When Bayes Error Rate can be Achieved by Multiple Quantile Classifier?

Yuanhao

2019-04-29

Contents

1 Introduction

Let an univariate observation x from one of the two populations P_1 and P_2 with prior probabilities π_1 and $\pi_2 = 1 - \pi_1$ and let $y \in \{1, 2\}$ be the population or class indicator. The two populations have cumulative distribution functions $F_1(x)$ and $F_2(x)$ and nonzero derivatives $f_1(x)$ and $f_2(x)$ on the same domain. Thus the corresponding quantile functions $q_1(\theta)$ and $q_2(\theta)$ for $\theta \in (0, 1)$ are continous. We will use the words "population" and "class" interchangeably and restrict the discussion within an univariate input.

The quantile-based classifier (QC) proposed by Hennig and Viroli [2016] was shown to achieve the Bayes error rate under the assumption that the log-odds of class 2 conditioned on x, $g(x) = \log(\pi_2/\pi_1) + \log(f_2(x)/f_1(x))$, has an unique root r_1. In other words, the Bayes decision boundary is only a single point at r_1. Figure 1 shows one possibility of such log-odds functions. In particular, it includes the case of the logistic regression with $\log(p_2/(1-p_2)) = \beta_0 + \beta_1 x$ where $p_2 = Pr(y = 2|x)$.

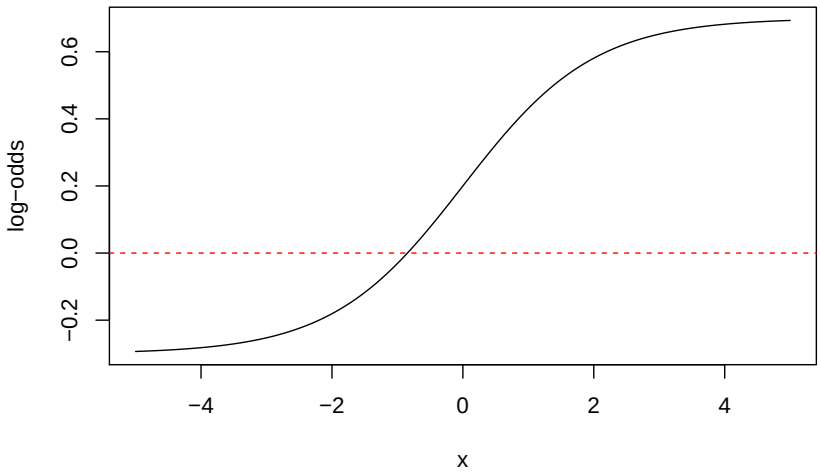

Figure 1: Log-odds function that statistifies the assumption of having the Bayes error rate

Here is an intuitive explanation of why the QC can achieve the Bayes error rate under this assumption. For an univariate x, the population QC is given by,

$$\hat{y} = \mathcal{G}(x \mid \theta) = \begin{cases} 1, & \text{if } s(x \mid \theta) \leq 0 \\ 2, & \text{if } s(x \mid \theta) > 0 \end{cases}, \tag{1}$$

where $\theta \in (0, 1)$ and the discriminant function

$$s(x \mid \theta) = \rho_\theta(x - q_1(\theta)) - \rho_\theta(x - q_2(\theta)), \tag{2}$$

$q_k(\theta)$ is the θ-quantile of P_k, $k = 1, 2$, and $\rho_\theta(u) = u(\theta - \mathrm{I}_{\{u<0\}})$.

The figure below shows the discriminant function $s(x \mid \theta)$ is a piecewise linear function, which intersects with the x-axis at

$$x_0(\theta) = \theta q_1(\theta) + (1 - \theta) q_2(\theta).$$

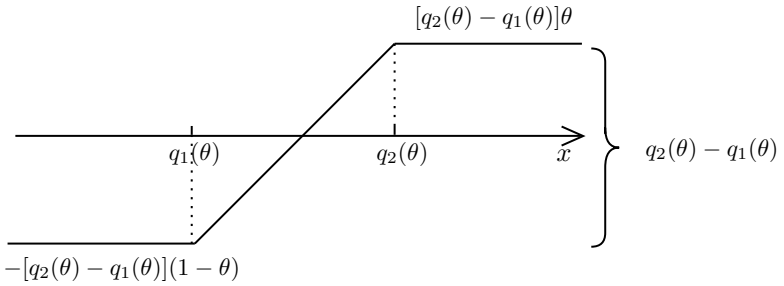

Figure 2: Quantile-based transformation when $q_1(\theta) < q_2(\theta)$.

Since $x_0(\theta)$ is a continous function of $\theta \in (0, 1)$ which can attain all values of the domain of x, then $\exists \theta_0 \in (0, 1)$, s.t., $x_0(\theta_0) = r_1$ and hence the Bayes error rate can be achieved. It also implies that r_1 lies between $q_1(\theta_0)$ and $q_2(\theta_0)$.

Although the assumption of having the log-odds of class 2 $g(x) = \log(\pi_2/\pi_1) + \log(f_2(x)/f_1(x))$ had a unique root is more general than a linear discriminant function as used in the logistic regression, it is still restrictive as $g(x)$ can have more than one root. For example, if $P_1 \sim N(0,1)$ and $P_2 \sim N(0,2)$, then the $g(x)$ is a quadratic function, which has two roots.

Is there a way of extending the QC to achieve the Bayes error rate in case that $g(x)$ have mutiple roots? In this paper, we investigate the possibility of such extensions and propose the multiple quantile classifier that uses multiple quantiles for each variable instead of a signle quantile used by the original QC. By incorporating multiple quantiles, we proved that the Bayes error rate can be achieved by using M quantiles if the log-odds function has M roots. In Section 4, we reveal the relationship between QC and multivariate adaptive regression splines (MARS) [Friedman et al., 1991, Hastie et al., 2009]. We show that the multiple quantile classifier is a restricted MARS and is specially for classification only. We can then use the stagewise approach of MARS to extend and implement the multiple quantile classifier and the resulting approach may be named as multivariate adaptive quantile classificatoin splines (MAQCS).

2 Log-Odds from Baye's Theorem

By the Bayes's theorem, we have the (conditional) log-odds of class 2 as a function of x,

$$g(x) = \log(\pi_2/\pi_1) + \log(f_2(x)/f_1(x)).$$

Then the optimal strategy regarding minimizing the classfication error is to predict $y = 2$ whenever $g(x) > 0$. Thus, the **key** of whether a classifier can achieve the Bayes error rate is whether their approximated log-odds can have the same signs or roots as the log-odds function derived from the Bayes's theorem.

