Information Sciences 292 (2015) 39–56

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

K-plane regression

Naresh Manwani [a,*], P.S. Sastry [b]

[a] GE Global Research, John F. Welch Technology Centre, # 122, EPIP Phase 2, Whitefield Road, Bangalore 560066, India
[b] Department of Electrical Engineering, Indian Institute of Science, Bangalore 560012, India

ARTICLE INFO

Article history:
Received 7 June 2013
Received in revised form 22 August 2014
Accepted 29 August 2014
Available online 8 September 2014

Keywords:
Piecewise linear regression
Cluster-wise linear regression
Expectation maximization
Mixture of experts

ABSTRACT

In this paper, we present a novel algorithm for piecewise linear regression which can learn continuous as well as discontinuous piecewise linear functions. The main idea is to repeatedly partition the data and learn a linear model in each partition. The proposed algorithm is similar in spirit to *k*-means clustering algorithm. We show that our algorithm can also be viewed as a special case of an EM algorithm for maximum likelihood estimation under a reasonable probability model. We empirically demonstrate the effectiveness of our approach by comparing its performance with that of the state of art algorithms on various datasets.

1. Introduction

In a regression problem, one is provided with training data consisting of feature vectors and the corresponding real-valued target outputs and the goal is to learn a function that captures the relationship between the feature vectors and the targets. Regression function learning has applications in many real world problems (e.g. wind speed prediction [2], business cycle forecasting [18], revenue forecasting [23,30], etc.).

Least squares regression and support vector regression are the two well-known and generic approaches [9,16,27,33,21,24]. In the least squares approach, nonlinear regression functions can be learnt using a user-specified fixed nonlinear mapping of feature vectors from the original space to some suitable high dimensional space. However, this could turn out to be computationally expensive and can also result in over-fitting. Learning rates and generalization error of regularized least squares regression are well studied [31]. In support vector regression (SVR), nonlinear models are learnt by using the kernel trick. SVR variants have been used for online learning of regression functions also [32]. SVR has a large margin flavor and has well studied performance guarantees. However, in general, for nonlinear problems, SVR solution is not easy to visualize in the original feature space.

A different approach to learning a nonlinear regression function is to approximate the target function by a piecewise linear function. The piecewise linear approach provides better understanding of the behavior of the regression surface in the original feature space as compared to the kernel-based approach of SVR. In piecewise linear approaches, the feature space is partitioned into disjoint regions and for every partition a linear regression function is learnt. The goal here is to simultaneously estimate the optimal partitions and the linear model for each partition. This problem is computationally intractable [25].

In this paper, we present a novel method of learning piecewise linear regression functions. In contrast to all the existing methods, our approach is capable of learning discontinuous functions. We show, through empirical studies, that this algorithm is attractive in comparison to the SVR approach as well as the hinge hyperplanes method [10].

* Corresponding author.
E-mail addresses: naresh.manwani@ge.com (N. Manwani), sastry@ee.iisc.ernet.in (P.S. Sastry).

http://dx.doi.org/10.1016/j.ins.2014.08.058
0020-0255/

Existing approaches for learning piecewise linear regression functions can be broadly classified into two categories. In the first set of approaches one assumes a fixed structure for the target function where as in the second set of approaches the form of the regression function is not fixed *a priori*.

In the fixed structure approaches we search over a parametrized family of piecewise linear regression functions and the parameters are learnt by solving an optimization problem to, typically, minimize the sum of the squared errors. Some examples of such methods are mixture of experts and hierarchical mixture of experts [19,29,22].

Among the approaches where no fixed structure is assumed, *regression tree* [11,20] is the most widely used method. A regression tree is built by binary or multivariate recursive partitioning of the feature space in a greedy fashion. Regression trees split the feature space at every node in such a way that fitting a linear regression function to each child node will minimize the sum of squared errors. In contrast to a decision tree where leaf nodes are assigned class labels, leaf nodes in a regression tree are associated with linear regression models. Most of the algorithms for learning regression trees are greedy in nature. At any node of the tree, once a hyperplane is learnt to split the feature space, it cannot be altered subsequently. The greedy nature of the method can lead to a suboptimal tree.

