# *Semi-parametric classification*

***Kevin Tao[a], Niyu Jia[b]***

[a] *taokevin@bu.edu*
[b] *nyjia@bu.edu*

***Keyword:*** Semi-parametric statistics, classification, bandwidth selection

## 1. Introduction

The problem of classification is becoming increasingly important in a variety of quantitative field such as bioinformatics, image processing, and statistical machine learning. For decades, great emphasis had been placed on parametric classification, where the classification rule (or the estimated posterior probability) is based on a model with a fixed number of parameter and stringent construct. Within the realm of parametric classifiers, there are two common approach to solving a binary classification problem. The first approach assumes certain distributional property on $f(x|y)$ the conditional density of each group. Examples of such methods are linear discriminant analysis (assuming equal covariance multivariate Gaussian distributions), and quadratic discriminant analysis (assuming unequal covariance multivariate Gaussian distributions). The second approach assumes no distributional property of $f(x|y)$, and in fact avoids estimating $f(x|y)$ altogether. Example of such method is the logistic regression, which assumes a that log-odds ($log(\frac{p}{1-p})$) are linearly related to each coordinates of the feature space. Parametric models are popular due to their relatively intuitive approach, lower computational costs, and useful asymptotic properties. However, assumptions imposed by parametric models are often too stringent, leading to potentially mediocre performance when the structure of the data violates assumptions. Hence, with the power of modern computers, it would be desirable to consider a non-parametric/semi-parametric approach toward classification that is local in nature.

In this report we examine a kernel density based binary and multi-class semi-parametric classifier that is local in nature, and requires no assumption on the underlying group-specific. More specifically, we are interested in solving classification problem of the form $Y \sim Bernoulli(p)$ and $Y|x \sim Bernoulli(p(x))$ for the binary case, and $Y \sim Multinomial(p_1, ..., p_T)$ and $Y|x \sim Multinomial(p_1(x), ..., p_T(x))$ for the multi-class case. In section 2, we provide a revision of key properties of the kernel density estimators. In section 3 and 4, we introduce the kernel based classifier, its properties, and the associated bandwidth selection procedures (its advantages and limitations). Finally, in section 5, we compare the kernel based classifier to other parametric models via application to data sets from the UCI machine learning repository.

## 2. Preliminary

A commonly used nonparametric model for multivariate density estimation is the Multivariate Kernel Density estimator. It estimates the density of a point of evaluation $x \in \mathbb{R}^d$ with the formula

$$\hat{f}(x) = \frac{1}{n|H|} \sum_{i=1}^{n} K(H^{-1}(x - x_i)) \tag{1}$$

where $H \in \mathbb{R}^{d \times d}$ is the bandwidth matrix, $|\cdot|$denote the matrix determinant, and $K(u)$ is the multivariate kernel density that satisfy the following properties

$$\begin{aligned} Normalized &: \int K(u)du = 1 \\ Symmetry &: \int uK(u)du = 0 \\ Second\ Moments &: \int uu^T K(u)du = \mu_2(K) I_{d\times d}\ ,\ \mu_2(K) \in \mathbb{R} \end{aligned} \tag{2}$$

Under the assumption that the underlying density $f(x)$ is sufficiently smooth (with a well defined Hessian matrix), it can be shown (Appendix A) that the bias and variance of $\hat{f}(x)$ are given by the following formulas

$$\begin{aligned} Bias\hat{f}(x) &= \frac{\mu_2(K)}{2} tr(H^T \nabla^2 f(x) H) \\ var\hat{f}(x) &= \frac{f(x) \|K\|_2^2}{n|H|} \end{aligned} \tag{3}$$

where $\nabla^2 f(x)$ denotes the Hessian Matrix evaluated at $x$ and $\|\cdot\|_p$ denotes the $L^p$ norm. A quick application of Markov's Inequality shows that $P(|\hat{f}(x) - f(x)| > \epsilon) \leq \frac{E|\hat{f}(x) - f(x)|^2}{\epsilon^2} = \frac{Bias^2\hat{f}(x) + var\hat{f}(x)}{\epsilon^2}$. Thus, the estimator $\hat{f}(x)$ is consistent, $\hat{f}(x) \xrightarrow{P} f(x)$ if $tr(H^T \nabla^2 f(x) H) \to 0$ (i.e. $H$ converges to the zero matrix term-wise, and $|H| \to 0$), and $n|H| \to \infty$

