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 Set up & Algorithm

In this section we describe the problem, the proposed algorithm and its objective and our methods for optimising the objective.

2.1 Problem Statement & Notation

Let $f : \mathcal{X} \rightarrow \mathbb{R}$ be the function of interest. Here $x = [x_1, \dots, x_D] \in \mathbb{R}^D$ 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 seek additive approximations to the function. That is, $\hat{f}$ can be expressed as,

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

where $x^{(j)} \in \mathcal{X}^{(j)} \subset \mathbb{R}^{d_j}$ and $\hat{f}^{(j)} : \mathcal{X}^{(j)} \rightarrow \mathbb{R}$. We shall refer to the $\mathcal{X}^{(j)}$'s as groups and the collection of all groups $\bigcup_j \mathcal{X}^{(j)}$ as the decomposition. 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 $\hat{f}$ as a sum of one dimensional components. The decomposition here corresponds to $x^{(j)} = x_j$, $d_j = d = 1\ \forall j$ and $M = D$. In this project, we would like to be more expressive than this model. We will consider decompositions for which $d > 1$ and more importantly allows for overlap between the groups. Ravikumar et al. [7] treat $\hat{f}$ as a sparse combination of one dimensional functions. While this is seemingly restrictive than [4], the sparse approximation may provide favourable bias-variance tradeoffs in high dimensions. Drawing inspiration from this, we will consider models where M is very large and seek a sparse collection of groups to approximate the function - i.e. $\hat{f}^{(j)} = \mathbf{0}$ for several j.

2.2 Kernel Ridge Regression

One method to formulate a regression problem is to consider the following optimisation problem.

2.3 Additive Kernel Ridge Regression

Therefore the optimisation objective can be written as,

$$F(\boldsymbol{\alpha}) = \frac{1}{2}\|Y - \sum_{j=1}^{m} K^{(j)}\alpha^{(j)}\|_2^2 + \frac{1}{2}\sum_{j=1}^{M} {\alpha^{(j)}}^\top K^{(j)}\alpha^{(j)} + \sum_{j=1}^{M}\|\alpha^{(j)}\|_2 \tag{2}$$

where $\alpha^{(j)} \in \mathbb{R}^n\ \forall j$, $\boldsymbol{\alpha} = [{\alpha^{(1)}}^\top, \dots, {\alpha^{(M)}}^\top]^\top \in \mathbb{R}^{nM}$ and $K^{(j)} \in \mathbb{R}^{n \times n}\ \forall j$.

3 Optimisation Methods

3.1 SubGradient Method

3.2 Proximal Gradient Method

Note that we can write the objective (2) as $F(\boldsymbol{\alpha}) = G(\boldsymbol{\alpha}) + \Psi(\boldsymbol{\alpha})$ where G is smooth and Ψ is not. Ψ is the popular group lasso penalty. Via the Moreau decomposition, and using the fact that the argument in the prox operator is separable, the prox operator can be shown to be,

$$[\text{prox}_{\Psi,t}(\boldsymbol{\alpha})]^{(j)} = \text{prox}_{\Psi,t}(\alpha^{(j)}) = \begin{cases} \alpha^{(j)} - t\frac{\alpha}{\|\alpha^{(j)}\|_2} & \text{if } \|\alpha^{(j)}\| < t \\ \mathbf{0} & \text{otherwise} \end{cases}$$

Obtaining the prox operator takes $O(nM)$ time and has cost comparable to computing the gradient of $G(\boldsymbol{\alpha})$. We use the above to implement proximal gradient method as described in the class notes. We use backtracking to determien the step size and experiment both with and without acceleration.

3.3 BCD - use full name

3.4 BCGD

4 Experiments

5 Conclusion

The method proposed seems promising. Going forward we wish to perform comparisons with other nonparametric regression methods such as Gaussian process regression, locally polynomial regression and other additive models [1, 4, 7] on other synthetic and real datasets. One challenge will be in scaling this to large datasets as we need to construct several $n \times n$ kernel matrices. We wish to explore some ideas on this front too.

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.