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 multiresolution model based on a novel stick-breaking prior placed on the dictionary weights. 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

Massive datasets arise from a variety of sources including neurological, image, video and biological applications. In all these applications, common models utilized for density estimation, classification, variable selection and predictions fail to be efficient and cannot be applied. Dealing with large amounts of data requires the introduction of new models able to process the data accurately and efficiently. In this paper, we will focus on conditional density estimation for massive datasets. Conditional density estimation aims to estimates the density of the response $y \in \mathcal{Y}$ given a set of predictors $(x_1, x_2, \ldots, x_p) \in \mathcal{X}$. Though, a variety of flexible models have been proposed in the last decade, density estimation remains challenging for large sample sizes and high dimensional predictors.

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

The need to deal with a large number of observations motivated the literature on divide-and-conquer techniques, a class of algorithms extensively used in density estimation, classification and prediction. Well known examples are classification and regression trees (CART) (Breiman et al., 1984) and multivariate adaptive regression trees (MARS) (Friedman, 1991). These algorithms fit surfaces to data by explicitly dividing the input space into a nested sequence of regions, and by fitting simple surfaces within these regions. Though these methods are appealing to reduce the dimensionality of the problem, single tree estimates are generally associated to high variance. A possible solution to this problem would be combining estimates resulting from different trees. Well known examples are bagging (Breiman, 1996), boosting (Shapire et al., 1998) and random forest (Breiman, 2001). Though these algorithms can substantially reduce the variance, they can be computationally intensive.

Mixture of experts (Jacobs et al., 1991) is another divide-and-conquer algorithm particularly useful to reduce the variance associated to single tree estimates. As opposed to other divide-and-conquer algorithms, mixture of experts rely on soft partitioning algorithms that allows observations to lie simultaneously in different subsets. A mixture of experts model is a mixture model in which the model parameters, including mixture weights, are functions of covariates. Several mixture of experts models have been proposed in the last twenty years. some of them gain flexibility by dealing with infinitely many experts (Rasmussen & Ghahramani, 2002) (Meeds & Osindero, 2006), others propose a hierarchical structure where a mixture model is fit in each subset (Jordan & Jacobs, 1994) (Bishop & Svensen, 2003).

A significant downside of all divide-and-conquer algorithms is their poor scalability to high dimensional predictors. As the number of features increases, the problem of finding the best splitting attribute becomes intractable so that CART, MARS and multiple trees models cannot be efficiently applied. Also mixture of experts models become computationally demanding, since both mixture weights and dictionary density are predictor dependent. In an attempt to make mixture of experts more efficient sparse extensions relying on

different variable selection algorithms have been proposed (Mossavat & Amft, 2011). However, performing variable selection in high dimensions is still a challenging problem, especially when multiple parameters involved in the model, such as weights and mean functions, depend on high dimensional predictors.

In order to efficiently deal with massive datasets, 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. The proposed approach is based on a two-stage algorithm where first the predictor space is recursively partitioned and then, considering the partition as fixed, a multiresolution stick-breaking process is estimated. According to the proposed process, observations can lie simultaneously in subsets located at different resolution levels. This results in a model that allows borrowing information across different resolution levels and reaches a good compromise in terms of the bias and variance trade-off. The tree partition is found by implementing a fast multiscale technique used for graph partitioning (Karypis & Kumar, 1999). We show that the algorithm scales efficiently to massive numbers of features, and can be implemented efficiently with gibbs sampling. State of the art predictive performance is demonstrated for toy examples and an important application to predicting drug response to anti-depressants.

2. Model Specification

2.1. Model Structure

This paper will mainly focus on the problem of estimating the conditional density $f(y|x)$ of the response given a high dimensional vector of predictors, concentrated near a lower-dimensional subspace $\mathcal{M}$ embedded in $\mathcal{X}$. In dealing with massive datasets, the idea of combining different local models defined on subsets of the predictor space have increased popularity in the last decade. These methods, generally known as divide-and-conquer algorithms, aim to replace complicated conditional density functions with a combination of simple densities defined locally, on subsets of the predictor space.

Suppose we define a multiscale partition of $\mathcal{X}$. Generation one corresponds to the entire $\mathcal{X}$ denoted as $\mathcal{X}^1$. At generation two, $\mathcal{X}^1$ is split into two mutually exclusive partition sets, $\mathcal{X}^1 = \left(\mathcal{X}_1^2, \mathcal{X}_2^2\right)$. Each subset is recursively partitioned into two subsets so that for a general partition level ℓ the partition will be given by $\mathcal{X}^\ell = \left(\mathcal{X}_1^\ell, \ldots, \mathcal{X}_{2^{\ell-1}}^\ell\right)$. Let us assume this process proceeds for k levels. Let (ℓ, s) be the node associated to the sth subset at resolution level ℓ. Let $ch(\ell, s)$ and $pa(\ell, s)$ be respectively the set of children and parents of node (ℓ, s). Let $A_\ell(x) \in \{1, \ldots, 2^{\ell-1}\}$ be the location of predictor x at level ℓ, with $A_1(x)$ 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. The dictionary density at generation j is f_{B_j} with $B_j = \{j, A_j\}$, for $j = 2, \ldots, k$. Then, $f(y|x)$ is defined as the convex combinations of densities $\{f_{B_j}\}_{j=1}^k$ with weights $\{\pi_{B_j(x)}\}_{j=1}^k$, i.e.

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

where $0 \leq \pi_{B_j(x)} \leq 1$ and $\sum_{j=1}^k \pi_{B_j(x)} = 1$.