## 3. The Kernel based Classifier

### 3.1. Binary setting

When faced with a binary classification problem, a very natural approach is to consider the posterior probability, and using Bayes' formula, it is not hard to see that

$$P(Y = 1|x) = \frac{p f_1(x)}{p f_1(x) + (1-p) f_0(x)} \tag{4}$$

where 1 and 0 denote the positive and negative group, $p$ is the prior probability of being in group 1, and $f_i(x)$ is the density function of group i. Since the maximum likelihood estimator of $p$ is known to be $\bar{Y}$, it only remains to find a good estimator for the group specific densities, and a non-parametric method would be to use the multivariate kernel density estimator [6]. Hence, if we abbreviate $P(Y = 1|x)$ with $m(x)$, we can define the estimator:

$$\hat{m}(x) = \frac{\hat{f}_1(x)\bar{Y}}{\bar{Y}\hat{f}_1(x) + (1-\bar{Y})\hat{f}_0(x)} \tag{5}$$

where $\hat{f}_i(x)$ is the group specific kernel density estimator given by $\frac{1}{n_i|H_i|} \sum_{j=1}^{n_i} K_i(H_i^{-1}(x - x_j))$, where $n_i$, $K_i$, and $H_i$ are the group specific sample size, kernel function, and bandwidth matrices. A key motivation of this estimator is the fact that the kernel density estimator is adaptive and local in nature, freeing us from unnecessary assumptions on the data structure, and the relationship between $P(Y = 1|x)$ and the feature space coordinates [4]. In this context, A test sample vector is classified to group 1 if $\hat{m}(x)$ exceeds a certain threshold say $\theta$ that ranges from 0 to 1.

### 3.2. Multi-class setting

Using similar argument as the binary classification problem, a natural approach to a multi-class (say T classes) classification problem is also to assess the posterior probabilities:

$$P(y = G_j|x) = \frac{f_j(x)p_j}{\sum_1^T f_i(x)p_i} \tag{6}$$

where $p_i$ are the prior probabilities of the associated multinomial distribution (when feature vectors x are not observed). Similarly, since sample proportion is the maximum likelihood estimator, we can create estimators for the posterior probabilities as follows:

$$\hat{P}(y = G_j|x) = \frac{\hat{f}_j(x)\frac{n_j}{N}}{\sum_1^T \hat{f}_i(x)\frac{n_i}{N}} \tag{7}$$

A intuitive classification rule would be to choose the group $G_k$ that maximizes the estimated posterior probabilities. It is also worth noting that every component of the quotient is consistent (converges in probability); hence, the vector of these components converges in probability, and so does the posterior probability since it is a continuous map.

## 4. Bandwidth selection

Just like any other non-parametric/semi-parametric models, the specific choice of each group specific bandwidth matrices is absolutely essential to the success of the model. Till this day, countless algorithms have been invented and discussed, ranging from MISE minimization, to variable bandwidth selection through weighting. Since most algorithms are computationally difficult to reproduce, we consider 3 well-known bandwidth selection procedures in this section

### 4.1. MISE Minimization

As discussed in the work [3], a commonly used approach in the early days of kernel classification is to select for bandwidths that minimizes the mean integrated square error of the density estimator itself. More specifically, we choose $H_i$ according to:

$$\begin{aligned}\hat{H}_i &= \arg\min_{H_i} \int E[\hat{f}_i(x) - f_i(x)]^2 dx \\ &= \arg\min_{H_i} \int MSE(\hat{f}_i(x)) \\ &= \arg\min_{H_i} \int Bias^2(\hat{f}_i(x)) + Var(\hat{f}_i(x)) dx\end{aligned} \tag{8}$$