The logistic regression assumes that this log-odds of class 2 is a linear function of x and the optimality of the QC only requires it to have a unique root. However, they are still restrictive as $g(x)$ can have multiple roots.

In the remain of this section, we will show various possibilities of $g(x)$ when comparing different distributions. Without loss of generality, we restrict the priors to be equal, $\pi_1 = \pi_2 = 0.5$.

2.1 Case 1: Beta vs Beta with one roots

When $P_1 \sim \text{Beta}(4,3)$ and $P_2 \sim \text{Beta}(1.5,4)$, we show their density functions and the log-odds function in Figure 3. The log-odds function has only one root $r_1 = 0.3965228$.

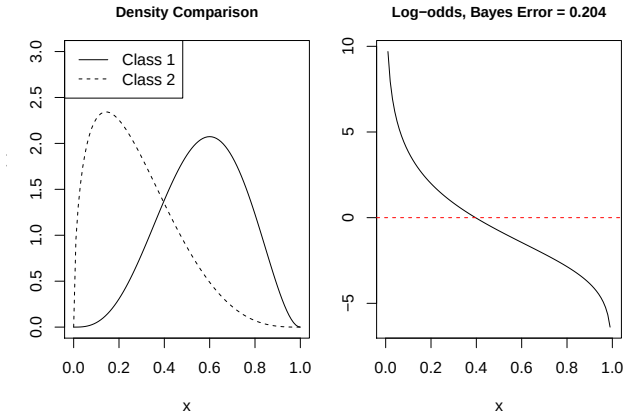

Figure 3: Density functions and log-odds for Beta(4,3) and Beta(1.5,4)

2.2 Case 2: Beta vs Beta with two roots

When $P_1 \sim \text{Beta}(0.6, 0.6)$ and $P_2 \sim \text{Beta}(2, 3)$, we show their density functions and the log-odds function in Figure 4. The log-odds function has two roots $r_1 = 0.1102811$ and $r_2 = 0.6962784$.

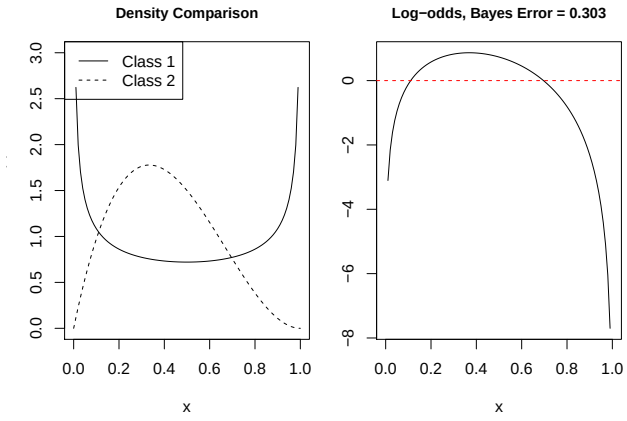

Figure 4: Density functions and log-odds for Beta(0.6,0.6) and Beta(2,3)

2.3 Case 3: Normal vs Normal with two roots

When $P_1 \sim \text{N}(0, 1)$ and $P_2 \sim \text{N}(0, 9)$, we show their density functions and the log-odds function in Figure 5. The log-odds function has two roots $r_1 = -1.572213$ and $r_2 = 1.572213$.

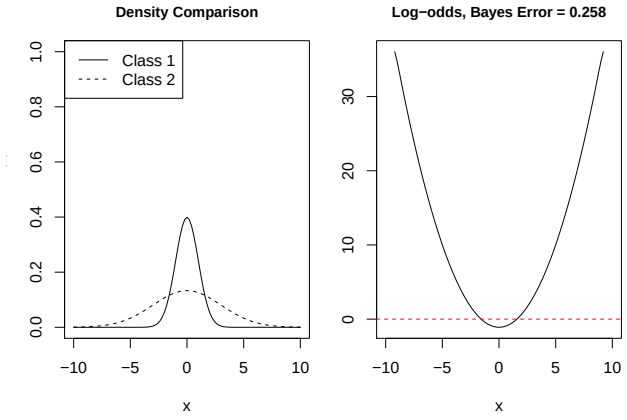

Figure 5: Density functions and log-odds for N(0,1) and N(0,9)

2.4 Case 4: Laplace vs Normal with two roots

When $P_1 \sim \text{N}(3, 1)$ and $P_2 \sim \text{ALD}(0, 0.5, 0.2)$, we show their density functions and the log-odds function in Figure 6. The log-odds function has two roots $r_1 = 1.667663$ and $r_2 = 5.132321$.

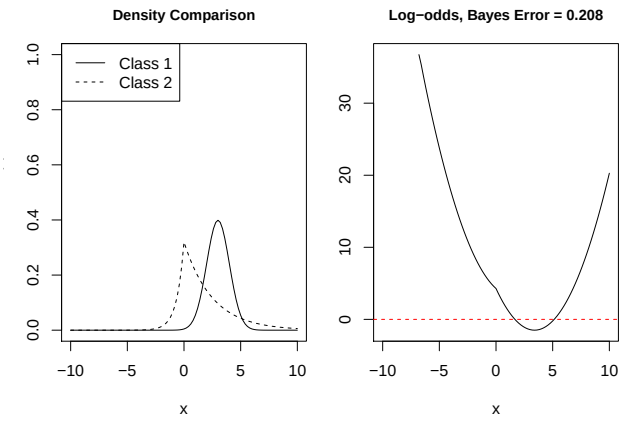

Figure 6: Density functions and log-odds for N(3,1) and ALD(0,0.5,0.2)

2.5 Case 5: Laplace vs Normal with four roots

When $P_1 \sim \mathrm{N}(0.9, 4)$ and $P_2 \sim \mathrm{ALD}(0, 0.55, 0.4)$, the log-odds function has four roots as shown in Figure 7. They are $r_1 = -5.682718$, $r_2 = -1.24455$, $r_3 = 1.082073$ and $r_4 = 6.536126$.

