,

$$f_K(x) = \sum_{k=0}^{K} c_k \, \phi_k(x).$$

- Flexible form, with a subset $k_1 \ldots k_m$ chosen to best fit the response, for a given number of terms m

$$\sum_{j=1}^{m} c_j \, \phi_{k_j}(x).$$

Fit by all-subset regression (if m and K are not too large) or by forward stepwise regression, selecting from the dictionary $\Phi = \{\phi_0, \phi_1, \ldots, \phi_K\}$

Multivariate function approximation: $d > 1$

- Multivariate product bases:

$$\phi_{\underline{k}}(\underline{x}) = \phi_{k_1,k_2,\ldots,k_d}(x_1, x_2, \ldots, x_d)$$

$$= \phi_{k_1}(x_1)\phi_{k_2}(x_2)\cdots\phi_{k_d}(x_d)$$

- Rigid approximation model

$$\sum_{k_1=0}^{K}\sum_{k_2=0}^{K}\cdots\sum_{k_d=0}^{K} c_{\underline{k}}\,\phi_{\underline{k}}(\underline{x})$$

- Exponential size: $(K+1)^d$ terms in the sum
- Requires exponentially large sample size $n >> (K+1)^d$ for accurate estimation
- Statistically and computationally problematic

Flexible multivariate function approximation: $d > 1$

By subset selection:

- A subset $\underline{k}_1 \ldots \underline{k}_m$ is chosen to fit the response, with a given number of terms m

$$\sum_{j=1}^{m} c_j \, \phi_{\underline{k}_j}(\underline{x})$$

- Full forward stepwise selection:
  - computationally infeasible for large d because the dictionary is exponentially large, of size $(K+1)^d$.
- Adhoc stepwise selection: (SAS stepwise polynomials, Friedman MARS, Barron-Xiao MAPS 1991)
  - Start with $m = 1$ with $\underline{k}_1 = (0, 0, \ldots, 0)$, then for $m > 1$ restrict the search for term m to those that incrementally modify existing terms in one variable, with a manageable number of choices $(m-1)Kd$.
  - Intuitively sensible and computationally fast, but not known if it approximates well in general.

Flexible multivariate function approximation: $d > 1$

By introducting internal parameters and nonlinear least squares

- Fit functions $f_m(x) = \sum_{j=1}^{m} c_j \phi(\underline{x}, \underline{\theta})$ in the span of a parameterized dictionary $\Phi = \{\phi(\cdot, \underline{\theta}) : \underline{\theta} \in \Theta\}$
- Parameterized product bases (with continuous powers, frequencies or thresholds)

$$\phi(\underline{x}, \underline{\theta}) = \phi_1(x_1, \theta_1)\, \phi_1(x_2, \theta_2) \cdots \phi_1(x_d, \theta_d)$$

- Paramterized ridge bases (shaped like ridges of mountain range) as in projection pursuit regression models, sinusoidal models, and single-hidden-layer neural nets:

$$\phi(\underline{x}, \underline{\theta}) = \phi_1(\theta_0 + \theta_1 x_1 + \theta_2 x_2 + \ldots + \theta_d x_d)$$

- Internal parameter vector $\underline{\theta}$ of dimension $d+1$.
- The univariate function $\phi(z) = \phi_1(z)$ is called the activation function or basic nonlinearity

Building Non-linear Dictionaries

Examples of activation functions $\phi(z)$

- Perceptron networks: $1_{\{z>0\}}$ or $\text{sgn}(z)$
- Sigmoidal networks: $e^z/(1+e^z)$ or $\tanh(z)$
- Sinusoidal models: $cos(z)$
- Hinging hyperplanes: $(z)_+$
- Quadratic splines: 1, z, $(z)_+^2$
- Cubic splines: 1, z, z^2, $(z)_+^3$
- Polynomials: $(z)^q$

Notation

- Response vector: $Y = (Y_i)_{i=1}^n$ in R^n
- Dictionary vectors: $\Phi_{(n)} = \left\{(\phi(\underline{X}_i, \underline{\theta}))_{i=1}^n \ : \ \underline{\theta} \in \Theta\right\} \subset R^n$
- Sample squared norm: $\|f\|_{(n)}^2 = \frac{1}{n}\sum_{i=1}^n f^2(\underline{X}_i)$
- Population squared norm: $\|f\|^2 = \int f^2(\underline{x})P(d\underline{x})$
- Normalized dictionary condition: $\|\phi\| \leq 1$ for $\phi \in \Phi$