where the bias and variance of $\hat{f}_i(x)$ is given in section 2. Intuitively, this method make the implicit assumption that by choosing the optimal bandwidth matrices for every group's density estimator, we optimize the classifier. This method, has several advantages. Namely, it provides the user with many previously available algorithms (eg: least square cross-validation, biased cross-validation, and etc.), and could be quite computationally friendly if the normal reference bandwidth is chosen. In addition, since most bandwidth selection procedure for classification requires heavy computation load, thus restricting bandwidth matrices to the form $H_i = h_i \times I_{d\times d}$, the normal reference bandwidth could be desirable as it considers the covariance between the coordinates. Furthermore, the fact that MISE minimization procedure can be generalized to classification with multiple classes should not be overlooked, as many more delicate methods are tailored toward binary classification, unsuitable for multi-class generalization.On the other hand, it is not hard to see that MISE minimization is in fact a rather ad-hoc method since it not only doesn't actually minimizes the MISE of the estimated posterior probability as it chooses bandwidth separately (ignoring possible interaction), but also fails to recognize that a classification problem is fundamentally different from a density estimation problem [3]; more specifically, we are more interested in the accuracy of the classifier, and a estimated posterior probability of 0.1 and 0.49 (assuming binary setting) makes little to no difference.

### 4.2. Difference between Density

Suppose that the prior probability $p$ is known and a threshold value of 0.5 is used for binary classification, a more delicate method, proposed by [5], is to examine the difference between the two competing densities. An alternative formulation of the classification rule then becomes: $g(x) \equiv pf_1(x) - (1-p)f_0(x)$, $\hat{g}(x) \equiv p\hat{f}_1(x) - (1-p)\hat{f}_0(x)$, classifying to positive if $\hat{g}(x) \geq 0$. Since $g(x)$ also defines the decision boundary, in the work [5], the authors argue that optimal bandwidth matrices should be chosen together (as suppose to independently), and they should minimize the $L^2$ distance between $\hat{g}(x)$ and $g(x)$. If we assume that $H_i = h_i \times I_{d\times d}$, and define the function $D(h_1, h_0) = \int [\hat{g}(z) - g(z)]^2 dz$, expanding the squares we get:

$$D(h_1, h_0) = \int \hat{g}(z)^2 dz - 2\int \hat{g}(z)g(z)dz + \int g(z)^2 dz \tag{9}$$

Since the last integral is a fixed constant, any minimize of $\delta(h_1, h_0) = \int \hat{g}(z)^2 dz - 2\int \hat{g}(z)g(z)dz$ is also the minimizer for the objective function; hence we can restate the optimization criterion as:

$$\delta(h_1, h_0) = \int [p\hat{f}_1(x) - (1-p)\hat{f}_0(x)]^2 dx - 2\int (p\hat{f}_1(x) - (1-p)\hat{f}_0(x))(pf_1(x) - (1-p)f_0(x))dx \tag{10}$$

Taking expectation of the second integral yields:

$$E[p^2 \int \hat{f}_1 f_1 + (1-p)^2 \int \hat{f}_0 f_0 - p(1-p)\int \hat{f}_1 f_0 - p(1-p)\int \hat{f}_0 f_1] \tag{11}$$

According to [5], an unbiased estimator is:

$$\frac{p^2}{m}\sum_{i=1}^{n_1}\hat{f}_1^{-i}(x_i)+\frac{(1-p)^2}{n_0}\sum_{i=1}^{n_0}\hat{f}_0^{-i}(x_i)-p(1-p)\{\frac{1}{n_1}\sum_{i=1}^{n_1}\hat{f}_0(x_i)+\frac{1}{n_0}\sum_{i=1}^{n_0}\hat{f}_1(x_i)\} \tag{12}$$

where $\hat{f}_j^{-i}(x)$ is density estimator with the $i^{th}$ observation from group j removed. Hence, we have obtained:

$$\begin{aligned}\hat{\delta}(h_1,h_0) = &\int[p\hat{f}_1(x)-(1-p)\hat{f}_0(x)]^2dx \\ &-2[\frac{p^2}{m}\sum_{i=1}^{n_1}\hat{f}_1^{-i}(x_i)+\frac{(1-p)^2}{n_0}\sum_{i=1}^{n_0}\hat{f}_0^{-i}(x_i)-p(1-p)\{\frac{1}{n_1}\sum_{i=1}^{n_1}\hat{f}_0(x_i)+\frac{1}{n_0}\sum_{i=1}^{n_0}\hat{f}_1(x_i)\}]\end{aligned} \tag{13}$$

