Boosting Based Conditional Quantile Estimation for Regression and Binary Classification

Songfeng Zheng

Department of Mathematics
Missouri State University, Springfield, MO 65897, USA
SongfengZheng@MissouriState.edu

Abstract. We introduce Quantile Boost (QBoost) algorithms which predict conditional quantiles of the interested response for regression and binary classification. Quantile Boost Regression (QBR) performs gradient descent in functional space to minimize the objective function used by quantile regression (QReg). In the classification scenario, the class label is defined via a hidden variable, and the quantiles of the class label are estimated by fitting the corresponding quantiles of the hidden variable. An equivalent form of the definition of quantile is introduced, whose smoothed version is employed as the objective function, which is maximized by gradient ascent in functional space to get the Quantile Boost Classification (QBC) algorithm. Extensive experiments show that QBoost performs better than the original QReg and other alternatives for regression and classification. Furthermore, QBoost is more robust to noisy predictors.

Keywords: Boosting, Quantile Regression, Classification.

1 Introduction

Least square regression aims to estimate the conditional expectation of the response Y given the predictor (vector) $\mathbf{x}$, i.e., $\mathrm{E}(Y|\mathbf{x})$. However, the mean value (or the conditional expectation) is sensitive to the outliers of the data [12]. Therefore, if the data is not homogeneously distributed, we expect the least square regression gives us a poor prediction.

The τ-th quantile of a distribution is defined as the value such that there is $100\tau\%$ of mass on the left side of it. Compared to the mean value, quantiles are more robust to outliers [12]. For a random variable Y, it can be proved [11] that

$$Q_\tau(Y) = \arg\min_c \mathrm{E}_Y[\rho_\tau(Y-c)],$$

where $Q_\tau(Y)$ is the τ-th quantile of Y, $\rho_\tau(r)$ is the "check function" [12] defined by

$$\rho_\tau(r) = rI(r \geq 0) - (1-\tau)r, \tag{1}$$

where $I(\cdot) = 1$ if the condition is true, otherwise $I(\cdot) = 0$.

G. Sidorov et al. (Eds.): MICAI 2010, Part II, LNAI 6438, pp. 67–79, 2010.

Given data $\{(\mathbf{x}_i, Y_i), i = 1, \cdots, n\}$, with predictor $\mathbf{x}_i \in \mathbf{R}^d$ and response $Y_i \in \mathbf{R}$, let the τ-th conditional quantile of Y given $\mathbf{x}$ be $f(\mathbf{x})$. Similar to least square regression, quantile regression (QReg) [12] aims at estimating the conditional quantiles of the response given predictor vector $\mathbf{x}$ and can be summarized as

$$f^*(\cdot) = \arg\min_f \frac{1}{n} \sum_{i=1}^{n} \rho_\tau \left(Y_i - f(\mathbf{x}_i)\right), \tag{2}$$

which can be solved by linear programming algorithms [12] or MM algorithms [11]. However, when the predictor $\mathbf{x}$ is in high dimensional space, the aforementioned optimization methods for QReg might be inefficient. High dimension problems are ubiquitous in applications, e.g., image analysis, gene sequence analysis, to name a few. To the best of our knowledge, the problem of high dimensional predictor is not sufficiently addressed in quantile regression literature.

Motivated by the basic idea of gradient boosting algorithms [8], we propose to estimate the quantile regression function by minimizing the objective function in Eqn. (2) with functional gradient descent. In each step, we approximate the negative gradient of the objective function by a base function, and grow the model along that direction. This results Quantile Boost Regression (QBR) algorithm. In the binary classification scenario, we define the class label via a hidden variable, and the quantiles of the class label can be estimated by fitting the corresponding quantiles of the hidden variable. An equivalent form of the definition of quantile is introduced, whose smoothed version is employed as the objective function for classification. Similar to QBR, functional gradient ascent is applied to maximize the objective function, yielding the Quantile Boost Classification (QBC) algorithm. The obtained Quantile Boost (QBoost) algorithms are computationally efficient and converge to a local optimum, more importantly, they enable us to solve high dimensional problems efficiently.

The QBoost algorithms were tested extensively on publicly available datasets for regression and classification. On the regression experiments, QBR performs better than the original QReg in terms of check loss function. Moreover, the comparative experiment on noisy data indicates that QBR is more robust to noise. On classification problems, QBC was compared to binary QReg on a public dataset, the result shows that QBC performs better than binary QReg and is more robust to noisy predictors. On three high dimensional datasets from bioinformatics, binary QReg is not applicable due to its expensive computation, while QBC performs better than or similar to other alternatives in terms of 5 fold cross validation error rates. Furthermore, both QBC and QBR are able to select the most informative variables, inheriting the feature selection ability of boosting algorithm.

2 Boosting as Functional Gradient Descent

Boosting [7] is well known for its simplicity and good performance. The powerful feature selection mechanism of boosting makes it suitable to work in high