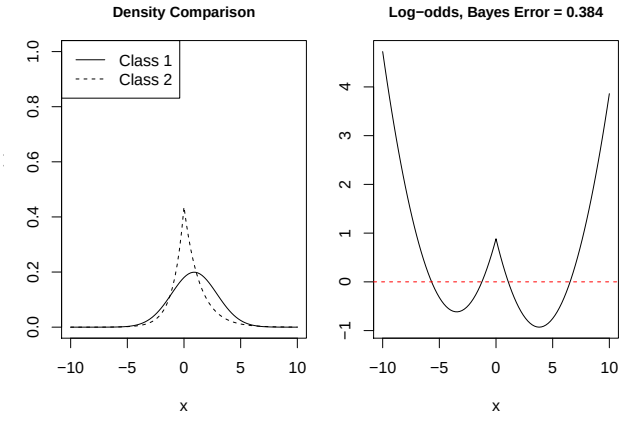

Figure 7: Density functions and log-odds for N(0.9,4) and ALD(0,0.55,0.4)

3 Discriminant Function of MQC

The original QC defined in equation (1) can only have one root in its discriminant function and hence lack of compatibility with Case 2 to 5 mentiond in Section 2.

To overcome this shortcoming, we found that it is effective to aggregate the discriminant functions of the QC's from using diffierent quantiles. The resulting population multiple quantile classifer (MQC) with M quantiles is defined by

$$\hat{y} = \mathcal{G}(x \mid \boldsymbol{\theta}) = \begin{cases} 1, & \text{if } s(x \mid \boldsymbol{\theta}) \leq 0 \\ 2, & \text{if } s(x \mid \boldsymbol{\theta}) > 0 \end{cases}, \tag{3}$$

where $\boldsymbol{\theta} \in (0,1)^H$ and the discriminant function

$$s(x \mid \boldsymbol{\theta}) = \sum_{m=1}^{M} \alpha_m [\rho_{\theta_m}(x - q_1(\theta_m)) - \rho_{\theta_m}(x - q_2(\theta_m))], \tag{4}$$

$\alpha_m \in \mathcal{R}$ for $m = 1, \ldots, M$, $q_k(\theta)$ is the θ-quantile of P_k, $k = 1, 2$, and $\rho_\theta(u) = u(\theta - \mathrm{I}_{\{u<0\}})$.

Our futher goal is to prove the following **Theorem**: Similar to the Bayes optimality of the QC when the log-odds function has a unique root, the MQC with M quantiles can be shown to have the Bayes optimality when the log-odds function has M roots.

We use Case 5 in Section 2 to illustrate that the MQC with four quantiles can achieve the Bayes error rate in case that the log-odds function has 4 roots.

3.1 One quantile

To help us (mannually) pick up quantiles for MQC, it is useful to see how the discriminant function of a single-quantile QC can vary by the choice of quantiles. From Figure 8, as θ increases from 0 to 1, the support of the non-constant component is moving to the right, and the direction of the non-constant component flips four times. They reflect that the QC has different class preferences in different domains of x, which agrees with the true log-odds function.

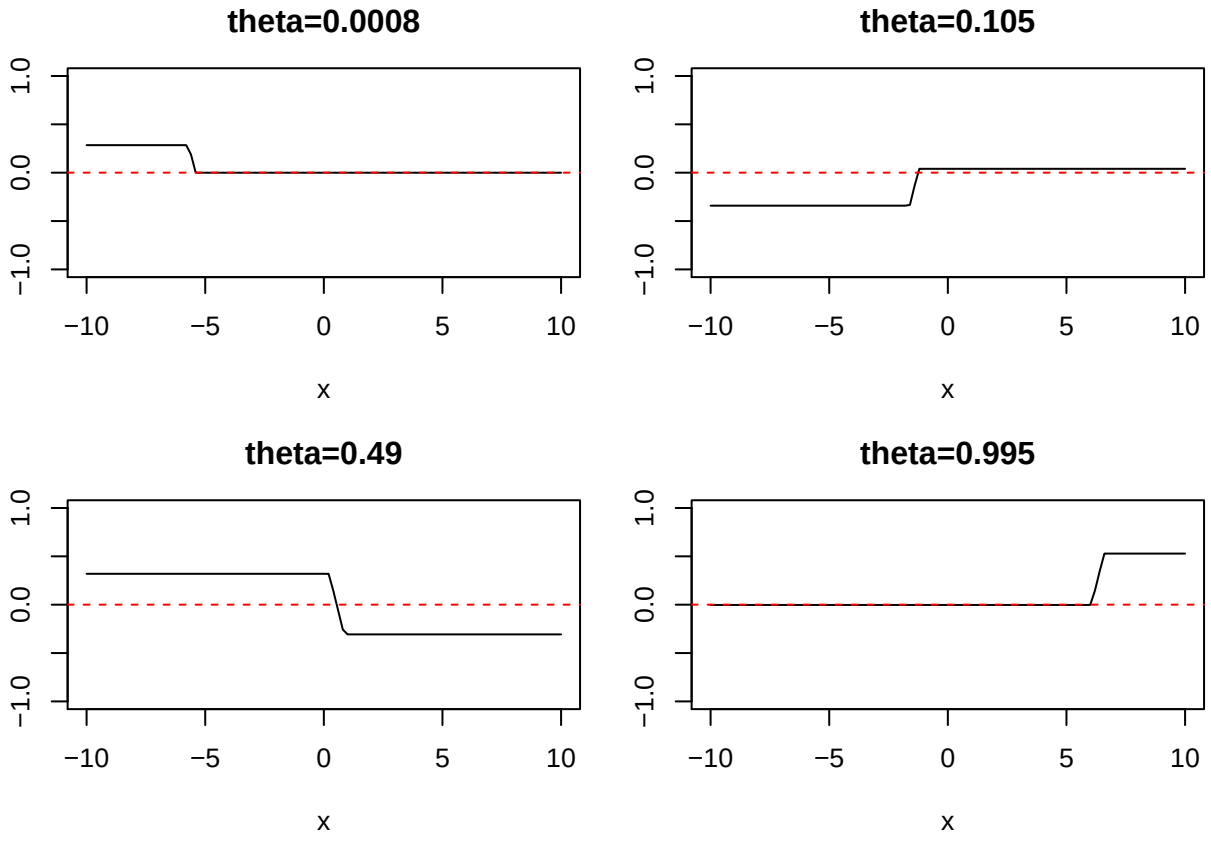

Figure 8: Disciminant functions of QC w.r.t. theta=0.001, 0.1,0.6,0.9

3.2 Four quantiles

One can play with the choice of $\boldsymbol{\theta}$ in order to have the same four roots $r_1 = -5.682718$, $r_2 = -1.24455$, $r_3 = 1.082073$ and $r_4 = 6.536126$ given by the Bayes optimum. Figure 9 show that the MQC with four quantiles at $\boldsymbol{\theta} = (0.0008, 0.105, 0.49, 0.995)$ has four roots that are close to (r_1, r_2, r_3, r_4).