Though not explicitly shown in the work [5], if we utilize the multivariate standard Gaussian kernel, we can in fact calculate the integral explicitly. Using the property of normal densities convolution, we can rewrite the integral as:

$$\frac{p^2}{n_1^2h_1^d}\sum_{i=1}^{n_1}\sum_{j=1}^{n_1}K^{(2)}(\frac{x_i-x_j}{h_1})+\frac{(1-p)^2}{n_0^2h_0^d}\sum_{i=1}^{n_0}\sum_{j=1}^{n_0}K^{(2)}(\frac{x_i-x_j}{h_0})-\frac{2p(1-p)}{n_1n_0h_1^d}\sum_{i=1}^{n_0}\sum_{j=1}^{n_1}\int K(z)K(\frac{x_j-x_i}{h_1}+\frac{zh_0}{h_1})dz \tag{14}$$

where $K^{(2)}$ is the convoluted standard multivariate normal density, namely the density of $N(0,2\times I)$. At a first glance the integral in the right seems formidable, yet it is not hard to see that $K(\frac{x_j-x_i}{h_1}+\frac{zh_0}{h_1})$ is equivalent to $(\frac{h_1}{h_0})^d$ multiplied by $w_t(z)$, where $t$ follows the multivariate normal density with mean $-\frac{x_j-x_i}{h_0}$ and covariance $\frac{h_1^2}{h_0^2}\times I$. Hence using the findings from [2], we observe that the integral evaluates to:

$$\int K(z)K(\frac{x_j-x_i}{h_1}+\frac{zh_0}{h_1})dz=\frac{h_1^d}{h_0^d}\frac{exp[-\frac{1}{2}(\alpha_{ij}^T(\frac{h_0^2}{h_1^2})+1)^{-1}\alpha_{ij}]}{\sqrt{(2\pi)^d(1+\frac{h_1^2}{h_0^2})^d}} \tag{15}$$

where $\alpha_{ij}$ is $-\frac{x_j-x_i}{h_0}$. Putting all pieces together will yield a criterion that minimizes $\hat{\delta}(h_1,h_0)$, and subsequently minimizes the estimated objective function. In pratice, since the piror probability often has to be estimated we can simply replace $p$ with the estimated sample proportion of positives. If we denote $(\hat{h_1},\hat{h_0})$ as the minimizer of $\hat{\delta}(h_1,h_0)$, [5] showed that this minimization criterion is optimal in the sense that:

$$\frac{D(\hat{h_1},\hat{h_0})}{infD(h_1,h_0)}\to 1 \tag{16}$$

almost surely when both $n_1$ and $n_0$ tends to infinite. The clear advantage of this procedure over the MISE minimization procedure is that there is a properly established objective function that is backed up by the context of classification. However, this procedure also has its drawbacks. In addition, as shown in [5], this procedure is also compatible with feature coordinates that are categorical, or a mixture of categorical and continuous, making it applicable to more situation. However, the procedure is computationally challenging since for any given pairs of bandwidth values, a full computation of the objective function requires time complexity of at least $O(n_1n_0)$, and the algorithm needs to search through a 2-dimensional grid of values. This also means that this procedure is unable to adjust for correlations between feature coordinates via off-diagonal entries of the bandwidth matrices, and is severely impacted if coordinates are not scaled. In addition, this method does not have a natural generalization to multi-class classification scenarios since the notion of difference between classes cannot be expressed by a single function.

