Modeling Quantiles

Claudia Perlich
IBM T.J. Watson Research, USA

Saharon Rosset
IBM T.J. Watson Research, USA

Bianca Zadrozny
Universidade Federal Fluminense, Brazil

INTRODUCTION

One standard Data Mining setting is defines by a set of n observations on a variable of interest Y and a set of p explanatory variables, or features, $x = (x_1,...,x_p)$, with the objective of finding a 'dependence' of Y on x. Such dependencies can either be of direct interest by themselves or used in the future to predict a Y given an observed x. This typically leads to a model for a conditional central tendency of $Y|x$, usually the mean $E(Y|x)$. For example, under appropriate model assumptions, Data Mining based on a least squares loss function (like linear least squares or most regression tree approaches), is as a maximum likelihood approach to estimating the conditional mean.

This chapter considers situations when the value of interest is not the conditional mean of a continuous variable, but rather a different property of the conditional distribution $P(Y|x)$, in particular a specific quantile of this distribution. Consider for instance the 0.9^{th} quantile of $P(Y|x)$, which is the function $c(x)$ such that $P(Y<c(x)|x) = 0.9$. As discussed in the main section, these problems (of estimating conditional mean vs. conditional high quantile) may be equivalent under simplistic assumptions about our models, but in practice they are usually not. We are typically interested in modeling extreme quantiles because they represent a desired 'prediction' in many business and scientific domains. Consider for example the motivating Data Mining task of estimating customer wallets from existing customer transaction data, which is of great practical interest for marketing and sales. A customer's wallet for a specific product category is the total amount this customer can spend in this product category. The vendor observes what the customers actually bought from him in the past, but does not typically have access to the customer's budget allocation decisions, their spending with competitors, etc. Information about customer's wallet, as an indicator of their potential for growth, is considered extremely valuable for marketing, resource planning and other tasks. For a detailed survey of the motivation, problem definition, see Rosset et al. 2005. In that paper we propose the definition of a customer's REALISTIC wallet as the 0.9^{th} or 0.95^{th} quantile of their conditional spending - this can be interpreted as the quantity that they may spend in the best case scenario. This task of modeling what a vendor can hope for rather than could expect turns out to be of great interest in multiple other business domains, including:

- When modeling sales prices of houses, cars or any other product, the seller may be very interested in the price they may aspire to get for their asset if they are successful in negotiations. This is clearly different from the 'average' price for this asset and is more in line with a high quantile of the price distribution of equivalent assets. Similarly, the buyer may be interested in the symmetric problem of modeling a low quantile.
- In outlier and fraud detection applications we may often have a specific variable (such as total amount spent on a credit card) whose degree of 'outlyingness' we want to examine for each one of a set of customers or observations. This degree can often be well approximated by the quantile of the conditional spending distribution given the customer's attributes. For identifying outliers we may just want to compare the actual spending to an appropriate high quantile, say 0.95.
- The opposite problem of the same notion of 'how bad can it get' is a very relevant component of financial modeling and in particular Value-at-Risk (Chernozhukov and Umantsev, 2001).

Addressing this task of quantile predictions, various researches have proposed methods that are often adaptations of standard expected value modeling approaches to the quantile modeling problem, and demonstrated that their predictions are meaningfully different from traditional expected value models.

BACKGROUND

Building and Evaluating Quantile Models

This section reviews some of the fundamental statistical and algorithmic concepts underlying the two main phases of predictive modeling - model building and model evaluation and selection - when the ultimate data mining goal is to predict high quantiles. Let us start from the easier question of model evaluation and model selection: given several models for predicting high quantiles and an evaluation data set not used for modeling, how can we estimate their performance and choose among them? The key to this problem is finding a loss function which describes well our success in predicting high quantile and evaluate the performance using this loss function. Clearly, the most important requirement from a loss function for evaluation is that the model which always predicts the conditional quantile correctly will have the best expected performance. Such a loss function indeed exists (Koenker, 2005).

Define the quantile loss function for the p^{th} quantile to be:

$$L_p(y,\hat{y}) = \begin{cases} p(y-\hat{y}) & \text{if } y \geq \hat{y} \\ (1-p)(\hat{y}-y) & \text{otherwise} \end{cases} \quad (1)$$

Figure 1 shows the quantile loss function for $p \in \{0.2, 0.5, 0.8\}$. With p=0.5 this is just absolute error loss. Expected quantile loss is minimized by correctly predicting the (conditional) p^{th} quantile of the conditional distribution. That is, if we fix a prediction point x, and define $c_p(x)$ to be the p^{th} quantile of the conditional distribution of Y given x:

$$P(Y \leq c_p(x) \mid x) = p, \forall x$$

then the loss is optimized in expectation at every point by correctly predicting $c_p(x)$:

$$\arg\min_c E(L_p(Y,c) \mid x) = c_p(x)$$

With p=0.5, the expected absolute loss is minimized by predicting the median, while when p=0.9 we are in

Figure 1. Quantile loss functions for some quantiles