Computational Statistics and Data Analysis 54 (2010) 1707–1718

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Fast kernel conditional density estimation: A dual-tree Monte Carlo approach

Michael P. Holmes *, Alexander G. Gray, Charles Lee Isbell Jr.

College of Computing, Georgia Institute of Technology, Atlanta, GA 30332, USA

ARTICLE INFO

Article history:
Received 26 February 2009
Received in revised form 7 January 2010
Accepted 8 January 2010
Available online 22 January 2010

Keywords:
Kernel conditional density estimation
Fast algorithms
Large datasets
Scalability
Dual-tree
Monte Carlo

ABSTRACT

We describe a fast, data-driven bandwidth selection procedure for kernel conditional density estimation (KCDE). Specifically, we give a Monte Carlo dual-tree algorithm for efficient, error-controlled approximation of a cross-validated likelihood objective. While exact evaluation of this objective has an unscalable $O(n^2)$ computational cost, our method is practical and shows speedup factors as high as 286,000 when applied to real multivariate datasets containing up to one million points. In absolute terms, computation times are reduced from months to minutes. This enables applications at much greater scale than previously possible. The core idea in our method is to first derive a standard deterministic dual-tree approximation, whose loose deterministic bounds we then replace with tight, probabilistic Monte Carlo bounds. The resulting Monte Carlo dual-tree algorithm exhibits strong error control and high speedup across a broad range of datasets several orders of magnitude greater in size than those reported in previous work. The cost of this high acceleration is the loss of the formal error guarantee of the deterministic dual-tree framework; however, our experiments show that error is still amply controlled by our Monte Carlo algorithm, and the many-order-of-magnitude speedups are worth this sacrifice in the large-data case, where cross-validated bandwidth selection for KCDE would otherwise be impractical.

1. Introduction

Conditional density estimation models the probability density $f(y|\mathbf{x})$ of a random variable y given a random vector $\mathbf{x}$. For example, in Fig. 1 each contour line perpendicular to the x axis represents a conditional density. This can be viewed as a generalization of regression: in regression we estimate the expectation $E[y|\mathbf{x}]$, while in conditional density estimation we model the full distribution. Fig. 1 illustrates a conditional bimodality such that $E[y|\mathbf{x}]$ is insufficiently descriptive for many tasks. Estimating conditional densities is much harder than regression, but having the full distribution is powerful because it allows one to extract almost any quantities of interest, including expectations, modes, prediction intervals, outlier boundaries, samples, expectations of non-linear functions of y, etc. Conditional densities also facilitate data visualization and exploration. Conditional density estimates are of fundamental utility, applicable to such problems as time series prediction, static regression with prediction intervals, learning in Bayes nets and other graphical models, and so on. The estimation problem is challenging, however, because the data from which $f(y|\mathbf{x})$ must be learned generally do not include any exact $\mathbf{x}$ for which $f(y|\mathbf{x})$ will be queried.

Nonparametric kernel techniques address this issue by interpolating between the points that have been seen, without strong assumptions on distributional forms. In nonparametric conditional density estimation, we make only minimal

* Corresponding author.
E-mail addresses: mpholmes@gmail.com (M.P. Holmes), agray@cc.gatech.edu (A.G. Gray), isbell@cc.gatech.edu (C.L. Isbell Jr.).