### 4.3. Misclassification rate

A third bandwidth selection procedure, described by [4], is to consider the criterion of misclassification rate, a natural criterion in the context of classification. If we denote

$$d(x)=\arg\max_j P(j|x)=\arg\max_j p_jf_j(x) \tag{17}$$

where $p_j$ is the prior probability of being in group $j$, and $f_j(x)$ is the group specific density. Using this notation and suppose that there are a total of $T$ many groups, we can express the misclassification rate as follows:

$$\Delta = \sum_{i=1}^{T} p_i P(d(x) \neq i | x \in G_i) \tag{18}$$

In practice, however, it is not possible to precisely calculate $\Delta(h_1, ..., h_T)$ since the probabilities must be computed using $(f_1, ...f_T)$ over a variable region that is defined by $(\hat{f}_1, ..., \hat{f}_T)$. Hence, a common approach as illustrated by [4] is to consider an estimate of misclassification rate via misclassified sample proportion, calculated by:

$$\hat{\Delta}(h_1, ...h_T) = \frac{1}{N} \sum_{i=1}^{N} 1_{\hat{d}(x_i) \neq y_i} \tag{19}$$

Where $N = \sum_1^T n_i$ the sum of group specific sample sizes, and $\hat{d}(x)$ is the predicted group based on posterior probabilities estimated from sample proportions and kernel density estimators. In practice, just like any other smoothing problem, we will end up at the trivial solution, where all bandwidths are 0, due to overfitting. Thus, a commonly used method, as discussed in [3], is to calculate the estimated misclassification rate based on leave-one-out cross-validation, arriving at the following objective function:

$$\hat{\Delta}(h_1, ...h_T) = \frac{1}{N} \sum_{i=1}^{N} 1_{\hat{d}^{-i}(x_i) \neq y_i} \tag{20}$$

Where $\hat{d}^{-i}(x)$ is the delete $i^{th}$ group predictor. Without loss of generality, if we assume that observation $i$ belongs to group $T$, the we can write:

$$\hat{d}^{-i}(x) = \arg\max\{\hat{p_1}\hat{f}_1(x), ..., \hat{p_{T-1}}\hat{f_{T-1}}(x), \hat{p_T}\hat{f_T}^{-i}(x)\} \tag{21}$$

since observation $i$ is only part of the group specific density estimator for group $T$. The misclassification bandwidth selection procedure is advantageous is the sense that it can be used with any number of classes naturally, optimizing for arguably the most important criterion in most context. Similar to the second method, misclassification bandwidth selection also requires a search through a grid (of dimension $T$), resulting in massive computational work load if the number of competing classes is high. In fact this is the reason why we only consider bandwidth matrices of the simplest form, even though a more ideal objective function could have been $\hat{\Delta}(H_1, ...H_T)$, where $H_i$ are any possible symmetric matrices with positive entries. Again, the computational challenge also prevent this procedure from capturing correlation between feature coordinates via off-diagonal entries of the bandwidth matrices, while requiring scaled coordinates to work optimally. In addition, another noticeable disadvantage of this procedure is that multiple minimums can be identified, and that the estimated misclassification rate function is always piece-wise constant, even thought he true misclassification rate function may be smooth [3]. An interesting observation discussed in multiple works [3] [4] [7] is that the minimizer of misclassification rate tends to be larger than the minimizer for least square criterion discussed prior. Many suggests that this arises due to the fundamentally different objective function that the two methods are optimizing, meanwhile they claim that a relatively larger bandwidth for the sole purpose of classification (maximizing accuracy and reducing misclassification rate) may not be a bad thing.

## 5. Data analysis

In this section we compare the kernel classifier against other parametric models using data sets from the UCI machine learning repository. Due to computational challenge, we only consider the MISE minimization and misclassification rate bandwidth selection procedures. Further, we restrict ourselves to binary classification problem, and a manageable sample size. In addition, we also want to investigate the benefit of dimension reduction, since kernel density estimators are known to suffer from the curse of dimensionality, and so does the kernel based classifier as discussed in [1].

