Abstract

The additive model is one of the most popularly used models for high dimensional nonparametric regression analysis. However, its main drawback is that it neglects possible interactions between predictor variables. In this paper, we reexamine the group additive model proposed in the literature, and rigorously define the intrinsic group additive structure for the relationship between the response variable Y and the predictor vector X, and further develop an effective structure-penalized kernel method for simultaneous identification of the intrinsic group additive structure and nonparametric function estimation. The method utilizes a novel complexity measure we derive for group additive structures. We show that the proposed method is consistent in identifying the intrinsic group additive structure. Simulation study and real data applications demonstrate the effectiveness of the proposed method as a general tool for high dimensional nonparametric regression.

Motivation

Existing methods to perform high dimensional nonparametric regression have some drawbacks:

- Old methods like Partial Linear Models fail to consider interactions between the predictors.
- Grouping the predictors using a tree can work well but is computationally expensive.
- Other methods like Group LASSO may identify a suboptimal group structure for high dimensional data.

Our focus is to propose new algorithms that can be applied in general high dimensional nonparametric regression.

Methodology

Partial Order between Group Additive Structures:

G is a sub group additive structure of G' (G \subseteq G') if for every group g \in G there is a group v \in G' such that g \subseteq v.

Group Additive Structures and Their Induced Function Spaces:

- G_1 \subseteq G_2 \Rightarrow \mathcal{L}_{G_1} \subseteq \mathcal{L}_{G_2}
- G_2 \subseteq G_3 \text{ and } X_1, \ldots, X_n \text{ are independent} \Rightarrow \mathcal{L}_{G_2} \subseteq \mathcal{L}_{G_3}

Different Types of Group Additive Structures:

1. **Amiable**: The induced function space of G contains the true non-parametric function of the model.
2. **Non-amiable**: A group additive structure that is not amiable.
3. **Intrinsic**: The smallest amiable group additive structure G*.

For statistical modeling, G* achieves the greatest dimension reduction for the relationship between Y and X.

Complexity Measure of Group Additive Structure:

- Use size of the RKHS as a proxy of the complexity of group additive structure.
- Size of RKHS is measured by covering number.
- Expression of the complexity of a group additive structure: C(G) = \sum_{g \in G} \phi(g)

Model Estimation Consists of Two Steps:

- **STEP 1**: When the group structure is known, it can be estimated by using kernel methods to solve:
  \[ \hat{f}_{g,\lambda} = \arg\min_{f} \frac{1}{n} \sum_{i=1}^{n} (y_i - f(x_i))^2 + \lambda \|f\|_{\mathcal{L}_{G}}^2 \]
  - **STEP 2**: Optimal group structure is chosen to minimize both fitting error and group structure complexity:
  \[ \hat{G} = \arg\min_{G} \{ \sum_{g \in G} \|f_{g,\lambda}\|_{\mathcal{L}_{G}}^2 + \mu C(G) \} \]

Real Data

- **Boston Housing Data**: Backward step-wise algorithm is used.
- **Crime Data**: Forward step-wise algorithm is used.

Simulation

- Exhaustive search algorithm without parameter tuning could often identify the intrinsic group structures.
- Stepwise algorithm could often identify the intrinsic group structures for some models.

Future Research

- Our complexity measure of group additive structure is based on the covering number of RKHSs. There exist other ways to define the complexity measure. A thorough comparison of defining the complexity measure with different methods deserves further research.
- The current backward step-wise algorithm may become unstable and fail to achieve the potential in identifying the intrinsic group additive structure. It is of great interest to further improve the algorithm or propose new algorithms that can be applied in general high dimensional nonparametric regression.