Non-parametric Method with Algebraic Expression:

An Algorithm Based on Weierstrass Theorem

Jingwei Ma[1], Wenqi Jiang[2]

[1]Johns Hopkins University

[2]Technical University of Berlin

November 30, 2022

Abstract

Interpretability of nonparametric methods is mandatory in areas like mortgage assessments. Because of complexity of connections among activation (base) functions, ANN is criticized for low interpretability. Our target is to establish a nonparametric method, where connections among base functions are simpler. This method has similar performances with ANN and higher interpretability than ANN.

This nonparametric method is featured with three convergences: *Bernstein Polynomials* converge to true function; overfitting polynomial functions converge to *Lagrange Interpolations*; average coefficients of orthogonal base functions in the overfitting functions converge to true parameters. These convergences are of help to shun complex connections among base functions, optimizing loss functions, and regularization techniques.

This paper only discusses about numerical data.

Keywords: Bernstein polynomials, Weierstrass Theorem, non-parametric methods

Introduction

All data, functions, samples, and populations in this paper are literally numeric. This paper doesn't involve any comments on pictures, texts, voices, moods, cognitions and et ectara. Expansion explanations and analogy explanations to the following contents lead to mistakes.

Forms and parameters: why prefer function approximations?

Parametric regressions focus on estimating true functions, whose forms are already known before performing regressions. In linear regressions, for instance, form of true function is defined to be linear, thus fitting function is restricted to be linear as well. To the opposite, in nonparametric regressions, forms of true functions are not pre-established. We need to put effort into forms and parameters of true functions simultaneously.

Function approximation can be of help in nonparametric regressions. Function approximations can decompose a (true) function into a combination of base functions. Base functions are typically in similar, even identical, forms. Connections among base functions are pre-established. By changing parameters of base functions, forms and parameters of true functions would change simultaneously.

We can understand nonparametric regressions from another perspective. Before performing regressions, forms of true functions are already known as a combination of base functions. As connections among base functions have been pre-established, our only task is to estimate parameters of base functions. Hence, boundary between parametric and nonparametric methods blurs, because in both situations we only need to consider estimating parameters of variables or base functions.

Some function approximations are approximately equivalent

Some function approximations are mutually approximate. For instance, as algebraic polynomials and ANN are both function approximations, we may use them to decompose the same uniformly continuous function. Please notice that, under the same condition (that, number of base functions, n, tends to infinity), difference between ANN and true function, and difference between algebraic polynomial and true function, are approximate to 0. Hence, ANN and algebraic polynomials are approximate, if n tends to infinity.

Three errors: typical problems in nonparametric regressions

Given limited data and computational resources, nonparametric regressions with function approximations typically have three types of errors: estimation error, optimization error and approximation error. In comparison, parametric regressions typically only consider about estimation error.

Similar to parametric regressions, estimators in nonparametric regressions are also different from true parameters, hence, estimation error happens. Please notice that overfitting is a special case of estimation error. From the perspective of orthogonal base function, overfitting can be defined as including irrelevant orthogonal base functions, expected values of whom are 0 (statistical insignificancy). However, each time when performing estimation, coefficients of irrelevant orthogonal base functions are not 0, because estimators are different from expected values. This is the underlying reason for overfittings.

Connections of base functions in function approximations are complex. In such situation, optimizing loss functions can be complex as well. If gradient decent doesn't successfully minimize loss functions, optimization error takes place.

Function approximation are typically under conditions, that n tends to infinity. Hence, given limited computational resources, they are different from true functions. In most case, these differences, defined as approximation error, are decreasing when computational resources are increasing.

Concerns about interpretability of ANN

ANN is in family of function approximations. Traditional ANN use sigmoid as activation (base) functions, connections among whom are complicated.

Low interpretability in ANN partly originates from how this function approximation is generated. ANN is a function composition of activation (base) functions, thus, connections among activation functions are complex, even implicit. Albeit complexity, such connections are still known before regression. In addition, after performing regression, all coefficients of all base functions are indicated. Hence, the whole network is implicitly known after regression. To some extent, we can regard the whole network as an implicit function. From another perspective, a nonparametric regression, where connections among base functions are more straightforward, would be more interpretable.