### 5.1. Sonar data

The sonar data set from the UCI machine learning repository features 60 coordinates and a binary response. The 60 coordinates indicate the reflection of sonar waves from different angles, while the binary response indicate whether the sonar bounced off a metal cylinder (1), or cylindrical rock (0). The data set contains a total

of 208 observations, where half are used for training and half are used for testing. To assess the impact of dimension reduction, we implemented linear discriminant analysis (LDA) and discovered that two discriminant eigenvectors are sufficient to encompass 99.9% of the variability between groups, thus the transformed data has only 2 coordinates. For testing, we consider normal reference bandwidth on the reduced data, and misclassification bandwidth in both the dimension reduced and unreduced data. In addition, we also compare the kernel classifier to logistic regression, LDA, and quadratic determinant analysis (QDA). Since logistic regression failed to converge using the 60 coordinates training set, and QDA requires a group-wise minimal sample size, all parametric models would be using the dimension reduced data.

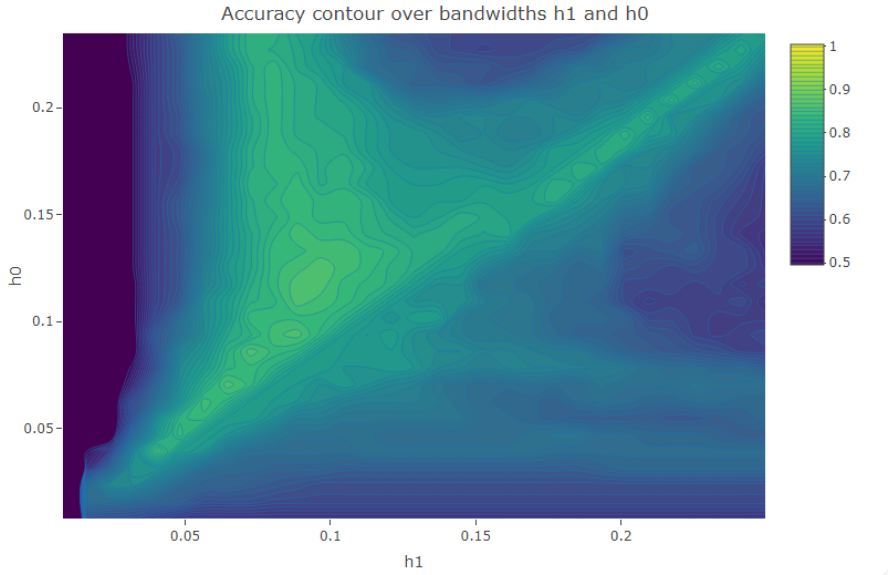


*Figure 1: cross validation accuracy contour of kernel density classifier on the sonar data*

Figure 1. shows the accuracy $(1 - misclassification\%)$ of the unreduced sonar data. Although multiple bandwidth combinations attained the maximum, it seems that the global maximum (givien more samples) would be located in the center; hence the bandwidth is chosen to be $(h_1, h_0) = (0.0886, 0.1252)$.

