A short note on quantile classifiers

Yen-Chi Chen
University of Washington

September 16, 2020

Quantile classification is a classification method using quantiles to form a classifier. It avoids the needs of making a parametric assumption on the distribution but can still achieve a fast convergence rate to the Bayes risk under appropriate assumptions. In this note, we will briefly describe the procedure of constructing a quantile classifier.

Consider a simple binary classification problem, where we observe a univariate covariate $X \in \mathbb{R}$ and a class label $Y \in \{0,1\}$. Let $p(x,y)$ be the joint distribution of (X,Y). Let $p(x|y)$ be the conditional PDF of X given $Y = y$ and let $F_y(x) = F(x|Y = y)$ be the CDF of X under class $Y = y$.

Let $\theta \in [0,1]$ be a quantile probability and let $q_y(\theta) = F_y^{-1}(\theta)$ be the quantile function of $F_y(x)$. Namely, $q_y(\theta)$ is the θ-th quantile of X given $Y = y$. We define a quantile discrepancy to the y-th group as

$$\begin{aligned}\eta_y(x;\theta) &= (\theta + (1-2\theta)I(x - q_y(\theta) > 0))|x - q_y(\theta)| \\ &= \begin{cases} \theta \cdot |x - q_y(\theta)|, & \text{if } z > q_y(\theta) \\ (1-\theta)\cdot|x - q_y(\theta)|, & \text{if } z \leq q_y(\theta) \end{cases}\end{aligned} \tag{1}$$

$\eta_y(x;\theta)$ measures how a value x deviates from the quantile value of y-the group.

One way to think about the quantity η_y is via the *quantile loss* function. A θ-quantile loss function is

$$L_\theta(a,b) = \begin{cases} \theta \cdot |a-b|, & \text{if } a > b \\ (1-\theta)\cdot|a-b|, & \text{if } a \leq b. \end{cases}$$

For a random variable Z, one can show that the minimizer $q^* = \mathsf{argmin}_q \mathbb{E}[L_\theta(Z,q)]$ will be the θ-quantile of Z. As a result, $q_y(\theta) = \mathsf{argmin}_q \mathbb{E}[L_\theta(X,q)|Y = y]$ so one can immediately see that

$$\eta_y(x;\theta) = L_\theta(x, q_y(\theta)).$$

With the discrepancy η_y, we construct the θ-quantile classifier c_θ as

$$c_\theta(x) = \begin{cases} 1, & \eta_1(x;\theta) < \eta_0(x;\theta) \\ 0, & \eta_1(x;\theta) \geq \eta_0(x;\theta) \end{cases} \tag{2}$$

The quantile θ here will behaves like a tuning parameter.

Consider the 0-1 loss function $L(a,b) = I(a \neq b)$. We will then choose θ that minimizes the $0-1$ loss, i.e.,

$$\theta^* = \mathsf{argmin}_\theta\ \mathbb{E}[L(c_\theta(X),Y)].$$

The final classifier will be

$$c_{\theta^*}(x) = \begin{cases} 1, & \eta_1(x;\theta^*) < \eta_0(x;\theta^*) \\ 0, & \eta_1(x;\theta^*) \geq \eta_0(x;\theta^*) \end{cases} \tag{3}$$

1 Optimality of quantile classifiers

Given the underlying probability model, the Bayes classifier is

$$c^*(x) = \begin{cases} 1, & p(x,1) > p(x,0) \\ 0, & p(x,1) \leq p(x,0). \end{cases}$$

And it leads to the Bayes risk

$$R^* = R(c^*) = \mathsf{min}_c R(c).$$

We will show that under suitable conditions, the classifier c_{θ^*} also achieves the Bayes risk.

Proposition 1 *Suppose that there exists a unique point x^* such that the Bayes classifier can be represented as*

$$c^*(x) = \begin{cases} 1, & x > x^* \\ 0, & x \leq x^* \end{cases} \quad \mathit{or} \quad c^*(x) = \begin{cases} 1, & x \leq x^* \\ 0, & x > x^* \end{cases}$$

and $p(x^*,1) = p(x^*,0)$. Then the risk of quantile classifier

$$R(c_{\theta^*}) = R(c^*).$$

Proof. In this setting, we only need to prove that the quantile classifier will result in a decision boundary x^*.

Consider a given θ. WLOG, we assume that the quantile $q_0(\theta) < q_1(\theta)$. In this scenario, one can easily see that whenever $x \leq q_0(\theta)$, $|x - q_0(\theta)| < |x - q_1(\theta)|$ so $\eta_0(x) < \eta_1(x)$ and $c_\theta(x) = 0$. Similarly, when $x \geq q_1(\theta)$, $c_\theta(x) = 1$.

