Fast nonparametric conditional density estimation

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

College of Computing
Georgia Institute of Technology
Atlanta, GA 30332-0760
{mph,agray,isbell}@cc.gatech.edu

Conditional density estimation. The idea of conditional density estimation is to construct a density estimate $\hat{f}(y|x)$ for a dependent variable y, conditional on a vector of variables x. This can be seen as a generalization of regression, where instead of estimating the expected value $E(y|x)$ alone, we instead model the full density. This is especially important for multi-modal densities, where the expected value might be nowhere near a mode, and for situations in which confidence intervals are preferred to point estimates. Some problems that can be addressed by conditional density estimates are: time series prediction, static regression with confidence bands, learning continuous k-Markov models, and collaborative filtering.

Nonparametric conditional density estimators address the common situation where $f(y|x)$ has no known parametric form. Though widely applicable, this class of estimators has received relatively little attention in the statistics community and little or none in the machine learning community. Following the lead of the statisticians [1, 2, 3, 4, 5], our approach is to use a double kernel estimator of the general form

$$\hat{f}(y|x) = \frac{\sum_i W_{h_2}(||x - X_i||)K_{h_1}(y - Y_i)}{\sum_j W_{h_2}(||x - X_j||)},$$

where K is a simple kernel function on y and W is a weight function on x that can be more complicated than a simple kernel. This form is important in continuous spaces where a given value of the vector x is not likely to be observed more than once. Naively, since we never get a sample set (Y_i, x), this would make estimation of $f(y|x)$ impossible; however, this can be overcome by allowing all observed values X_i to contribute in conditionally weighted fashion to any value x for which $f(y|x)$ is queried. Furthermore, by constructing estimates only for univariate y, we can significantly reduce the explosive data requirements incurred by attempting to model $f(y|x)$ with the multivariate densities in $f(x, y)/f(x)$ [6].

Theoretical and algorithmic contributions. We introduce an extended version of the double kernel estimator that constructs the weights W on x using multivariate locally linear smoothing (previous locally linear smoothers appear to have been limited to the simpler case of univariate x [1, 4]). The locally linear factor is important for added accuracy over the basic locally constant method, especially when extrapolating at the edges of the data. We use two bandwidth selection methods, one that optimizes over a standard cross-validated estimate of the integrated squared error (ISE), and the other optimizing a cross-validated likelihood. The first has better robustness properties, but the latter is faster to compute and does not appear to have been previously used for conditional density bandwidths. Direct computation of these quantities is $O(N^3)$ for ISE and $O(N^2)$ for likelihood, which is probably the reason that applications of previous work appear to have been confined to small, bivariate datasets (i.e. univariate x and univariate y) [1, 2, 3, 4, 5]. We render the cross-validation tractable through a new multi-tree-based fast approximation algorithm [7], making it possible to handle datasets of greater dimensionality.

The contributions of this work are: locally linear smoothing for the case of multivariate conditioning vectors x, first application of maximum likelihood for conditional bandwidth selection, fast algorithms for data-driven bandwidth selection using both ISE and likelihood criteria, and the first application to datasets with multivariate conditioning.

Applications. We present results from applying kernel conditional density estimation to several problems: some synthetic datasets that demonstrate good performance where the answer is known; a dataset of geyser eruption lengths for comparison with previous work; a dataset from Sloan Digital Sky Survey, in which we estimate confidence intervals for distances to various astronomical objects involved in mapping the large-scale structure of the universe; and a time series problem in which we forecast bookings of flights and train itineraries, with data coming from an industrial price management system that uses such forecasts in a larger optimization framework.

References

[1] Jianqing Fan and Tsz Ho Yim. A crossvalidation method for estimating conditional densities. *Biometrika*, 91(4):819–834, 2004.

[2] Jan G. De Gooijer and Dawit Zerom. On conditional density estimation. *Statistica Neerlandica*, 57(2):159–176, May 2003.

[3] David M. Bashtannyk and Rob J. Hyndman. Bandwidth selection for kernel conditional density estimation. *Computational Statistics & Data Analysis*, 36(3):279–298, May 2001.

[4] Bruce E. Hansen. Nonparametric estimation of smooth conditional densities. Unpublished manuscript, May 2004.

[5] Bruce E. Hansen. Nonparametric conditional density estimation. Unpublished manuscript, November 2004.

[6] B. W. Silverman. *Density Estimation for Statistics and Data Analysis*, volume 26 of *Monographs on Statistics and Probability.* Chapman & Hall, 1986.

[7] Alexander G. Gray and Andrew W. Moore. N-body problems in statistical learning. In *Advances in Neural Information Processing Systems (NIPS) 13*, 2000.

Topic: estimation, prediction, and sequence modeling
Preference: oral/poster