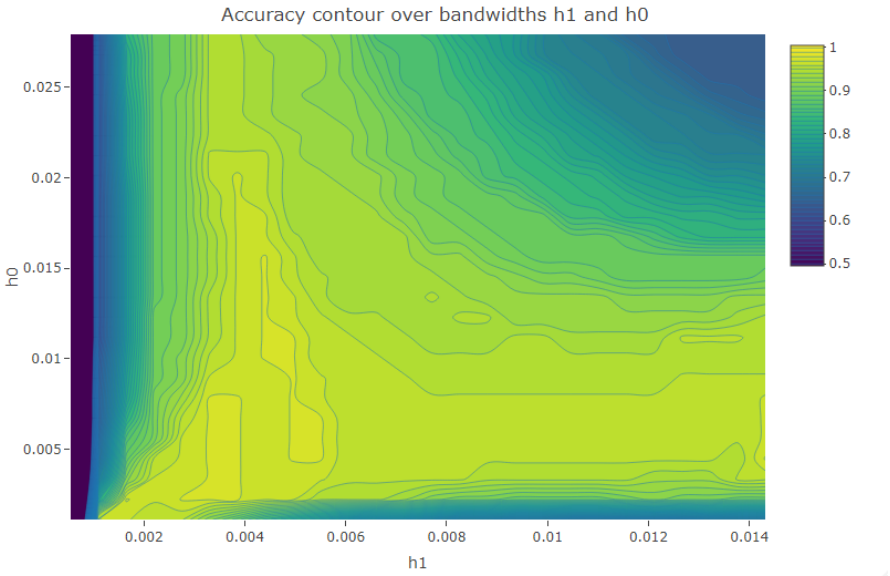


*Figure 2: cross validation accuracy contour of kernel density classifier on the dimension reduced sonar data*

Figure 2. shows the accuracy $(1 - misclassification\%)$ of the dimension reduced sonar data. Again, multiple bandwidth combinations attained the maximum, yet this time it is impossible to visually distinguish (with some

confidence) between a local max and a potential global max; hence, we are forced to random select one of the maximum, which turns out to be $(h_1, h_0) = (0.0044, 0.0112)$

### 5.2. Model assessment

To compare between the models, we use criterion such as the Reciever Operator Curve (ROC), area under curve (AUC), accuracy, positive prediction value (PPV), F1 score, and Mathew Correlation Coefficient (MCC).

The Reciever Operator curve is a plot in the FPR and TPR plane, where TPR and FPR refers to false positive rate and true positive rate respectively. Calculated as: $TPR = \frac{TP}{TP+FN}$, and $FPR = \frac{FP}{FP+TN}$, where (T) and (F) indicate True/False, and (P) and (N) indicate Positive/Negative. The ROC of a binary classifier indicates its change in FPR and TPR as threshold value of the classification rule changes. a perfectly random model will achieve the 45 degree line, while the higher a model's curve above the 45 degree line indicates higher robustness of the model.In this study, since all models return a probability value between 0 and 1, we consider threshold values from 0 to 1, in increments of 0.05. Figure 3. shows the ROC of all 7 models.

Positive Prediction Value (PPV) is a measure of the performance of a classifier when the predicted output is positive. PPV is calculated as $\frac{TP}{TP+FP}$, and is often considered as an essential aspect in biomedical context.

F1 score, and Mathew Correlation Coefficient (MCC) are overall performance indices of a binary classifier, calculated as: $F1 = \frac{2TP}{2TP+FP+FN}$, and $MCC = \frac{TP \times TP - FN \times FP}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$. F1 is bounded between 0 and 1, playing a similar role as accuracy, giving equal weights to FN and FP. MCC is bounded between -1 (worst) and 1 (best), taking into account the ratio of the four components of a confusion matrix, making it resistant to the effect of prevalence (the true proportion of positives in the sample).

Area under curve (AUC) is another measurement of the robustness of a binary classifier, defined as the definite integral of the ROC curve on the support [0,1]. Figure 4. shows the AUC, accuracy, PPV, F1, and MCC indices of the 7 models. Note that all indices besides AUC are calculated based on a threshold of 0.5.

