Computational Statistics and Data Analysis 77 (2014) 371–382

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Asymmetric ν-tube support vector regression

Xiaolin Huang[a,*], Lei Shi[a,c], Kristiaan Pelckmans[b], Johan A.K. Suykens[a]

[a] KU Leuven, Department of Electrical Engineering, ESAT-STADIUS, B-3001, Leuven, Belgium
[b] Department of Information Technology, Uppsala University, SE-751 05, Uppsala, Sweden
[c] School of Mathematical Sciences, Fudan University, 200433, Shanghai, PR China

ARTICLE INFO

Article history:
Received 11 September 2013
Received in revised form 27 March 2014
Accepted 27 March 2014
Available online 1 April 2014

Keywords:
Robust regression
ν-tube support vector regression
Asymmetric loss
Quantile regression

ABSTRACT

Finding a tube of small width that covers a certain percentage of the training data samples is a robust way to estimate a location: the values of the data samples falling outside the tube have no direct influence on the estimate. The well-known ν-tube Support Vector Regression (ν-SVR) is an effective method for implementing this idea in the context of covariates. However, the ν-SVR considers only one possible location of this tube: it imposes that the amount of data samples above and below the tube are equal. The method is generalized such that those outliers can be divided asymmetrically over both regions. This extension gives an effective way to deal with skewed noise in regression problems. Numerical experiments illustrate the computational efficacy of this extension to the ν-SVR.

1. Introduction

Since its introduction by Schölkopf et al. (2000), the ν-tube Support Vector Regression (ν-SVR) has become a standard tool in nonparametric regression tasks. The ν-SVRs extend standard Support Vector Regression techniques given by Vapnik (1995) via (i) enforcing a fraction of the data samples to lie inside a tube, as well as (ii) minimizing the width of this tube. Mathematically, training such a tube $[f(x) - \varepsilon, f(x) + \varepsilon]$ can be formulated as the following optimization problem,

$$\begin{aligned} \min_{f,\varepsilon} \quad & \varepsilon \\ \text{s.t.} \quad & \sum_{i=1}^{n} \mathcal{I}\left(y_i \in [f(x_i) - \varepsilon, f(x_i) + \varepsilon]\right) \geq \rho n, \end{aligned} \tag{1}$$

where $\{(x_i, y_i)\}_{i=1}^n$, $x_i \in \mathbb{R}^d$, $y_i \in \mathbb{R}$ are the data samples, $0 \leq \rho \leq 1$ is a user defined constant, and $\mathcal{I}(a)$ stands for an indicator function, which equals one when a is true and equals zero otherwise. Unlike traditional point-regression methods, (1) focuses on estimation of the confidence region directly, which is called *support tube* by Pelckmans et al. (2009). One can find the corresponding statistical discussion therein.

Apparently, the values of the data samples falling outside the tube have no direct influence on the result of (1), which is quite robust to *outliers*, since the outliers probably fall outside. In fact, this idea has appeared in robust regression and is known as the least median squares regression, which is proposed by Rousseeuw (1984), Rousseeuw and Leroy (1987). Denote the kth maximum of $\{u_i\}_{i=1}^n$ by $\max^k_{1\leq i\leq n}\{u_i\}$:

$$\max\nolimits^k_{1\leq i\leq n}\{u_i\} = u_{\Gamma(k)} \quad \text{with } u_{\Gamma(1)} \geq u_{\Gamma(2)} \geq \cdots \geq u_{\Gamma(n)}.$$

* Correspondence to: ESAT-STADIUS, Kasteelpark Arenberg 10, Bus 2446, 3001 Heverlee, Belgium. Tel.: +32 16328653; fax: +32 16321970.
E-mail addresses: huangxl06@mails.tsinghua.edu.cn (X. Huang), leishi@fudan.edu.cn (L. Shi), kp@it.uu.se (K. Pelckmans), johan.suykens@esat.kuleuven.be (J.A.K. Suykens).

http://dx.doi.org/10.1016/j.csda.2014.03.016
0167-9473/

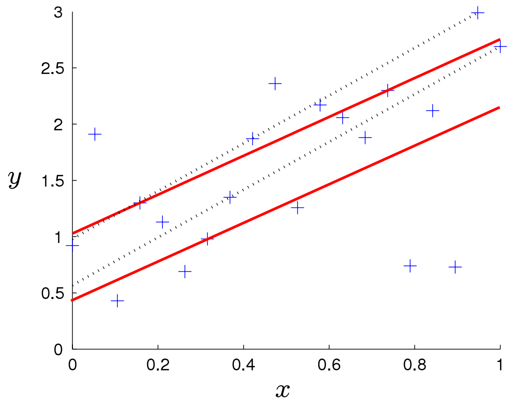

Fig. 1. We require 50% of the data (blue crosses) are covered by a tube. The result of the ν-SVR (3) is illustrated by red solid lines. Black dotted lines display the optimal solution to (1). The two tubes cover the same amount of the data samples, but the distributions of the outliers are different.