When $q_0(\theta) < x < q_1(\theta)$, the loss are linearly changing but with a slope θ in η_0 and a slope $1-\theta$ in η_1. One can easily see that

$$x < q_0(\theta) + (q_1(\theta) - q_0(\theta)) \cdot (1-\theta) \Rightarrow \eta_0(x;\theta) < \eta_1(x;\theta)$$

so we see that the decision boundary will be

$$\omega(\theta) = q_0(\theta) + (q_1(\theta) - q_0(\theta)) \cdot (1-\theta) = (1-\theta)q_1(\theta) + \theta q_0(\theta).$$

Namely,

$$c_\theta(x) = \begin{cases} 1, & x > \omega(\theta) \\ 0, & x \leq \omega(\theta). \end{cases}$$

Thus, the risk function will be

$$R(c_\theta) = \mathbb{E}[L(c_\theta(X),Y)] = \underbrace{P(Y=1)}_{\pi_1} F_1(\omega(\theta)) + \underbrace{P(Y=0)}_{\pi_0}(1 - F_0(\omega(\theta))).$$

The first-order condition (taking the first derivative and set it to be 0) leads to

$$\frac{\partial}{\partial\theta} R(c_\theta) = \pi_1 p_{(}\omega(\theta)|1)\omega'(\theta) - \pi_0 p(\omega(\theta)|0)\omega'(\theta) = [\pi_1 p_{(}\omega(\theta)|1) - \pi_0 p(\omega(\theta)|0)]\omega'(\theta) = 0.$$

One can easily see that $\omega'(\theta) \neq 0$ so the case that the first-order condition holds is $\omega(\theta^*) = \omega^*$ such that

$$\pi_1 p(\omega^*|1) = \pi_0 p(\omega^*|0).$$

By assumption, the only value ω^* that solves the above equality is $\omega^* = x^*$, the Bayes decision boundary. Thus, we conclude that the choice θ^* will satisfy $\omega(\theta^*) = x^*$ so the resulting classifier achieves the Bayes risk.

□

2 Empirical quantile classifier

To see why the quantile classifier would be useful, consider the empirical version of it. When a random sample

$$(X_1, Y_1), \cdots, (X_n, Y_n)$$

is given, the quantiles can be estimated easily with the sample quantile

$$\widehat{q}_y(\theta) = \widehat{F}^{-1}(\theta|Y = y), \qquad \widehat{F}(x|y) = \frac{\sum_{i=1}^n I(X_i \leq x, Y_i = y)}{\sum_{j=1}^n I(Y_i = y)}.$$

Thus,

$$\widehat{\eta}_y(x;\theta) = \begin{cases} \theta \cdot |x - \widehat{q}_y(\theta)|, & \text{if } z > \widehat{q}_y(\theta) \\ (1-\theta) \cdot |x - \widehat{q}_y(\theta)|, & \text{if } z \leq \widehat{q}_y(\theta) \end{cases}$$

With this, the θ-quantile classifier can be estimated using

$$\widehat{c}_\theta(x) = \begin{cases} 1, & \widehat{\eta}_1(x;\theta) < \widehat{\eta}_0(x;\theta) \\ 0, & \widehat{\eta}_1(x;\theta) \geq \widehat{\eta}_0(x;\theta) \end{cases}$$

We estimate the optimal θ via empirical risk minimization:

$$\widehat{\theta} = \mathsf{argmin}_\theta\, \widehat{R}(\widehat{c}_\theta) = \mathsf{argmin}_\theta\, \frac{1}{n}\sum_{i=1}^n L(\widehat{c}_\theta(X_i), Y_i).$$

The final classifier is

$$\widehat{c}_{\widehat{\theta}}(x) = \begin{cases} 1, & \widehat{\eta}_1(x;\widehat{\theta}) < \widehat{\eta}_0(x;\widehat{\theta}) \\ 0, & \widehat{\eta}_1(x;\widehat{\theta}) \geq \widehat{\eta}_0(x;\widehat{\theta}) \end{cases}$$

The statistical errors of $\widehat{c}_{\widehat{\theta}}(x)$ comes from two parts: the first one is estimating the quantile, i.e., $|\widehat{q}_y(\theta) - q_y(\theta)|$ and the second part is the empirical risk minimization, i.e., $|\widehat{R}(c_\theta) - R(c_\theta)|$. Both quantities can be estimated uniformly at rate $O_P\left(\sqrt{\frac{\log n}{n}}\right)$. Thus, we conclude that

$$|R(\widehat{c}_{\widehat{\theta}}) - R^*| = O_P\left(\sqrt{\frac{\log n}{n}}\right).$$

Note that we did not make any any parametric assumptions of $p(x,y)$. So this is a nonparametric classifier but still has a nearly parametric classification rate!

3 Remarks

A limitation of the quantile classifier is that quantiles are not well-defined in multivariate case. Namely, when $X \in \mathbb{R}^p$, it is unclear how do we define the quantile. One possible way to generalize this idea to multivariate case is to consider a direction u and we use the quantile of $u^T X$. We may search over different possible directions to use the best one. Another possible way is to perform a componentwise quantile and aggregate them together. See the following papers for ideas related to this:

1. Hall, P., Titterington, D. M., & Xue, J. H. (2009). Median-based classifiers for high-dimensional data. Journal of the American Statistical Association, 104(488), 1597-1608.
2. Hennig, C., & Viroli, C. (2016). Quantile-based classifiers. Biometrika, 103(2), 435-446.

However, the conditions that guarantees consistency to the Bayes risk become very restrictive in multivariate case. It would greatly constraint the shape of the distribution. Fortunately, if the distributions of the two classes are differ by a locational-shift, then the quantile classification still works. See the following paper for more details (and a generalized version that uses multiple directions):

> Farcomeni, A., Geraci, M., & Viroli, C.. (2020). Directional quantile classifiers. arXiv preprint arXiv:2009.05007.

Although the quantile classifier seems to be very powerful, it seems that the resulting classifier is always going to be a half-space classifier, which turns out to be a parametric classifier and this is why we get a good convergence rate. Note that quantile classifier does not assume a parametric form of the underlying quantile or density but the resulting classifier is in fact a parametric classifier.