0167-9473/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2010.01.011

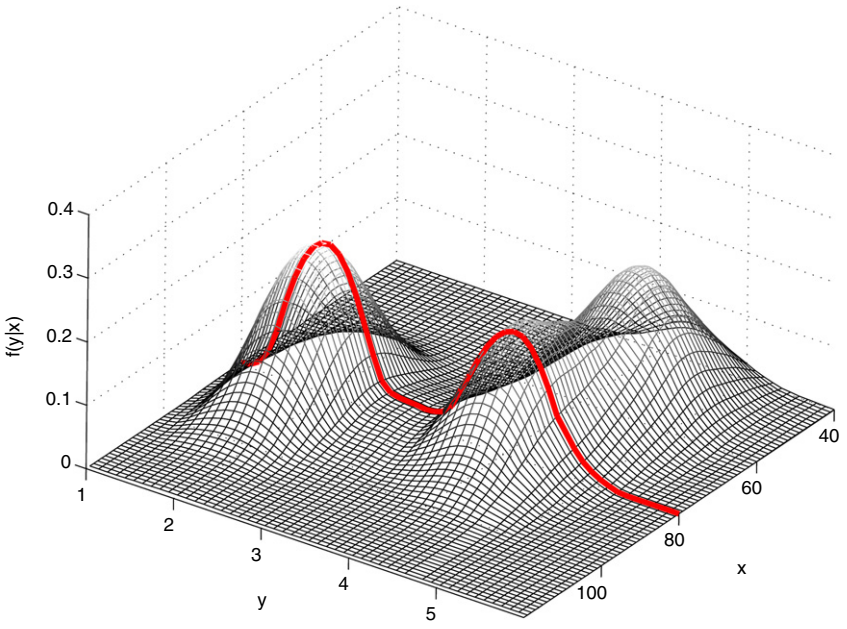

Fig. 1. Distribution $f(y, x)$ for which $f(y|x)$ can be either bimodal or unimodal, depending on x. The bold curve represents $f(y|x = 80)$.

assumptions about the smoothness of $f(y|\mathbf{x})$ without assuming any parametric form. Freedom from parametric assumptions is very often desirable when dealing with complex data, as we rarely have knowledge of true distributional structure. Nonparametric conditional density estimation has received some attention from statisticians and econometrics researchers (Gooijer and Zerom, 2003; Fan and Yim, 2004; Hansen, 2004; Bashtannyk and Hyndman, 2001; Hyndman et al., 1996; Rosenblatt, 1969), though relatively little when compared to nonparametric regression. Perhaps the main obstacle to wider adoption has been its computational cost, which is the problem addressed in this paper. Note that what we mean by nonparametric conditional density estimation is different from other techniques with similar names, such as conditional probability estimation (which refers to outputting class probabilities in the classification setting, also referred to as class-conditional probabilities).

In the present work, we consider the standard kernel conditional density estimator that first received serious attention in the work of Fan et al. (1996) and Hyndman et al. (1996), though it was originally proposed by Rosenblatt (1969). This is a direct kernel estimator of conditional densities, as opposed to approaches that separately estimate $f(y, \mathbf{x})$ and $f(\mathbf{x})$, which are combined to estimate $f(y|\mathbf{x}) = f(y, \mathbf{x})/f(\mathbf{x})$ (see Stender, 2006). Direct estimation of conditional densities allows parameter estimation to be formulated as the optimization of a single, unified objective function, whereas separate estimation of $f(y, \mathbf{x})$ and $f(\mathbf{x})$ optimizes two different objective functions that may not produce the highest-quality *conditional* densities.

Although the direct estimator we use is consistent given mild conditions on its bandwidths, practical use has been hampered by the lack of an efficient data-driven bandwidth selection procedure, upon which any kernel estimator depends critically. We propose a new method for efficiently selecting bandwidths to maximize cross-validated likelihood, an objective with some advantages over the squared-error criteria used in prior work. The speedup of this method is obtained by combining Monte Carlo techniques with a dual-tree-based approximation (see Gray and Moore, 2000) of the likelihood function. This approximation approach belongs to a new class of multi-tree Monte Carlo methods (Holmes, 2009). We present two versions of fast likelihood approximation, one analogous to previous dual-tree algorithms with deterministic error control, which gives speedups on the order of 1.5–10-fold in our experiments, and the other with a new, probabilistic Monte Carlo error control mechanism, which gives much larger speedups—as high as 286,000-fold on one million points.

With this fast learning procedure we can address datasets that are both higher in dimension and several orders of magnitude larger in size than those reported in previous work, which has been confined to bivariate datasets of size no greater than 1000 (Fan and Yim, 2004). We present results that validate the accuracy and speedup of our likelihood approximation on real datasets possessing a variety of sizes and dimensionalities. Most of these datasets were previously impractical to address with naively computed data-driven techniques. Thus, our fast bandwidth optimization method enables applications at scales that were previously unreachable. We conclude that kernel conditional density estimation is a powerful technique that is made substantially more efficient by our fast approximate optimization procedure, with many opportunities for application in a variety of statistical fields.

In the remainder of the paper we first describe the standard kernel conditional density estimator; this is followed by a discussion of the bandwidth selection problem and our choice of likelihood cross-validation as a bandwidth selection objective, at which point we derive our dual-tree approximation algorithms (both deterministic and Monte Carlo), show experimental performance, and conclude with a summary of results and implications.