Each $B(x)$ is a set encoding the path through the partition tree up to generation k specific to predictor value x. According to model 1, one observation can simultaneously lie in subsets located at different resolution levels. This is particularly useful to reach a good compromise between the bias and variance trade-off as explained in the next section. Moreover, it allows borrowing information across different resolution levels. Though the proposed approach reminds a mixture of experts model (Jacobs et al., 1991), the two approaches are completely different since under (1) neither mixture weights nor dictionary densities directly depend on predictors. This allows our model to scale efficiently to high dimensional predictors.

Now let us examine the implications of model (1). 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 $B(x) = B(x')$ and the path through the tree will be the same. Moreover, in this case, we 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.

To derive mixture weights, a natural choice corresponds to a stick-breaking process (Sethuraman, 1994). For each node $B_j(x_i)$ in the binary partition tree, define a stick length $V\{B_j(x_i)\} \sim \text{beta}(1, \alpha)$. The parameter α encodes the complexity of the model, with $\alpha = 0$ corresponding to the case in which $f(y|x) = f(y)$. We relate the weights in (1) to the

stick-breaking random variables as follows:

$$\pi_{B_j(x)} = V\{B_j(x)\} \prod_{B_h \in pa\{B_j\}} [1 - V\{B_h(x)\}],$$

with $B_j(x) = \{j, A_j(x)\}$ and $V\{B_k(x)\} = 1$ to ensure that $\sum_{j=1}^{k} \pi_{B_j}(x) = 1$.

2.2. Theoretical properties

3. Estimation

The proposed approach is based on a two-stage algorithm where first the predictor space is recursively divided in different subsets using an efficient partitioning algorithm and then, considering the partition as fixed, a multiresolution stick-breaking process is estimated. The predictor space is partitioned applying metis (Karypis & Kumar, 1999), a fast multiscale technique used for graph partitioning. Though more complicated densities can be considered, dictionary densities f_{B_j} will be estimated by assuming a normal form. In particular, densities corresponding to a particular partition set will be estimated considering observations for all subjects having predictors in that partition set. To be specific, for estimating density $f_{B_j}(y)$, we use the data $\{y_i : x_i \in \mathcal{X}^j_{A_j}\}$. We then conduct the analysis treating the dictionary elements and partition sets as fixed and placing a prior on the weights π_{B_j}.

Parameters involved in the dictionary density can be estimated using either frequentist or bayesian methods. Both methodologies have advantages and disadvantages. The frequentist approach, relying on maximum likelihood estimation, would allow to obtain parameter estimates in a faster and more efficient way. On the other hand, bayesian methods can avoid singularities associated with traditional maximum likelihood inference. Both methodologies will be implemented and compared for all data examples considered.

In any case, inference on stick breaking weights will be carried out using the Gibbs sampler. For this purpose, introduce the latent variable $S_i \in \{1, \ldots, k\}$, for $i = 1, \ldots, n$, denoting the multiscale level used by the ith subject. Let $n(B_j)$ be the number of observations allocated to node B_j. Then, at each gibbs sampling iteration, stick breaking are sampled as follows

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

$$\Pr(S_i = j \,|\, -) = \frac{\pi_{B_j(x_i)} f_{B_j(x_i)}(y_i)}{\sum_{h=1}^{k} \pi_{B_h(x_i)} f_{B_h(x_i)}(y_i)}$$

2. Update stick-breaking random variable $V_{B_j(x_i)}$, for $j = 1, \ldots, k$ and $i = 1, \ldots, n$, from $Beta(a_p, b_p)$ with $a_p = 1 + n\{B_j(x_i)\}$ and $b_p = \alpha + \sum_{B_h(x_i) \in ch\{B_j(x_i)\}} n\{B_h(x_i)\}$.

4. Simulation Studies

5. Real data examples

References

Bishop, C.M. and Svensen, M. Bayesian Hierarchical mixtures of experts. *Nineteenth Conference on Uncertainty in Artificial intelligence*, pp. 57–64, 2003.

Breiman, L. Bagging predictors. *Machine Learning*, 24:123140, 1996.

Breiman, L. Random Forests. *Machine Learning*, 45: 5–32, 2001.

Breiman, L., Friedman, J., Stone, C. J., and Olshen, R. A. *Classification and regression trees*. Chapman & Hall/CRC, 1984.

Friedman, J. H. Multivariate adaptive regression trees. *The annals of statistics*, 19:1–141, 1991.

Jacobs, R. A., Jordan, M. I., Nowlan, S. J., and Hinton, G. E. Adaptive mixture of local experts. *Neural Computation*, 3:79–87, 1991.

Jordan, M. I. and Jacobs, R. A. Hierarchical mixtures of experts and the EM algorithm. *Neural Computation*, 6:181–214, 1994.

Karypis, G. and Kumar, V. A fast and high quality multilevel scheme for partitioning irregular graphs. *SIAM Journal on Scientific Computing 20*, 1:359392, 1999.

Meeds, E. and Osindero, S. Bayesian Hierarchical mixtures of experts. *Advances in Neural Information Processing Systems*, 2006.

Mossavat, I. and Amft, O. Sparse bayesian hierarchical mixture of experts. *IEEE Statistical Signal Processing Workshop (SSP)*, 2011.

Rasmussen, C. E. and Ghahramani, Z. Infinite mixtures of Gaussian process experts. *Advances in neural information processing systems 14*, 2002.

Sethuraman, J. A Constructive Denition of Dirichlet Priors. *Statistica Sinica*, 4:639–650, 1994.

Shapire, R., Freund, Y., Bartlett, P., and Lee, W. Boosting the margin: A new explanation for the effectiveness of voting methods. *Annals of Statistics*, 26:16511686, 1998.