High Dimensional Nonparametric Regresssion via Additive Kernel Ridge Models

Anonymous Author(s)
Affiliation
Address
email

Abstract

We describe additive kernel ridge regression, a generalisation of the kernel ridge method for nonparametric regression. ...

1 Introduction

Regression in high dimensions is an inherently difficult problem with known lower bounds depending exponentially in dimension [3]. In this project we intend to make progress in this problem by treating the function as an additive model of lower dimensional components. Using additive models is fairly standard in high dimensional regression literature [4, 5, 7]. However, in this work we wish to consider additive models which are more general/ expressive than previous work.

There are a number of potential nonparametric methods for modeling the low-order interaction terms. One option is to use multidimensional splines. Thin plate splines [10] extend spline-based nonparametric regression to multiple covariates, although computational complexity increases dramatically with the order of interaction. Natural thin plate splines extend one-dimensional smoothing splines. These can be fit via penalized regression, and complexity can be reduced by choosing knots on a grid rather than at each data point. Thin plate (penalized) regression splines can alternatively be used; this approach requires truncated SVD, but avoids choice of knots. Tensor product splines [8] can be used to construct multivariate splines as tensor products of single-dimensional splines. The number of basis functions grows exponential with the interaction order, but they can be fit via penalized regression.

Another option is to model the low-order interaction basis functions implicitly using kernels. There is existing research on using linear combinations of kernels for kernel learning, called multiple kernel learning [2]. Computationally efficient use of additive kernels has also been explored in Bayesian settings [1].

Our work extends Sparse Additive Models (SpAM) [7] to multi-dimensional nonparametric basis functions. For Sparse Additive Models, parameters are typically optimized using the backfitting algorithm. Our work also extends recent work on Generalized Additive Models plus Interactions [6]. However, in this work the interaction model was assumed to follow a specific functional form, leading to an optimization method tailored to their interaction model.

2 Problem Statement

Let $f : \mathcal{X} \to \mathbb{R}$ be the function of interest. Here $x = [x_1, \ldots, x_D] \in \mathbb{R}^D$ and $\mathcal{X} \subset \mathbb{R}^D$. We have data $(X_i, Y_i)_1^n$ and wish to obtain an estimate $\hat{f}$ of f. In this work, we will be assuming that f has the form

$$f(x) = f^{(1)}(x^{(1)}) + f^{(2)}(x^{(2)}) + \cdots + f^{(M)}(x^{(M)}), \tag{1}$$

where $x^{(j)} \in \mathcal{X}^{(j)} \subset \mathbb{R}^{d_j}$ and $f^{(j)} : \mathcal{X}^{(j)} \to \mathbb{R}$. We shall refer to the $\mathcal{X}^{(j)}$'s as groups. In this work, we are particularly interested in the case where D is very large and the group dimensionality is bounded– i.e. $d_j \leq d \ll D$.

The work in Hastie and Tibshirani [4] treats f as a sum of one dimensional components. In Equation 1 this corresponds to $x^{(j)} = x_j$, $d_j = d = 1 \; \forall j$ and $M = D$. We would like to be more expressive than this. One option would be to allow for all interactions of up to order d. But this requires $M \approx \mathcal{O}(D^d)$ which poses computational challenges both for training and evaluation during run time.

In this work, we will, at least initially, be focusing on kernel regression. The Nadaraya Watson estimator [9] is a popular kernel smoothing method which estimates the function via,

$$\hat{f}(t) = \frac{\sum_{i=1}^{n} \mathcal{K}(t, X_i) Y_i}{\sum_{i=1}^{n} \mathcal{K}(t, X_i)}$$

A natural way to handle additive models is to use an additive kernel of the form

$$\mathcal{K}(x, x') = \alpha_1 \mathcal{K}^{(1)}(x^{(1)}, {x^{(1)}}') + \alpha_2 \mathcal{K}^{(2)}(x^{(2)}, {x^{(2)}}') + \ldots \alpha_M \mathcal{K}^{(M)}(x^{(M)}, {x^{(M)}}')$$

The optimization problem here then is to learn $\alpha \in \mathbb{R}^M$ and the hyperparameters of the kernels $\mathcal{K}^{(j)}$ by minimizing the cross validation error. If we fix the hyperparameters of the kernel, the problem is convex in α but otherwise the problem is generally nonconvex.

As a first step, we wish to study the problem when M is manageable. For this, we could select the groups either randomly or via some greedy procedure. Alternatively, [1] use a trick based on elementary symmetric polynomials to efficiently compute additive kernels of up to all orders of interaction. This however, will require some parameter sharing between the kernels and not be as expressive. We shall first explore these paths before delving deep into more sophisticated additive models. Outside optimization, we also wish to study some of the statistical properties of the function such as rate of convergence and minimaxity.

References

[1] David K. Duvenaud, Hannes Nickisch, and Carl Edward Rasmussen. Additive gaussian processes. In *Advances in Neural Information Processing Systems*, 2011.

[2] Mehmet Gönen and Ethem Alpaydın. Multiple kernel learning algorithms. *The Journal of Machine Learning Research*, 12:2211–2268, 2011.

[3] László Györfi, Micael Kohler, Adam Krzyzak, and Harro Walk. *A Distribution Free Theory of Nonparametric Regression*. Springer Series in Statistics, 2002.

[4] T. J. Hastie and R. J. Tibshirani. *Generalized Additive Models*. London: Chapman & Hall, 1990.

[5] John D. Lafferty and Larry A. Wasserman. Rodeo: Sparse Nonparametric Regression in High Dimensions. In *NIPS*, 2005.

[6] Yin Lou, Rich Caruana, Johannes Gehrke, and Giles Hooker. Accurate intelligible models with pairwise interactions. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 623–631. ACM, 2013.

[7] Pradeep Ravikumar, John Lafferty, Han Liu, and Larry Wasserman. Sparse Additive Models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2009.

[8] Charles J Stone. The use of polynomial splines and their tensor products in multivariate function estimation. *The Annals of Statistics*, pages 118–171, 1994.

[9] Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Publishing Company, Incorporated, 2008.

[10] Simon N Wood. Thin plate regression splines. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(1):95–114, 2003.