Flexible $m-$term nonlinear optimization

Impractical one-shot optimization

- Sample version

$$\hat{f}_m \quad \text{achieves} \quad \min_{(\underline{\theta}_j, c_j)_{j=1}^m} \| Y - \sum_{j=1}^m c_j \, \phi_{\underline{\theta}_j} \|_{(n)}^2$$

- Population version

$$f_m \quad \text{achieves} \quad \min_{(\underline{\theta}_j, c_j)_{j=1}^m} \| f - \sum_{j=1}^m c_j \, \phi_{\underline{\theta}_j} \|^2$$

- Optimization of $(\underline{\theta}_j, c_j)_{j=1}^m$ in $R^{(d+2)m}$.

Flexible $m-$term nonlinear optimization

Greedy optimizations

- Step 1: Choose c_1, $\underline{\theta}_1$ to achieve $\min \| Y - c\phi_{\underline{\theta}} \|^2_{(n)}$
- Step $m > 1$: Arrange

$$\hat{f}_m = \alpha \hat{f}_{m-1} + c\, \phi(\underline{x}, \underline{\theta}_m)$$

with α_m, c_m, $\underline{\theta}_m$ chosen to achieve

$$\min_{\alpha, c, \underline{\theta}} \| Y - \alpha \hat{f}_{m-1} - c\, \phi_{\underline{\theta}} \|^2_{(n)}.$$

- Also acceptable, with $res_i = Y_i - \hat{f}_{m-1}(\underline{X}_i)$,
  - Choose $\underline{\theta}_m$ to achieve $\max_{\underline{\theta}} \sum_{i=1}^n res_i\, \phi(\underline{X}_i, \underline{\theta})$
  - Reduced dimensionality of the search space
  - Foward stepwise selection of $S_m = \{\phi_{\underline{\theta}_1}, \ldots, \phi_{\underline{\theta}_m}\}$. Given S_{m-1}, combine the terms to achieve

$$\min_{\underline{\theta}} d(Y, span\{\phi_{\underline{\theta}_1}, \ldots, \phi_{\underline{\theta}_{m-1}}, \phi_{\underline{\theta}}\})$$

Basic $m-$term approximation and computation bounds

For either one-shot or greedy approximation

- Population version:

$$\|f - f_m\| \leq \frac{\|f\|_\Phi}{\sqrt{m}}$$

and moreover

$$\|f - f_m\|^2 \leq \inf_g \left\{ \|f - g\|^2 + \frac{2\|g\|_\Phi^2}{m} \right\}$$

- Sample version:

$$\|Y - \hat{f}_m\|_{(n)}^2 \leq \|Y - f\|_{(n)}^2 + \frac{2\|f\|_\Phi^2}{m}$$

where $\|f\|_\Phi$ is the variation of f with respect to Φ (as will be defined on the next slide).

ℓ_1 norm on coefficients in representation of f

- Consider the range of a neural net, expressed via the bound,

$$\Big|\sum_j c_j \,\mathrm{sgn}(\theta_{0,j} + \theta_{1,j}x_1 + \ldots + \theta_{d,j}x_d)\Big| \le \sum_j |c_j|$$

equality if $\underline{x}$ is in polygon where $\mathrm{sgn}(\underline{\theta}_j \cdot \underline{x}) = \mathrm{sgn}(c_j)$ for all j

- Motivates the norm

$$\|f\|_\Phi = \lim_{\epsilon \to 0} \inf \Big\{ \sum_j |c_j| \ : \ \|\sum_j c_j \phi_{\underline{\theta}_j} - f\| \le \epsilon \Big\}$$

called the variation of f with respect to Φ (B. 1991)

$$\|f\|_\Phi = V_\Phi(f) = \inf\{V \ : \ f/V \in \mathit{closure}(\mathit{conv}(\pm\Phi))\}$$