This example implies that it is still possible to achieve the Bayes error rate with MQC when the true log-odds function has multiple roots.

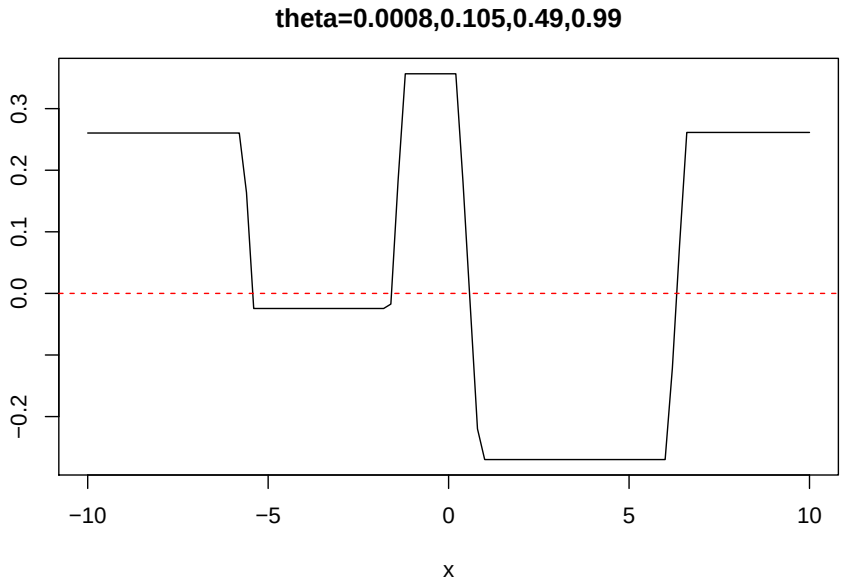

Figure 9: Disciminant functions of MQC with quantiles at theta=(0.0008, 0.105, 0.49, 0.995)

4 MQC as Restricted MARS

Multivariate adaptive regression splines (MARS) Friedman et al. [1991], Hastie et al. [2009] is an adaptive procedure for regression. It uses a stepwise approach similar to CART to do automatic variable selection and fitting. It can also be extended for classification with a proper link function to incorporate the generalized linear models.

We briefly introduce the representaion of MARS in the univariate case here. Details of the estimation can be found in Friedman et al. [1991] and Hastie et al. [2009].

MARS uses the following collection of piecewise linear basis functions

$$\mathcal{C} = \{(x-t)_+, (t-x)_+, t = x_1, \ldots, x_n\},$$

where "+" means the positive part.

Then the MARS can be represented as a function

$$f(x) = \beta_0 + \sum_{m=1}^{M} \beta_m h_m(x),$$

where $h_m(x)$ is a function in $\mathcal{C}$ or a product of two or more such functions.

In contrast, the discriminant function of a single-quantile QC can be re-expressed as a linear combination of the piecewise linear basis functions used by MARS,

$$\begin{aligned} s(x \mid \theta) &= \rho_\theta(x - q_1(\theta)) - \rho_\theta(x - q_2(\theta)) \\ &= (x - q_1(\theta))(\theta - 1_{\{x<q_1(\theta)\}}) - (x - q_2(\theta))(\theta - 1_{\{x<q_2(\theta)\}}) \\ &= \theta(x - q_1(\theta))_+ + (1-\theta)(q_1(\theta) - x)_+ - \theta(x - q_2(\theta))_+ - (1-\theta)(q_2(\theta) - x)_+. \end{aligned}$$

Thus, the MQC with the discriminant function in equation (4) can be viewed as a restricted MARS by aggregating some of its basis functions without products.

One important adavantage of using the piecewise linear basis function for MARS is their ability to operate locally and they are zero over part of their range, and hence the model can be built parsimoniously in high dimension. With the piecewise linear basis function used for MQC, we reduce the number of candidate baisis functions from $2Np$ to $Np/2$ in the balanced case, and ensure the population optimalty in the univariate case by the Theorem we are going to prove.

By following the implmentation of MARS in case of multivariate inputs, we can further extend MQC and call this new approach multivaraite addaptive quantile classification splines (MAQCS).

5 Simulation Study

We access the performance of the linear logistic regression, the polynominal logistic regression, MQC and MARS through the simulation study of the examples mentioned in Section 2.

Training samples of size n were simulated from two populations where half belong to Class 1 and half belong to Class 2. To obtain an accurate estimate of the test error rate, balanced test samples of size 10^6 were used.

For the MQC method, we mannually select $\boldsymbol{\theta}$ to optimize the result so there will be a bias. However, our goal here is to see how low the test error rate the MQC can achieve.

Table 1 shows the detailed set-ups of the simulation study in each scenario. The summary of test errors for all scenarios can be found in Table 8 at the end of this section.

Table 1: Simulation Set-Up for each Scenario

Case	#Roots	n	n-class1	n-class2
Beta(4,3) vs Beta(1.5,4)	1	100	50	50
Beta(0.6,0.6) vs Beta(2,3)	2	100	50	50
N(3,1) vs ALD(0,0.5,0.2)	4	100	50	50
N(3,1) vs ALD(0,0.5,0.2)	4	200	100	100
N(3,1) vs ALD(0,0.5,0.2)	4	400	200	200
N(3,1) vs ALD(0,0.5,0.2)	4	2000	200	200

