Multiresolution dictionary learning for conditional distributions

Abstract

Nonparametric estimation of the conditional distribution of a response given high-dimensional features is a challenging problem. In many settings it is important to allow not only the mean but also the variance and shape of the response density to change flexibly with features, which are massive-dimensional with a distribution concentrated near a lower-dimensional subspace or manifold. We propose a novel multiresolution approach which starts by learning a multiscale dictionary of densities, constructed as Gaussian within each set of multiscale partition tree for the features. We place a novel multiresolution stick-breaking process prior on the dictionary weights to construct a conditionally Bayesian approach. The algorithm scales efficiently to massive numbers of features, and can be implemented efficiently with slice sampling. State of the art predictive performance is demonstrated for toy examples and an important application to predicting drug response to anti-depressants.

Key words: Density regression; Dictionary learning, Manifold learning; Mixture of experts; Multiresolution stick-breaking; Nonparametric

1. Introduction

2. Model Specification

2.1. Model Structure

Let $x_i = (x_{i1}, \ldots, x_{ip})' \in \mathcal{X}$ denote a vector of predictors, with p large, and let $y_i \in \mathcal{Y}$ denote a response variable. We are interested in estimating the conditional density $f(y|x)$ of the response given predictors, while taking into account that the predictors are high-dimensional but thought to be concentrated near a lower-dimensional subspace $\mathcal{M}$ embedded in $\Re$ that may or may not correspond to a Riemannian manifold. There is a literature on obtaining multiscale representations of such subspaces.

Suppose we define a multiscale partition of $\mathcal{X}$. Generation one corresponds to the entire $\mathcal{X}$. At generation two, $\mathcal{X}$ is split into two mutually exclusive partition sets, $\mathcal{X}_1$ and $\mathcal{X}_2$ with $\mathcal{X} = \mathcal{X}_1 \bigcup \mathcal{X}_2$ and $\mathcal{X}_1 \bigcap \mathcal{X}_2 = \emptyset$. At each subsequent generation, each set is further refined into two children. For example, in generation $j = 3$, $\mathcal{X}_1$ is split into $\mathcal{X}_{11}$ and $\mathcal{X}_{12}$ while $\mathcal{X}_2$ is split into $\mathcal{X}_{21}$ and $\mathcal{X}_{22}$. This proceeds for k generations.

Any predictor value $x \in \mathcal{X}$ has a corresponding ancestral history $A(x) \in \{1,2\}^k$ encoding their location in the partition tree, with $A_j(x) \in \{1,2\}^j$ denoting the elements up to generation j and the first element equal to 1 by definition. We characterize the conditional density $f(y|x)$ as a convex combination of multiscale *dictionary* densities. At level one, the global parent density is denoted f_1. In generation two, the dictionary densities corresponding to partition sets $\mathcal{X}_1$ and $\mathcal{X}_2$ are denoted f_{11} and f_{12}. For $j = 2, \ldots, k$, the dictionary density at generation j for ancestry $a_j \in \{1,2\}^j$ is f_{a_j}. The resulting conditional density is characterized as

$$f(y|x) = \sum_{j=1}^{k} \pi_{A_j(x)} f_{A_j(x)}(y), \tag{1}$$

where $0 \leq \pi_{a_j} \leq 1$ and $\sum_{j=1}^{k} \pi_{a_j} = 1$ for any $a \in \{1,2\}^k$ with a_j the first j elements of a.

Each $A_j(x)$ is a $j \times 1$ binary vector encoding the path through the partition tree up to generation j specific to predictor value x. For two predictor values x and x' located close together, it is expected that the paths will be similar, which leads to similar weights on the dictionary densities. In the extreme case in which x and x' belong to the same leaf partition set, we have $A(x) = A(x')$ and the path through the tree will be the same. In this case, we also will have $f(y|x) = f(y|x')$ so that up to k levels of resolution the densities $f(y|x)$ and $f(y|x')$ are identical. If the paths through the tree differ only in the final generation or two, the weights will typically be similar but the resulting conditional densities will not be identical.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

2.2. Multiscale Stick-Breaking

In the first stage, apply a multilevel partitioning approach to obtain a multilevel partition that we will then consider as fixed. In addition, for $j = 1, \ldots, k$ and $a \in \mathcal{A}_j$ the dictionary densities are estimated by assuming a normal form. This is done by taking the data for all subjects having predictors in that partition set and estimating the dictionary density based on only these data. To be specific, for estimating density $f_{a_j}(y)$, we use the data $\{y_i : x_i \in \mathcal{X}_{a_j}\}$, noting that $a_j \in \{1, 2\}^j$ is a j-dimensional binary vector.