Estimation error can also lead to interpretability problems. When including irrelevant orthogonal base functions (overfitting), forms of regressions are ponderous and difficult to be interpreted. In addition, coefficients of relevant orthogonal base functions are also different from true parameters, which would also make regressions confusing. For example, even though true parameters are positive numbers, if variances are large, estimators are possible to be negative numbers.

All these problems can be alleviated with our new model.

Basic ideas of a new non-parametric method

In some situations, for example mortgage assessments, interpretations are mandatory by legislations. Hence, introducing an interpretable nonparametric method is necessary.

As prementioned, a nonparametric regression with simple connections among base functions is more interpretable. Hence, our target is to establish an interpretable non-parametric method, where connections among base functions are linear. This method is based on the three convergences:

1. When using least squares as loss function, all estimators are featured with zero biases and high variances in polynomial functions. As all components in polynomial functions are orthogonal, repeatedly running (overfitting) regressions and taking averages of coefficients would simultaneously mitigate estimation errors of all orthogonal base functions.
2. *Lagrange Interpolation*, generating functions going through every point, *may* be a special case of overfitting. Besides, different from other types of overfitting regressions, loss functions are not necessary for generating *Lagrange Interpolation*.
3. *Lagrange Interpolation* generates algebraic polynomial function. To compare this polynomial function with true function, we need to decompose true function into a combination of algebraic polynomials as well. Thus, it would be of help to use *Bernstein Polynomials* as function approximation.

From another aspect, this new method can be regarded as a combination of *Bagging (Bootstrap Aggregating)* and overfitting *Bernstein Polynomials* regressions. Due to linearity among base functions, *Bagging* would have a good performance.

Calculations of the non-parametric method

1) Collect k numbers of datasets, independently and randomly sampled from a population. Each dataset contains $(m+1)$ numbers. Total amounts of samples are $k(m+1)$. Suppose that $k \gg 0$ and $m \gg 0$. Please notice that k and m are integers.

2) For $dataset\ 1$, using *Lagrange Interpolation*, a polynomial function can be generated, which can go through $(m+1)$ points in dataset 1:

$$y_1 = a_{10} + a_{11}x^1 + a_{12}x^2 + \cdots + a_{1m}x^m$$

Likewise, for dataset 2, dataset 3 and et ectara,

$$y_2 = a_{20} + a_{21}x^1 + a_{22}x^2 + \cdots + a_{2m}x^m$$

$$y_3 = a_{30} + a_{31}x^1 + a_{32}x^2 + \cdots + a_{3m}x^m$$

$$\ldots$$

$$y_k = a_{k0} + a_{k1}x^1 + a_{k2}x^2 + \cdots + a_{km}x^m$$

3) Finally, calculate averages of coefficients collected from different datasets. These averages will converge to true parameters.

$$\widehat{\alpha_0} = \frac{a_{10} + a_{20} + a_{30} + \cdots + a_{k0}}{k}$$

$$\widehat{\alpha_1} = \frac{a_{11} + a_{21} + a_{31} + \cdots + a_{k1}}{k}$$

$$\ldots$$

$$\widehat{\alpha_m} = \frac{a_{1m} + a_{2m} + a_{3m} + \cdots + a_{km}}{k}$$

Hence,

$$\hat{y} = \widehat{\alpha_0} + \widehat{\alpha_1}x^1 + \widehat{\alpha_2}x^2 + \cdots + \widehat{\alpha_m}x^m$$

Proof of this nonparametric method

In the following contents, we concentrate on three convergences, when m and k tend to infinity.

Assumptions

1. Denote the true function as $f(x)$. Suppose that $f(x)$ is uniformly continuous.

2. Denote a *Bernstein Polynomial* function as $B_n(f,x)$, where the highest power is n.

3. There are k different datasets and each of them contains $(m+1)$ data points. These datasets are independently and randomly selected from a true population P_{true}.