5.1 Case 1: Beta vs Beta with one roots and n=100

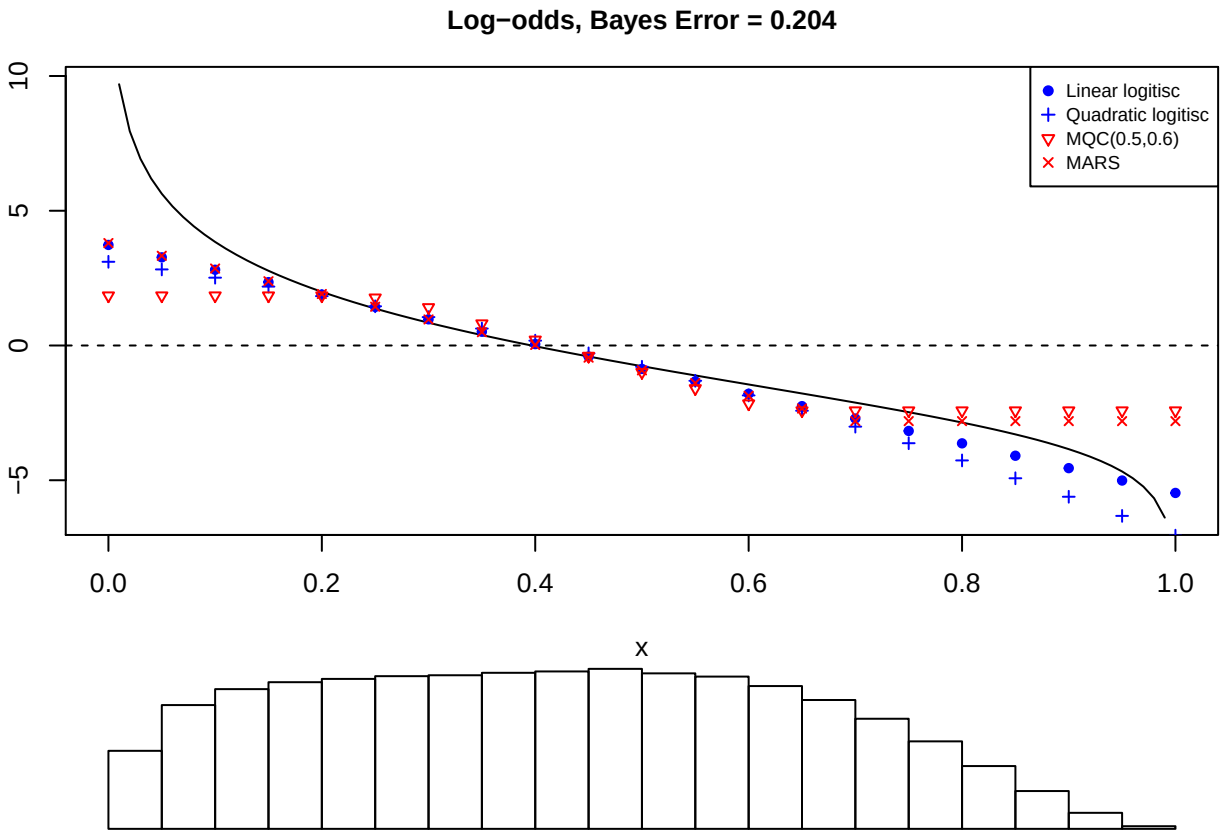

Figure 10: log-odds for Beta(4,3) and Beta(1.5,4) and the approximations, n=100

Table 2: Test Error Rate for Classification between Beta(4,3) and Beta(1.5,4), n=100

Method	TestError	SD
Linear logitisc	0.2037	0.0004
Quadratic logitisc	0.2048	0.0004
QC(0.78)	0.2076	0.0004
MQC(0.5,0.6)	0.2046	0.0004
MARS	0.2036	0.0004

5.2 Case 2: Beta vs Beta with two roots and n=100

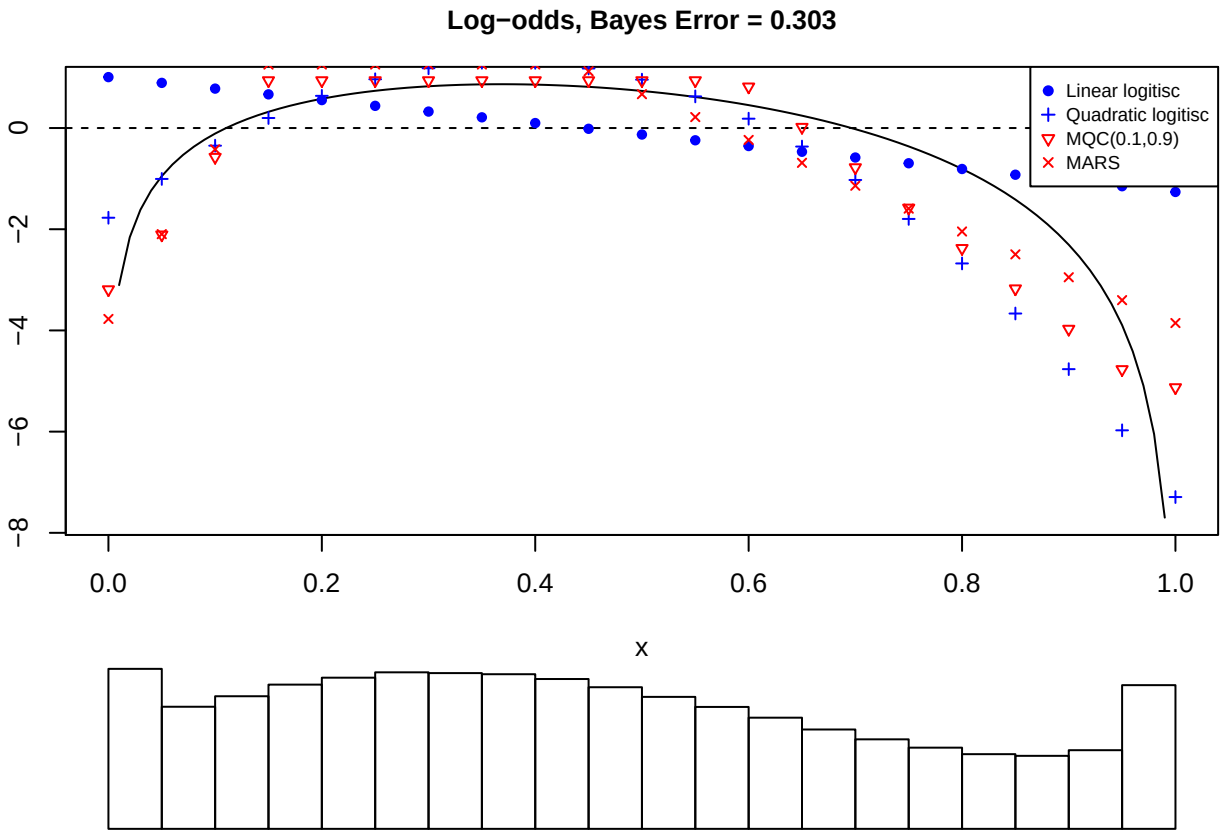

Figure 11: log-odds for Beta(0.6,0.6) and Beta(2,3) and the approximations, n=100