- It appears in the bound $\|f - f_m\| \le \frac{\|f\|_\Phi}{\sqrt{m}}$

ℓ_1 norm on coefficients in representation of f

- Finite sum representations, $f(\underline{x}) = \sum_j c_j \phi(\underline{x}, \underline{\theta}_j)$. Variation $\|f\|_\Phi = \sum_j |c_j|$, which is the ℓ_1 norm of the coefficients in representation of f in the span of Φ
- Infinite integral representation $f(\underline{x}) = \int e^{i\,\underline{\theta}\cdot\underline{x}}\, \tilde{f}(\underline{\theta})\, d\theta$ (Fourier representation), for $\underline{x}$ in a unit cube. The variation $\|f\|_\Phi$ is bounded by an L_1 spectral norm:

$$\|f\|_{cos} = \int_{R^d} |\tilde{f}(\underline{\theta})|\, d\underline{\theta}$$

$$\|f\|_{step} \leq \int |\tilde{f}(\underline{\theta})|\, \|\underline{\theta}\|_1\, d\underline{\theta}$$

$$\|f\|_{q-spline} \leq \int |\tilde{f}(\underline{\theta})|\, \|\underline{\theta}\|_1^{q+1}\, d\underline{\theta}$$

- As we said, this $\|f\|_\Phi$ appears in the numerator of the approximation bound.

Statistical Risk

- The population accuracy of function estimated from sample
- Statistical risk $E\|\hat{f}_m - f\|^2 = E(\hat{f}_m(\underline{X}) - f(\underline{X}))^2$
- Expected squared generalization error on new $\underline{X} \sim P$ of the estimator trained on the data $(\underline{X}_i, Y_i)_{i=1}^n$
- Minimax optimal risk bound

$$E\|\hat{f}_m - f\|^2 \leq \|f_m - f\|^2 + c\frac{m}{n}\log N(\Phi, \delta_n).$$

Here $\log N(\Phi, \delta_n)$ is the metric entropy of Φ at $\delta_n = 1/n$; with Φ of metric dimension d, it is of order $d\log(1/\delta_n)$, so

$$E\|\hat{f}_m - f\|^2 \leq \frac{\|f\|_\Phi^2}{m} + \frac{cmd}{n}\log n$$

- Need only $n >> md$ rather than $n >> (K+1)^d$.
- Best bound is $2\|f\|_\Phi\sqrt{\frac{cd}{n}\log n}$ at $m^* = \|f\|_\Phi\sqrt{n/cd\log n}$

Adaptation

- Adapt network size m and choice of internal parameters
- Complexity penalized least squares criterion.
  Let $\hat{m}$ achieve

$$\min_m \left\{ \|Y - \hat{f}_m\|^2_{(n)} + 2c\frac{m}{n} \log N(\Phi, \delta_n) \right\}$$

- Then the statistical risk (generalization error) satisfies

$$E\|\hat{f}_{\hat{m}} - f\|^2 \leq \min_m \left\{ \|f_m - f\|^2 + 2c\frac{m}{n} \log N(\Phi, \delta_n) \right\}$$

- Performs as well as if the best m^* were known in advance.

Confronting the computational challenge

- Greedy search reduces dimensionality of optimization from *md* to just *d* to obtain the current $\underline{\theta}_m$ maximizing

$$J_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} res_i \, \phi(\underline{X}_i, \underline{\theta}).$$

- This surface can still have many maxima. It provides a computational challenge. We might get stuck at an undesirably low local maximum.

- Seek insight from a special case in which the set of maxima can be identified.

Identifying the maxima

- Insight from a special case:
  - Sinusoidal dictionary: $\phi(\underline{x}, \underline{\theta}) = e^{-i\underline{\theta}\cdot\underline{x}}$
  - Gaussian design: $\underline{X}_i \sim \text{Normal}(0, \tau I)$
  - Target function: $f(\underline{x}) = \sum_{j=1}^{m_o} c_j \, e^{i\underline{\alpha}_j \cdot \underline{x}}$
- For step 1, with large n, the objective function becomes near its population counterpart

$$J(\theta) = E\left[f(\underline{X}) e^{-i\underline{\theta}\cdot\underline{X}}\right] = \sum_{j=1}^{m_o} c_j \, E\left[e^{i\underline{\alpha}_j \cdot \underline{X}} e^{-i\underline{\theta}\cdot\underline{X}}\right]$$

  which simplifies to

$$\sum_{j=1}^{m_o} c_j \, e^{-(\tau/2)\|\alpha_j - \theta\|^2}.$$

- For large τ it has precisely m_o maxima, one at each of the α_j in the target function.

summary

- Flexible approximation models
  - Subset selection
  - Nonlinearly parameterized bases as with neural nets
  - ℓ_1 control on coefficients of combination
- Accurate approximation with moderate number of terms
- Accurate estimation with a moderate sample size
- Computational challenges remain