We then conduct the analysis treating the dictionary elements and partition sets as fixed and placing a prior on the weights $\pi_a^{(j)}$ for all $a \in \{1, 2\}^k$. A natural choice corresponds to a multiscale stick-breaking process that is defined as follows. For each node in the binary partition tree including the root node, define a stick length, $V_{a_j} \sim \text{beta}(1, \alpha)$, for $a_1 = 1$ and $a_j \in \{1, 2\}^j, j = 2, \ldots, k$. The parameter α encodes the complexity of the model, with $\alpha = 0$ corresponding to the case in which $f(y|x) = f_q(y)$. We relate the weights in (1) to the stick-breaking random variables as follows:

$$\pi_{a_j} = V_{a_j} \prod_{h=1}^{j-1} (1 - V_{a_h}), \quad j = 1, \ldots, k-1, \qquad (2)$$

with $V_{a_k} = 1$ for the leaf stick-breaking random variables to ensure that $\sum_{j=1}^{k} \pi_{a_j} = 1$ for any $a \in \{1, 2\}^k$ with $a_1 = 1$ and a_j denoting the first j elements of a.

2.3. Estimation

1. Step Estimate dictionary densities
2. Step Run Gibbs-Sampler to estimate multiscale stick-breaking weights. Introduce $S_i \in \{1, \ldots, k\}$ for $i = 1, \ldots, n$ denoting the level of the dictionary density that subject i uses. Let n_{a_j} be the number of observations allocated to the a_jth subgroup in the jth level. Then,

1. Update S_i by sampling from the multinomial full conditional with

$$\Pr(S_i = j \mid -) = \frac{\pi_{A_j(x_i)} f_{A_j(x_i)}(y_i)}{\sum_{h=1}^{k} \pi_{A_h(x_i)} f_{A_h(x_i)}(y_i)}$$

2. Update stick-breaking random variable V_{a_j}

3. Simulation Studies

In order to test the predictive performance of the proposed model the following simulation examples were considered. First, let us define n_T and n_P as the number of observations in the training and testing sets. Let y_i, for $i = 1, \ldots, n$, be the response variable and $x_i \in \mathcal{R}^p$ a set of predictors. For all examples below, dictionary densities were considered to be normal with parameters obtained through maximum likelihood estimation. To sample stick breaking weights, the gibbs sampler was run for $20,000$ iterations with a burn-in of $1,000$.

Table 1. Predictive Mean Squared Error (MSE) under multiscale stick-breaking (MSB), CART and random forest (RF)

STUDY	(n_T, n_P)	MSB	$CART$	RF
(1)	(50, 20)	0.04	0.06	0.17
	(90, 20)	0.07	0.09	0.10
	(150, 20)	0.03	0.04	0.07
(2)	(50, 20)	5.37	23.30	11.87
	(90, 20)	5.45	7.41	9.14
	(150, 20)	7.84	11.35	8.66

In the first simulation study, x_{i1} was drawn from a two components mixture of equally weighted normals with mean and variance $(-4, 1)$ and $(4, 1)$ respectively. Variables x_{ij}, for $j = 2, \ldots, 10$, were drawn from a standard normal density. Then, draw y_i from a normal with parameter $(3, 1)$ if $x_{i1} \sim N(-4, 1)$ and parameter $(-3, 1)$ otherwise.

In the second simulation study, x_{i1} was drawn from a two components mixture of equally weighted normals with mean and variance $(-2, 1)$ and $(2, 1)$. Variable x_{i2} was drawn from a two components mixture of equally weighted normals with mean and variance $(5, 1)$ and $(10, 1)$. Let $x_{ij} \sim N(0, 1)$ for $j = 3, \ldots, 5$. Let y_i only depend on regressors (x_{i1}, x_{i2}) through the linear model $y_i = 2x_{i1} + 2x_{i2} + \epsilon_i$ with $\epsilon_i \sim N(0, 1)$.

Table 1 shows predictive mean squared errors under the proposed model, classification and regression trees (CART) and random forest (RF) for different values of (n_T, n_P).

4. Real Data

We tested the predictive performance of the proposed multiscale model on several datasets. The first one is a study involving 110 patients with depression who have been taking anti-depressive medications for eight week. Subjects started taking drug at baseline so the baseline measurement gives a measure of how depressed they are before they start the current treatment. Two anti

Table 2. Predictive Mean Squared Error under Metabolomics data for multiscale stick-breaking (MSB), CART and random forest (RF)

(n_T, n_P)	MSB	$CART$	RF
$(50, 60)$	0.68	1.99	0.90
$(60, 50)$	0.72	1.76	0.94
$(70, 40)$	0.71	1.80	0.91
$(80, 30)$	0.73	1.50	0.85
$(90, 20)$	0.70	1.70	0.81
$(100, 10)$	0.89	2.11	0.89

depressive indexes, i.e. QIDS and HAMD, were measured at baseline, four weeks and eight weeks. This study aimed to predict the change in depression indexes using metabolite data. For each subjects a set of covariates involving medication dose and subject characteristics such as age, race and gender were observed. We will start considering a univariate measurement given by the change occurring in QIDS between the baseline and the fourth week. Table 2 shows predictive mean squared error under the proposed approach, classification and regression trees (CART) and random forest (RF).

5. Citations and References