Table 3: Test Error Rate for Classification between Beta(0.6,0.6) and Beta(2,3), n=100

Method	TestError	SD
Linear logitisc	0.4292	0.0005
Quadratic logitisc	0.3110	0.0005
QC(0.58)	0.4078	0.0005
MQC(0.1,0.9)	0.3057	0.0005
MARS	0.3191	0.0005

5.3 Case 5: Laplace vs Normal with four roots and n=100

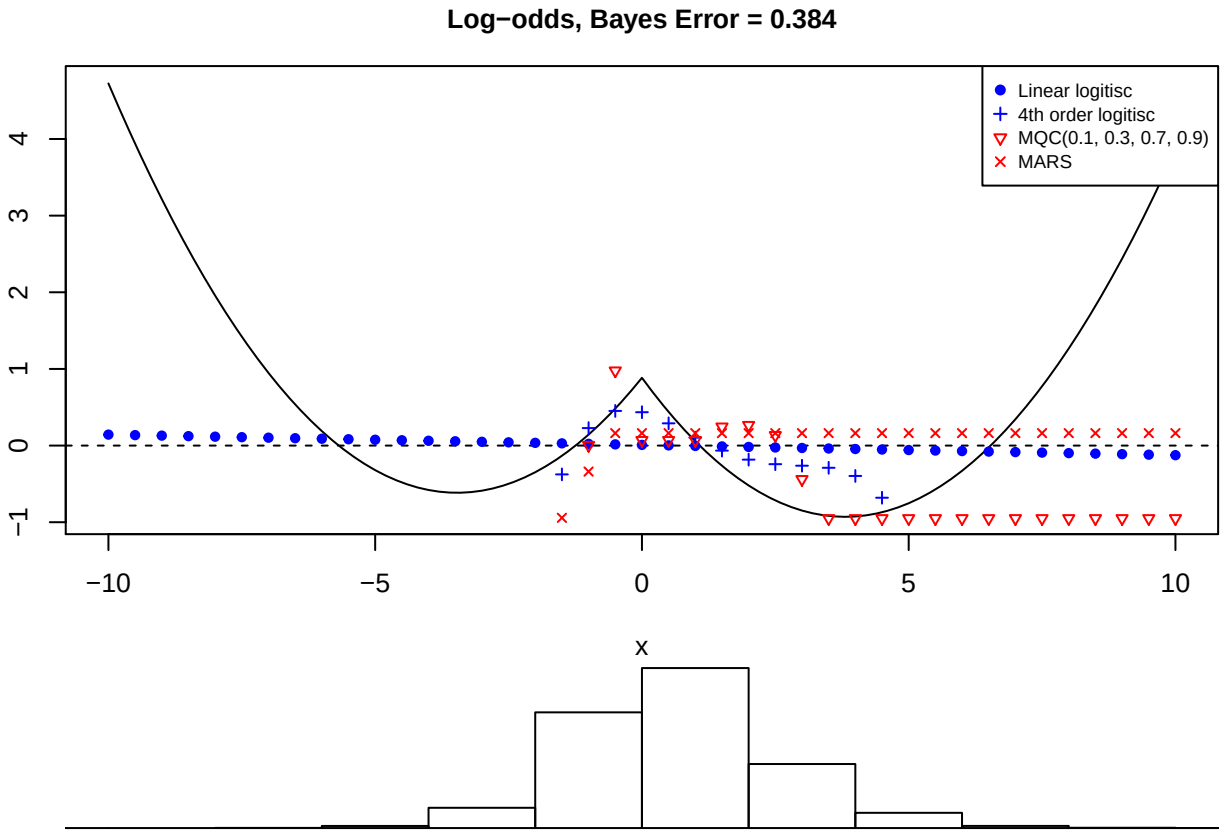

Figure 12: log-odds for N(0.9,4) and ALD(0,0.55,0.4) and the approximations, n=100

Table 4: Test Error Rate for Classification between N(0.9,4) and ALD(0,0.55,0.4), n=100

Method	TestError	SD
Linear logitisc	0.4116	0.0005
Quadratic logitisc	0.3868	0.0005
QC(0.04)	0.4816	0.0005
MQC(0.1, 0.3, 0.7, 0.9)	0.4291	0.0005
MARS	0.4871	0.0005