Then the least median squares estimator can be written as

$$\min_{f} \quad \max^{k}_{1\le i\le n} \left\{ (y_i - f(x_i))^2 \right\}. \tag{2}$$

One can observe the equivalence between (2) and (1) when $\rho = k/n$. If the median squared error is minimized, (2) is regarded as the most robust estimator in view of the breakdown point defined by Donoho and Huber (1982). The idea of minimizing the median error also has been discussed for classification tasks by Ma et al. (2011) and Tsyurmasto et al. (2013). For the least median squares regression, there have been some approximation algorithms proposed by Tichavsky (1991), Boček and Lachout (1995), Olson (1997), Verardi and Croux (2009), and Winker et al. (2011). The most popular method for computing the least median squares estimator is PROGRESS suggested by Rousseeuw and Leroy (1987) and modified by Rousseeuw and Hubert (1997). Besides, several algorithms have been developed for finding the global optimum, see the works of Steele and Steiger (1986) and Stromberg (1993).

Regression method (1) enjoys robustness to outliers, but it is non-convex. It can be modeled as a mixed integer linear programming (MILP), which has been proved to be NP-hard, see the discussion given by Huang et al. (2012). The solving time of an NP-hard problem is not acceptable for large-scale problems. Thus, we need a convex proxy for (1) for computational efficacy. Actually, the ν-SVR can be regarded as such a convex approximation, i.e., one can get a narrow tube, though not the optimal one, to cover ρ percentage samplings via solving a convex problem. Let us consider the case that f is chosen from the set of affine functions. Then the ν-SVR in primal space refers to the following optimization problem,

$$\begin{aligned} \min_{w,b,\varepsilon} \quad & \frac{1}{2\gamma} w^T w + \nu\varepsilon + \frac{1}{n}\sum_{i=1}^{n} L_\varepsilon \left(y_i - (w^T x_i + b) \right) \\ \text{s.t.} \quad & \varepsilon \ge 0, \end{aligned} \tag{3}$$

where $\nu \ge 0$ is a user defined parameter and $L_\varepsilon(u)$ is the ε insensitive zone loss as defined below:

$$L_\varepsilon(u) = \begin{cases} u - \varepsilon, & u \ge \varepsilon, \\ 0, & -\varepsilon < u < \varepsilon, \\ -u - \varepsilon, & u \le -\varepsilon. \end{cases}$$

It has been proved by Schölkopf et al. (2000) that the minimizer of (3) satisfies:

$$\sum_{i=1}^{n} \mathcal{I} \left(y_i \in [w^T x_i + b - \varepsilon, w^T x_i + b + \varepsilon] \right) \ge (1 - \nu)n,$$

which means that by setting $\nu = 1 - \rho$, the ν-SVR provides a feasible solution to (1). The difference between (1) and (3) can be observed by a linear regression task shown in Fig. 1. In this example, we pursue a tube covering half of the data samples, which are displayed by blue crosses. With a suitable ν, the ν-SVR (3) results in a good solution to (1). This tube is shown by red solid lines and covers 50% of the data. This solution is not yet the optimal to (1). Since this problem scale is small, we can get the optimal tube via solving the MILP formulation of (1) by iLog CPLEX. The optimal tube is shown by black dotted lines, which has the smallest width among all the tubes covering 50% of the data. One noticeable point is that there are 8 points above the optimal tube and 2 points below it. However, the amount of outliers above and below the red solid tube are imposed to be equal by (3), which reserves from the symmetry of L_ε.

Motivated by these observations, this paper extends L_ε to an asymmetric loss. Then an asymmetric ν-tube support vector regression (asymmetric ν-SVR) is established. By the proposed method, we can find an asymmetric tube, above and below which the outliers are distributed asymmetrically. The asymmetric tube is more flexible and can give a better solution to (1). This is especially suitable for dealing with asymmetric noise. Many applications have asymmetric noise. For example, when