Computational Statistics and Data Analysis 79 (2014) 149–164

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Random average shifted histograms

M. Bourel [a,c], R. Fraiman [b], B. Ghattas [c,*]

[a] IMERL, Facultad de Ingenieria, Universidad de la República, Montevideo, Uruguay
[b] CMAT, Facultad de Ciencias, Universidad de la República, Montevideo, Uruguay
[c] Université d'Aix Marseille, Institut de Mathématiques de Marseille, Marseille, France

ARTICLE INFO

Article history:
Received 27 August 2013
Received in revised form 7 March 2014
Accepted 7 May 2014
Available online 27 May 2014

Keywords:
Machine learning
Histogram
Density estimation
Bootstrap
Bagging
Average shifted histograms

ABSTRACT

A new density estimator called *RASH*, for *Random Average Shifted Histogram*, obtained by averaging several histograms as proposed in *average shifted histograms*, is presented. The principal difference between the two methods is that in RASH each histogram is built over a grid with random shifted breakpoints. The asymptotic behavior of this estimator is established for the one-dimensional case and its performance through several simulations is analyzed. *RASH* is compared to several classic density estimators and to some recent ensemble methods. Although *RASH* does not always outperform the other methods, it is very simple to implement, being also more intuitive. The two dimensional case is also analyzed empirically.

1. Introduction

There is no doubt that, in regression and classification, ensemble learning, which consists on combining several models, gives rise to more complex models that largely outperform the classical simple methods. Algorithms like Bagging (Breiman, 1996a), Boosting (Freund and Schapire, 1997), Stacking (Breiman, 1996b; Wolpert, 1992) and Random Forests (Breiman, 2001) have been deeply studied both from the standpoint of theory and that of applications and they have evolved into many variants achieving very high performances when tested over tens of different data sets from the machine learning benchmark. These algorithms have been designed for supervised learning, initially restricted to regression or binary classification. Several extensions are actually under study: multivariate regression and multi-class learning, among others.

Nevertheless, there exist very few extensions of ensemble methods for unsupervised learning such as clustering analysis or density estimation. In this work we present a contribution to this last case, which is an important problem in statistics. Some extensions of Boosting (Di Marzio and Taylor, 2004), Bagging (Ridgeway, 2002; Rosset and Segal, 2002) and Stacking (Smyth and Wolpert, 1999) to density estimation have been already considered. Other approaches inspired from Bagging and Stacking have also been studied empirically (Bourel and Ghattas, 2013).

We suggest a new simple algorithm, *Random Average Shifted Histogram* (*RASH*) for density estimation aggregating histograms which are "weak learners" in this context. Our idea arises from the average shifted histogram introduced by Scott (1992): to avoid the problem of the histogram's origin choice, this method averages several histograms built using different shifts of the breakpoints over a fixed grid. The main difference is that in *RASH* the breakpoints are randomly shifted. We introduce thus a random breakpoints histogram (RH-estimate) and study its asymptotic properties. Then we build up our aggregation estimate by averaging M RH-estimates. We show by extensive simulations that this kind of aggregation gives

* Corresponding author.
E-mail addresses: mbourel@fing.edu.uy (M. Bourel), rfraiman@cmat.edu.uy (R. Fraiman), badih.ghattas@univ-amu.fr (B. Ghattas).

http://dx.doi.org/10.1016/j.csda.2014.05.004
0167-9473/

rise to better estimates. We compare our algorithm to Average Shifted Histogram (*ASH*) and to several classic algorithms used in the literature. Besides being simple, our approach seems to be more intuitive and shows very high accuracy.

Section 2 gives a brief description of the histogram and its main asymptotic properties. Our algorithm is presented in Section 3. In Section 4 we provide asymptotic results for the RH-estimate in the one-dimensional case: consistency, asymptotic normality and rates of convergence. Section 5 describes the simulation study, where we compare our proposal with several competitors and provide an extension to the two dimensional configurations. All proofs are given in the Appendix.

2. Some density estimators

We start fixing some notations and describing the algorithms we will compare with *RASH*. We also recall some important results about the histogram, kernel density estimator, average shifted histogram, and an aggregated model selection introduced in Samarov and Tsybakov (2007).