5.4 Case 5: Laplace vs Normal with four roots and n=200

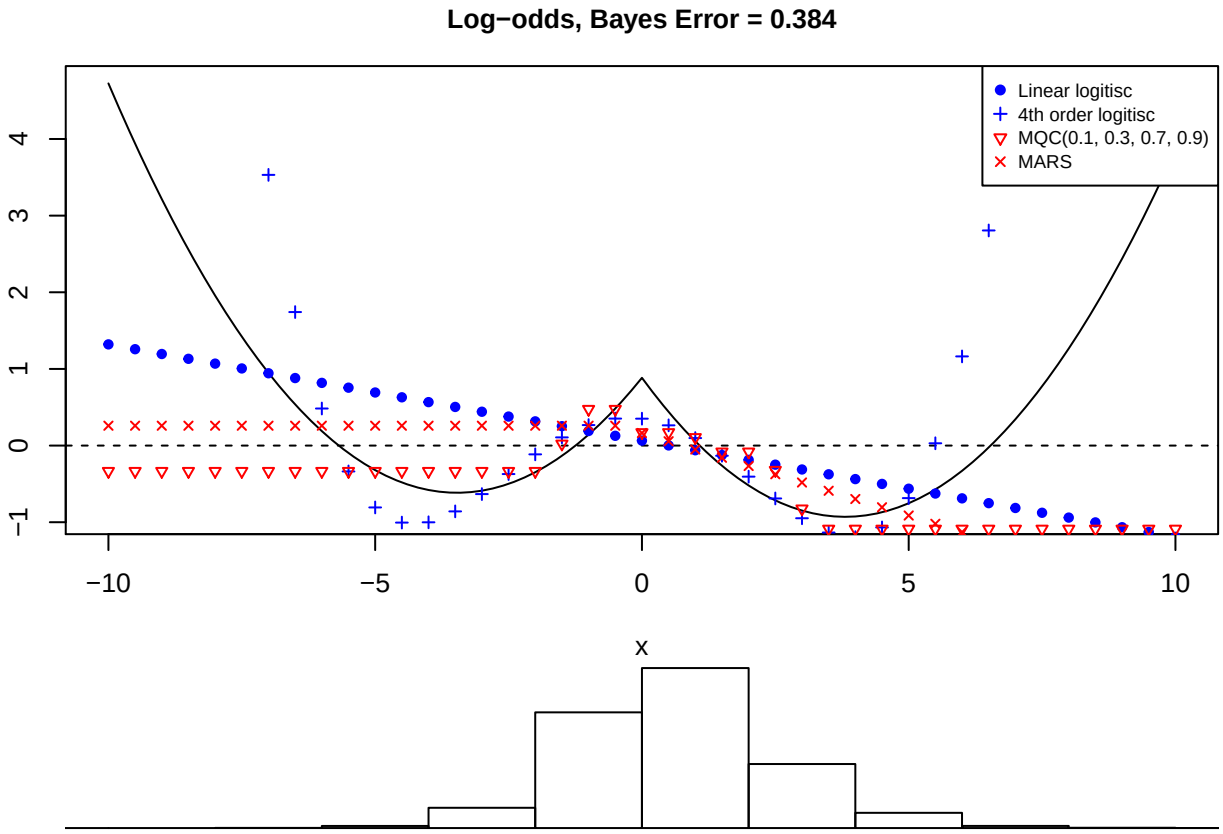

Figure 13: log-odds for N(0.9,4) and ALD(0,0.55,0.4) and the approximations, n=200

Table 5: Test Error Rate for Classification between N(0.9,4) and ALD(0,0.55,0.4), n=200

Method	TestError	SD
Linear logitisc	0.4179	0.0005
Quadratic logitisc	0.3888	0.0005
QC(0.71)	0.4051	0.0005
MQC(0.1, 0.3, 0.7, 0.9)	0.3872	0.0005
MARS	0.4087	0.0005

5.5 Case 5: Laplace vs Normal with four roots and n=400

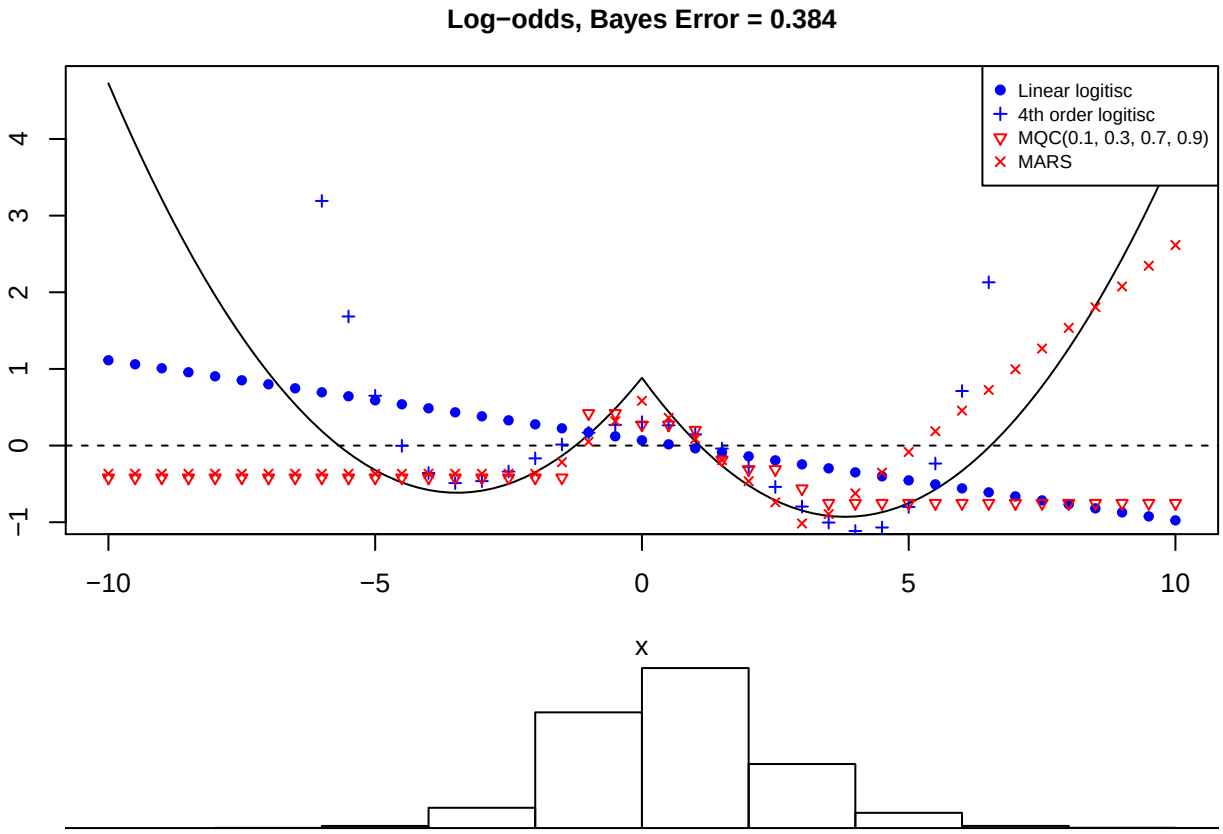

Figure 14: log-odds for N(0.9,4) and ALD(0,0.55,0.4) and the approximations, n=400

Table 6: Test Error Rate for Classification between N(0.9,4) and ALD(0,0.55,0.4), n=400

Method	TestError	SD
Linear logitisc	0.4121	0.0005
Quadratic logitisc	0.3893	0.0005
QC(0.5)	0.4208	0.0005
MQC(0.1, 0.3, 0.7, 0.9)	0.3865	0.0005
MARS	0.3875	0.0005