4. Assume that $m \geq n \gg 0$. Assume that $k \gg 0$. Notice that k, m and n are all integers.

Preparations

1) In $dataset\ 1$, there are $(m+1)$ data points, denote as (x_{10}, y_{10}),(x_{11}, y_{11}) ... (x_{1m}, y_{1m}). According to *Lagrange Interpolation Theorem*, a polynomial function can be generated, which can go through every point in $dataset\ 1$:

$$y_1 = \sum_{j=0}^{m} y_j\, L_j$$

Where

$$L_j = \prod_{i=0, i\neq j}^{m} \frac{x - x_i}{x_j - x_i} = \frac{x - x_0}{x_j - x_0} \cdots \frac{x - x_{j-1}}{x_j - x_{j-1}} \frac{x - x_{j+1}}{x_j - x_{j+1}} \cdots \frac{x - x_m}{x_j - x_m}$$

Coefficients are not important at this step. Denote as a_{10}, a_{11} ... a_{1m}

$$y_1 = a_{10} + a_{11}x^1 + a_{12}x^2 + \cdots + \ a_{1m}x^m$$

Likewise, using *Lagrange Interpolation Theorem* in $dataset$ 2, $dataset$ 3 ... $dataset\ k$.

$$y_2 = a_{20} + a_{21}x^1 + a_{22}x^2 + \cdots + \ a_{2m}x^m$$

$$\cdots$$

$$y_k = a_{k0} + a_{k1}x^1 + a_{k2}x^2 + \cdots + \ a_{km}x^m$$

2) (The first convergence) According to *Weierstrass Theorem*, as true function $f(x)$ is uniformly continuous, it is approximate to a combination of algebraic polynomials. *Bernstein Polynomials* can transform uniformly continuous functions into algebraic polynomials:

$$B_n(f,x) = \sum_{l=0}^{n} \binom{n}{l} f(\frac{l}{n}) x^l (1-x)^{n-l}$$

$f(\frac{k}{n})$ can change shapes of *Bernstein Polynomials*, while n changes the accuracy of approximations. Similarly, coefficients are not important at this step, denote as $b_0, b_1 \ldots b_n$

$$B_n(f,x) = b_0 + b_1x^1 + b_2x^2 + \cdots + \ b_nx^n$$

Unless n tends to infinity, $B_n(f,x)$ may not be considered as equivalent to $f(x)$. $B_n(f,x)$ uniformly converges to $f(x)$ as n increases. The convergence rate of $B_n(f,x)$ to $f(x)$ is

$$|B_n - f| \leq cW_f\left(n^{-\frac{1}{2}}\right)$$

Please notice that, in some cases, for example, that true function is linear, true function can be completely decomposed by B_n and approximation error is 0. In other cases, given limited computational resources, true functions cannot be completely decomposed. Hence, to mitigate approximation error, n should be as large as possible. When $n = m,$ the convergence rate is:

$$|B_m - f| \sim m^{-\frac{1}{2}}$$

3) Unless n (or m) tend to infinity, each y_i is generated from $f(x)$, instead of $B_n(f,x)$. For each x_i,

$$y_i = f(x_i) + residuals \neq B_n(f,x_i) + residuals$$

Hence, for $B_n(f,x_i)$, to some extent, y_i are biased, since expected values are different. In the next chapter, we consider regress $B_n(f,x_i)$ with '*biased'* data. At this stage, when proving convergences, as m tend to infinity, $B_n(f,x)$ converge to $f(x)$ and bias is out of consideration.

Proof

As we don't consider bias, data points can be regarded as generated from B_n. Consider estimating B_n with these data points. Although the highest power in B_n is n, we still insist running the following regression with highest power m ($where\ m \geq n$):

$$y = \beta_0 + \beta_1 x^1 + \beta_2 x^2 + \cdots + \beta_m x^m + residuals$$

If $m > n$, this regression includes irrelevant terms:

$$\mathrm{E}(\beta_{n+1}) = \mathrm{E}(\beta_{n+2}) = \cdots = \mathrm{E}(\beta_m) = 0$$

In matrix form:

$$Y^* = X^* \beta^* + u$$

Where Y^*, β^* and u are $(m \times 1)$ vectors, and X^* is a $(m \times m)$ matrix.