2.1. Histogram

We consider, for an i.i.d. sample $X_1, \ldots, X_n$ of random variables with density f, L_n intervals $I_{1,n} \ldots, I_{L_n,n}$ where $|I_{j,n}| = \left|\left[a_j(n), a_{j+1}(n)\right)\right| = h_n$ for all n, and $h_n \downarrow 0$ as $n \to +\infty$.

The ordinary histogram (*Hist*) is defined as:

$$\widehat{f}_{n,0}(x) = \frac{1}{nh_n} \sum_{i=1}^{n} \sum_{j=1}^{L} \mathbb{1}_{I_{j,n}}(X_i) \mathbb{1}_{I_{j,n}}(x).$$

If $x \in I_{j,n}$, we have that:

$$\mathbb{E}\left(\widehat{f}_{n,0}(x)\right) = \frac{1}{nh_n} \sum_{i=1}^{n} \sum_{j=1}^{L} \mathbb{P}(X_i \in I_{j,n}) \mathbb{1}_{I_{j,n}}(x) = \frac{1}{nh_n} n\mathbb{P}(X_i \in I_{j,n}) = \frac{1}{|I_{j,n}|} \int_{I_{j,n}} f(t)dt$$

$$Var\left(\widehat{f}_{n,0}(x)\right) = \frac{1}{n^2 h_n^2} nVar\left(\mathbb{1}_{I_{j,n}}(X_i)\right) \leq \frac{1}{nh_n^2} \mathbb{P}(X_1 \in I_{j,n}) = \frac{1}{nh_n} \frac{\mathbb{P}(X_1 \in I_{j,n})}{h_n}.$$

When $h_n \to 0$ and $nh_n \to \infty$ we get the classical properties for the histogram:

$$\mathbb{E}\left(\widehat{f}_{n,0}(x)\right) \to f(x), \qquad Var\left(\widehat{f}_{n,0}(x)\right) \to 0.$$

The histogram depends on two parameters: the bin width h_n and the origin x_0. There is a huge literature that proposes several optimal choices for h_n. If we suppose that the underlying density f is Gaussian, it can be shown (see Scott, 1979) that an optimal choice for h is:

$$h_{\text{opt}} = 3.5\widehat{\sigma} n^{-1/3},$$

where $\widehat{\sigma}$ is an estimate of the standard deviation.

2.2. Average shifted histogram

The histogram estimate may change significantly when the origin x_0 changes, even if h_n is fixed. Scott (1985), introduced the Average Shifted Histogram (*ASH*) algorithm which aims to avoid choosing x_0. It is a nonparametric density estimator which averages several histograms with different origins. Consider for example an histogram with origin x_0, bin width h, and support $\left\{[jh, (j+1)h)\right\}_{j\in\mathbb{Z}}$. For $M > 0$, let $\delta = \frac{h}{M}$ and divide each interval $\left[jh, (j+1)h\right)$ into M new subintervals $B_k = \left[k\delta, (k+1)\delta\right)$ obtaining a finer grid. For instance, if $k = 0$, we divide $[0, h)$ into M intervals $B_0 = [0, \delta), B_1 = [\delta, 2\delta), \ldots, B_{M-1} = \left[(M-1)\delta, M\delta\right) = \left[(M-1)\delta, h\right)$. Let ν_k be the number of observations falling in B_k and $x \in B_0 = [0, \delta)$. Then there are M "shifted" histograms with bin width $h = M\delta$ which cover B_0. The value of the first one at x is:

$$\widehat{f}_1(x) = \frac{\nu_{1-M} + \nu_{2-M} + \cdots + \nu_0}{nh}.$$

The value of the second one at x is

$$\widehat{f}_2(x) = \frac{\nu_{2-M} + \nu_{3-M} + \cdots + \nu_0 + \nu_1}{nh}.$$

The Mth final shifted histogram which covers $[0, \delta)$ takes at x the value:

$$\widehat{f}_M(x) = \frac{\nu_0 + \cdots + \nu_{M-1}}{nh}.$$