A more refined regression tree approach is the *hinging hyperplanes* method [10,26] which overcomes several drawbacks of regression tree approach. A hinge function is the simplest piecewise linear function defined as maximum or minimum of two affine functions [10]. In the hinging hyperplanes approach, any regression function is approximated as a sum of such hinge functions where the number of hinge functions is not fixed *a priori*. The algorithm starts with fitting a single hinge function on the training data using the hinge finding algorithm [10]. Then, the residual error is calculated for every example and, based on this, a new hinge function may be added to the model. Every time a new hinge function is added, its parameters are found by fitting the residual error. This algorithm overcomes the greedy nature of the regression tree approach by providing a mechanism for re-estimation of the parameters of each of the earlier hinge functions whenever a new hinge is added. Overall, the hinge hyperplanes algorithm tries to learn an optimal regression tree, given the training data. In a generalization of this idea, an approach is proposed in [4] to fit piecewise linear functions represented as maximum or minimum of linear functions.

A different greedy approach for piecewise linear regression is the *bounded error approach* [1,7,8]. In the bounded error approaches, for a given bound $\epsilon > 0$, on the tolerable error, a piecewise linear regression function is learnt such that for every point in the training set, the absolute difference between the target value and the predicted value is less than ϵ. This property is called the bounded error property. Greedy heuristic algorithms [7,8] have been proposed to find such a piecewise linear function. These algorithms start with finding a linear regression function which should satisfy the bounded error property for as many points in the training set as possible. This problem is known as the *maximum feasible sub-system problem* (MAX-FS) and is shown to be NP-hard [1]. MAX-FS problem is repeated on the remaining points until all points are exhausted. At present, there are no theoretical results to characterize the quality of the solution given by the model.

Most of the existing approaches for learning regression functions find a continuous approximation for the regression surface even if the actual surface is discontinuous. In this paper, we present an algorithm which is able to learn both continuous as well as discontinuous functions.

We start with a simple algorithm that is similar, in spirit, to the *k*-means clustering algorithm. The idea is to repeatedly keep partitioning the training data and learning a hyperplane for each partition. In each such iteration, after learning the hyperplanes, we repartition the training data so that all feature vectors in a partition have least prediction error with the hyperplane of that partition. We call this the *simple K-plane regression* algorithm. This approach is also known as clusterwise linear regression (CLR) in the literature [28,14,5,12,15]. A stochastic version of CLR is discussed under the title *mixtures of linear regression* models [9, Chapter 14].

Simple *K*-plane regression algorithm (or CLR) is attractive because it is conceptually very simple. However, it suffers from some serious drawbacks in terms of convergence to non-optimal solutions, sensitivity to additive noise and lack of a model function. Based on these insights, we propose a new and modified *K*-plane regression algorithm. In the modified algorithm also, we keep repeatedly partitioning the data and learning a linear model for each partition. However, we try to separately and simultaneously learn the centers of the partitions and the corresponding linear models. Through empirical studies we show that this algorithm is very effective for learning piecewise linear regression surfaces and it compares favorably with other state-of-art regression function learning methods.

The rest of the paper is organized as follows. In Section 2 we discuss the simple *K*-plane regression algorithm, its drawbacks and possible reasons behind them. We then propose the modified *K*-plane regression algorithm in Section 3. We show that the modified *K*-plane regression algorithm monotonically decreases the error function after every iteration. In Section 4 we show the equivalence of our algorithm with a special case of an expectation maximization (EM) algorithm. Experimental results are given in Section 5. We conclude the paper in Section 6.

2. Simple *K*-plane regression

We begin by defining a *K*-piecewise affine function. We use the notation that a hyperplane in $\mathbb{R}^d$ is parametrized by $(\mathbf{w}, b)$ where $\mathbf{w} \in \mathbb{R}^d$ and $b \in \mathbb{R}$. We sometimes denote this parameter vector by $\tilde{\mathbf{w}} = [\mathbf{w}^T \ b]^T \in \mathbb{R}^{d+1}$.

Definition 1. A function $f : \mathbb{R}^d \rightarrow \mathbb{R}$, is called ***K*-piecewise affine** if there exists a set of K hyperplanes with parameters $(\mathbf{w}_1, b_1), \ldots, (\mathbf{w}_K, b_K)$ $((\mathbf{w}_i, b_i) \neq (\mathbf{w}_j, b_j), \ \forall i \neq j)$, and sets $\widetilde{S}_1, \ldots, \widetilde{S}_K \subset \mathbb{R}^d$, which form a partition of $\mathbb{R}^d$, such that, $f(\mathbf{x}) = \mathbf{w}_k^T \mathbf{x} + b_k, \forall \mathbf{x} \in \widetilde{S}_k, \ k = 1, \ldots K$.