1. (The second convergence) Running this regression, using only least squares as loss function,

$$Loss\ function = u^T \cdot u = (Y^* - X^* \beta^*)^T \cdot (Y^* - X^* \beta^*)$$

$$= {Y^*}^T Y^* - 2{Y^*}^T X^* \beta^* + (X^* \beta^*)^T \cdot (X^* \beta^*)$$

By minimizing loss function, estimator is

$$\widehat{\beta^*} = ({X^*}^T X^*)^{-1} {X^*}^T Y^*$$

In this case, loss function can decrease to 0, if

$$Y^* = X^* \beta^* \Leftrightarrow {Y^*}^T Y^* - 2{Y^*}^T X^* \beta^* + (X^* \beta^*)^T \cdot (X^* \beta^*) = 0$$

Please notice that loss function decreases to 0 is an extreme case of overfitting. According to *Lagrange Interpolation Theorem*, a polynomial function, with highest power of m, can goes through m+1 data points and satisfies:

$$Y^* = X^* \beta^*$$

Hence, the polynomial function generated from *Lagrange Interpolation*, with highest power of

m, is an extreme case of overfitting regressions.

2. (The third convergence) If $m = n$, then $\widehat{\beta^*} = ({X^*}^T X^*)^{-1} {X^*}^T Y^*$ is unbiased.

When $m > n$,

$$X^* \overset{\text{def}}{=} [X, X_{irrevelant}]$$

$$\beta^* \overset{\text{def}}{=} [\beta, 0]$$

Hence, $Y^* \overset{\text{def}}{=} [Y, 0]$, where $X, \beta \; and \; Y$ exclude irrelevant terms. Hence, substitute

$$\widehat{\beta^*} = ([X, X_{irrevelant}]^T [X, X_{irrevelant}])^{-1} [X, X_{irrevelant}]^T [Y, 0]$$

If overfitting doesn't happen, when using least squares as loss function, estimators are unbiased.

If substitute $(X^T X)^{-1} X^T Y$ with $\hat{\beta}$,

$$\widehat{\beta^*} = [\hat{\beta}, 0]$$

$$E(\hat{Y}) = E([X, X_{irrevelant}]\widehat{\beta^*}) = E(X\hat{\beta}) = Y$$

Hence, estimators in overfitting polynomial regressions are unbiased.

3. Using *Lagrange Interpolation* in $dataset\ 1$ to k,

$$y_1 = a_{10} + a_{11}x^1 + a_{12}x^2 + \cdots + \ a_{1m}x^m$$

$$\ldots$$

$$y_k = a_{k0} + a_{k1}x^1 + a_{k2}x^2 + \cdots + \ a_{km}x^m$$

As mentioned above, a_{10}, ... a_{1m}, a_{20} ... a_{km} are unbiased estimators. Since data points are independently and randomly sampled from true population, estimators, for example, a_{10}, ... a_{k0}, are mutually independent from as well. According to *Central Limitation Theorem*, as $k \gg 0$, averages of a_{10}, ... a_{k0}, or, a_{11} ... a_{k1} et cetera will approximately be normally distributed. At this stage, when proving convergences, as m tend to infinity, we may ignore the impact of approximation error on estimators' variance, thus, convergence rate only depends on $k^{-\frac{1}{2}}$. When k tends to infinity, averages of estimators converge to true parameters.

Comparing the new method with counterparty (ANN)

ANN and algebraic polynomials are approximately equivalent

Apart from what is mentioned above, we can understand approximate equivalence between ANN and algebraic polynomials from another perspective. Traditional activation function used in ANN, sigmoid function, is uniformly continuous; subsequently, it can be approximate to algebraic polynomials. In addition, when using polynomial functions as activation functions, the whole network will become a polynomial function, because function compositions of polynomials are still polynomials. Directly performing polynomial regressions can bypass using function compositions, which lead to low interpretability.