5.6 Case 5: Laplace vs Normal with four roots and n=4000

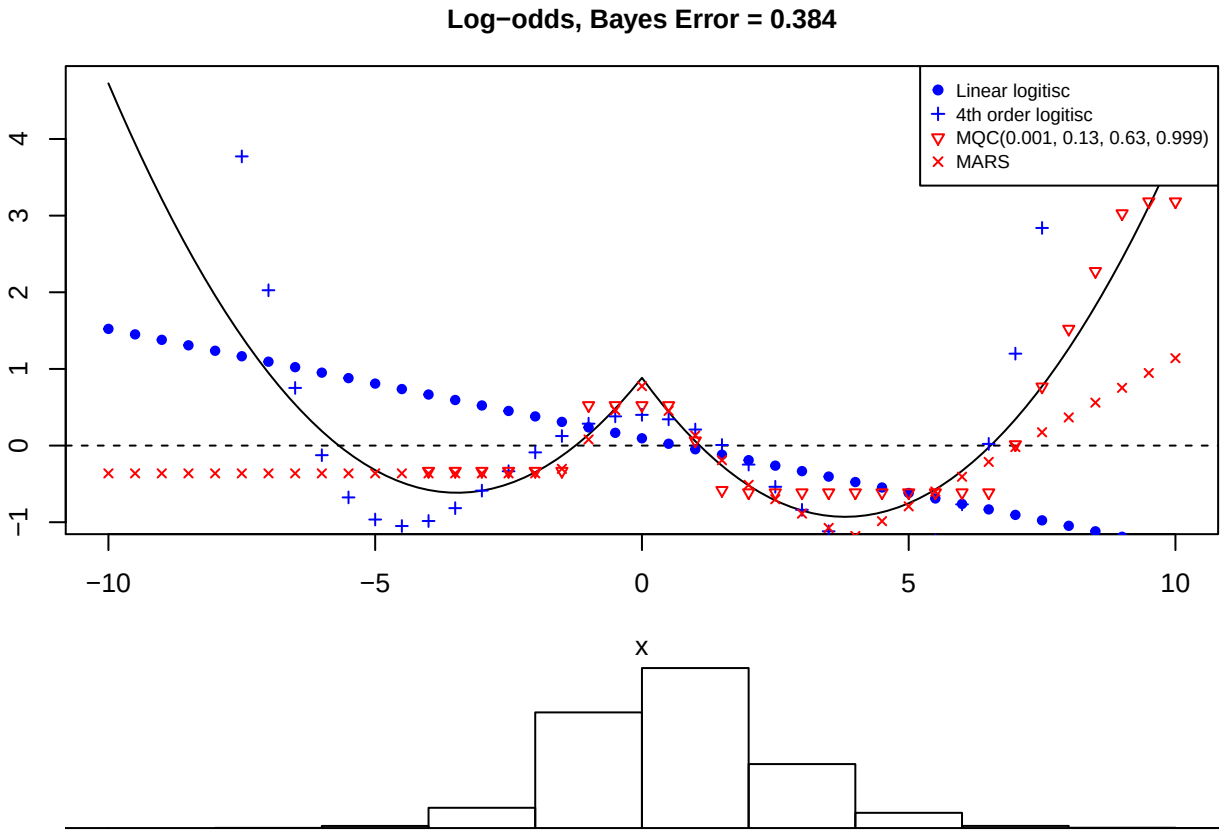

Figure 15: log-odds for N(0.9,4) and ALD(0,0.55,0.4) and the approximations, n=4000

Table 7: Test Error Rate for Classification between N(0.9,4) and ALD(0,0.55,0.4), n=4000

Method	TestError	SD
Linear logitisc	0.4117	0.0005
Quadratic logitisc	0.3926	0.0005
QC(0.663)	0.4051	0.0005
MQC(0.001, 0.13, 0.63, 0.999)	0.3841	0.0005
MARS	0.3854	0.0005

5.7 Summary

From Table 8, we see that the MQC has performed much better than the linear logistic regression and the QC. This agrees with our theoretical discussion as the linear logistic regression and the QC can only achieve optimality if the true log-odds function has a unique root.

The estimated log-odds function of MQC and MARS are closed most of the time as seen from Figure 10 to 15. This supports our argument that MQC is a restricted MARS. Meanwhile, MQC may outperform MARS with a small sample in case of comparing N(3,1) and ALD(0,0.5,0.2).

Table 8: Summary of the Simulation Study

case	n	Bayes	logistic	polylogisitc	QC	MQC	MARS
Beta(4,3) vs Beta(1.5,4)	100	0.2041	0.2037(4e-04)	0.2048(4e-04)	0.2076(4e-04)	0.2046(4e-04)	0.2036(4e-04)
Beta(0.6,0.6) vs Beta(2,3)	100	0.3034	0.4292(5e-04)	0.311(5e-04)	0.4078(5e-04)	0.3057(5e-04)	0.3191(5e-04)
N(3,1) vs ALD(0,0.5,0.2)	100	0.3837	0.4116(5e-04)	0.3868(5e-04)	0.4816(5e-04)	0.4291(5e-04)	0.4871(5e-04)
N(3,1) vs ALD(0,0.5,0.2)	200	0.3837	0.4179(5e-04)	0.3888(5e-04)	0.4051(5e-04)	0.3872(5e-04)	0.4087(5e-04)
N(3,1) vs ALD(0,0.5,0.2)	400	0.3837	0.4121(5e-04)	0.3893(5e-04)	0.4208(5e-04)	0.3865(5e-04)	0.3875(5e-04)
N(3,1) vs ALD(0,0.5,0.2)	4000	0.3837	0.4117(5e-04)	0.3926(5e-04)	0.4051(5e-04)	0.3841(5e-04)	0.3854(5e-04)

6 Further Work

In conclusion, we now have two goals to achieve.

1. Deduce the Theorem that the M-quantiles MQC can achieve the Bayes error rate if the log-odds function has M roots in the univariate case.
2. Implement the multivaraite adaptive quantile classification spline (MAQCS).

Some questions may arise.

1. One may be worried about the accuracy of the estimated in a small sample. However, this might be not a big issue if they were viewd as a component of the basis functions.
2. One may ask why we prefer MQC or MARS to the polynomial logistic regression if log-odds have multiple roots. The answer is that the polynomial basis functions would produce a nonzero product everywhere (non-parsimonious) in high-dimensional data and the resulting decision boundary can be overfitting to the training data.

References

Jerome H Friedman et al. Multivariate adaptive regression splines. *The annals of statistics*, 19(1):1–67, 1991.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer-Verlag, New York., 2 edition, 2009.

C. Hennig and C. Viroli. Quantile-based classifiers. *Biometrika*, 103(2):435–446, 2016. doi: 10.1093/biomet/asw015. URL +http://dx.doi.org/10.1093/biomet/asw015.