Potential drawbacks of new model: a trade-off

This new method is designed for high interpretability. Because three convergences, estimators in this method are asymptotically unbiased. However, there are potential concerns on accuracy. In the following contents, we consider about accuracy under the background of multivariate. Hence, we no longer assume that m and k tends to infinity. Meanwhile, we suppose there are p numbers of variables. Obviously, m, k and p are all integers and total numbers of samples are $k(m+1)p$.

There is a potential concern of the trade-off between approximation accuracy and computational resources. In *Bernstein Polynomials*, convergence rate depends on $m^{-\frac{1}{2p}}$. When using sigmoid function as activation function, convergence rate of ANN, on the contrary, is dependent on only on $m^{-\frac{1}{2}}$ and not related with p. Hence, if p is large, given the same computational resources, ANN with sigmoid function has lower approximation error.

However, as mentioned before, approximation error is only a part of total errors:

$$\underbrace{f_{true} - \hat{f}}_{total\ error} = \underbrace{[f_{true} - f_{approximation}]}_{approximation\ error} + \underbrace{\left[f_{approximation} - f_{\widehat{est\imath mated}}\right]}_{estimation\ error} + \underbrace{\left[f_{\widehat{est\imath mated}} - f_{\widetilde{opt\imath mz\imath ed}}\right]}_{optimization\ error}$$

The polynomial algorithm, introduced in this paper, have advantages when handling another two types of errors. Hence, if given the same computational resources, total accuracy of the new algorithm is possible to be higher than that of ANN.

Estimation error is about variances of both relevant and irrelevant estimators. In polynomial regressions, as $k \gg 0$, averages of estimators are approximately normally distributed. If ignore the approximation bias, convergence rate of estimators to true parameters only depends on $k^{-\frac{1}{2}}$, equal to Monte Carlo rate. When taking bias into consideration, standard error of estimators will be lower. Hence, holding other factors constant, convergence rate of estimators to true parameters would be even better than Monte Carlo rate. Another byproduct of taking averages of estimators is that statistical inferences (or cross-validations) may not be necessary, as variances of estimators are extremely small.

As a contrast, taking coefficients' averages may not of help in ANN, where activation functions (base functions) are not orthogonal, and connections among them are not additive. Meanwhile, regularizations, pervasively used in ANN, make all estimators biased. Hence, ANN is not advantageous when handling estimation error.

Optimization error is related with process of finding optimized estimators. Please notice that *Lagrange Interpolation* doesn't use loss functions, subsequently doesn't involve optimization processes.

In summary, as convergences are proved, estimators in our model are asymptotically unbiased. Hence, investing more computational resources can improve accuracy. ANN can be regarded as a tool to make approximation error lower, however, our method needs less computational resources in other steps, including cross-validation, optimizing loss functions and regularizations.

Meanwhile, as prementioned, ANN and algebraic polynomials are approximately equivalent. If we are required to make predictions with low approximation errors in limited time, while interpretations are mandatory by legislations, we may use ANN to predict and polynomials to interpret. After making predictions, use extra time to translate ANN into polynomials, by substituting activation functions with approximate algebraic polynomials.

Conclusion

Complexity of connections among activation functions in ANN lead to low interpretability. Our new non-parametric method based on *Weierstrass Theorem* is aimed to perform non-parametric estimations with algebraic polynomial functions, hence has higher interpretability than ANN. Overfitting problems, including irrelevant orthogonal base functions, are detrimental to interpretability as well. Our new method also alleviates overfitting problems.

This method is approximately equivalent to a combination of *Bagging (Bootstrap Aggregating)* and polynomial regression. Because three convergences, estimators in this method are asymptotically unbiased. In terms of accuracy, comparing with ANN, it could have relatively higher approximation error in some situations, but have advantages when handling optimization error and estimation error.

Other advantages may include: statistical inferences (or cross-validations) may not be necessary, as variances of estimators are extremely small; *Bagging (Bootstrap Aggregating)* is already embedded in this method.