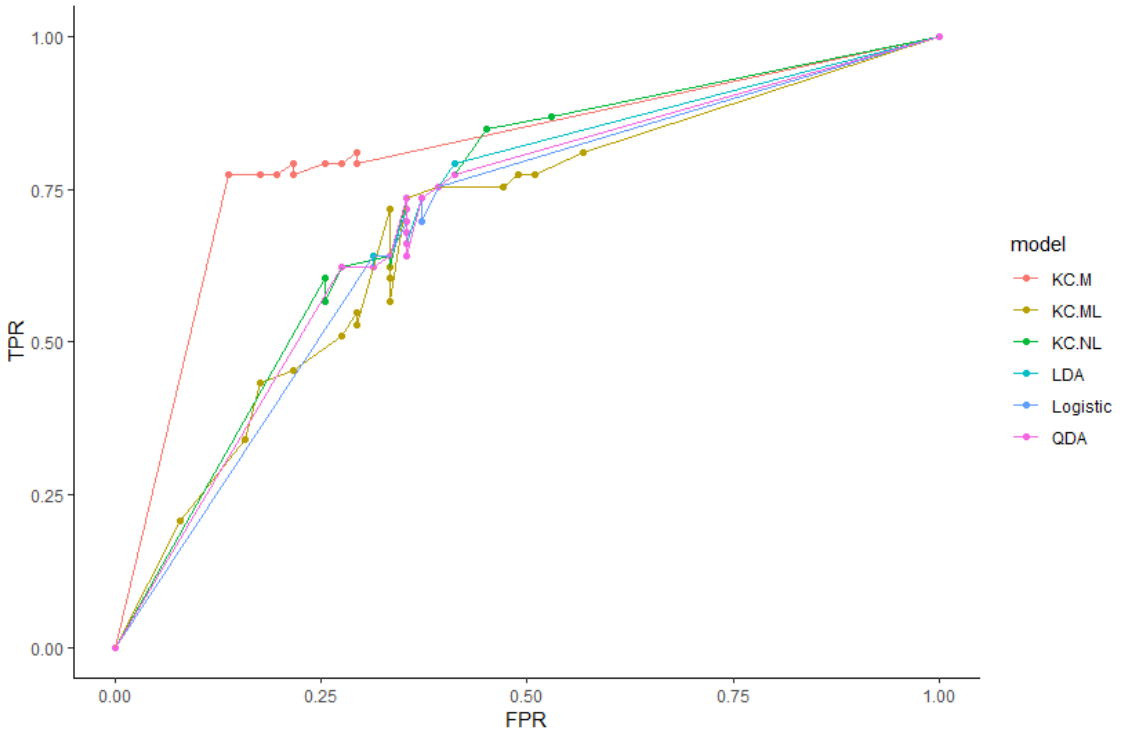


*Figure 3: ROC of the 7 models*

Figure 3. shows the ROC of all 7 models. KC indicates the kernel classifier, while M and N indicate the

bandwidth selection procedure (misclassification, or normal reference), and L indicate the presence of dimension reduction via LDA. The curves are rather rough due to small sample size of the test, hence we see massive overlap of TPR and TFR at various thresholds, yet all is not lost. From the curves, we can see that all three versions of the kernel based classifier are as least as good as the parametric models in terms of robustness, while misclassification bandwidth model on unreduced dimension shows the most competitive performance, achieving high TPR and low FPR at almost any given thresholds. From this graph, however we do not see the curse of dimensionality entering the play, since the performance of the unreduced model is in fact better.

| | AUC | ACC | PPV | MCC | F1 |
|---|---|---|---|---|---|
| KC.M | 0.815 | 0.788 | 0.792 | 0.577 | 0.792 |
| KC.ML | 0.685 | 0.635 | 0.653 | 0.271 | 0.627 |
| KC.NL | 0.716 | 0.673 | 0.673 | 0.346 | 0.685 |
| LDA | 0.694 | 0.673 | 0.673 | 0.346 | 0.685 |
| Logistic | 0.687 | 0.673 | 0.673 | 0.346 | 0.685 |
| QDA | 0.701 | 0.663 | 0.667 | 0.326 | 0.673 |

*Figure 4: Performance indices of the 7 models*

Figure 4. shows the model performance indices of all 7 models. The kernel based classifier using misclassification bandwidth on the unreduced data exhibit the most competitive performance, achieving a AUC of 0.815, in support of the ROC graph, and attaining significantly higher performance indices compared to other kernel classifiers and parametric models. To ensure that these findings are not a mere manifest of data selection bias, we also calculated bootstrapped version of some of the performance indices, retrieving both mean and standard error. Ideally, bootstrapping should include reasampling both a new training and test sets; however, this will incur a much greater computational burden (for bandwidth selection), and as we have seen, bandwidth selection may require visualization of the accuracy contour. Hence, we fix the training set in all runs to the present training set, and only vary the testing set, where we resample 150 observations from the sonar data in every run.

| | ACC | | PPV | | MCC | | F1 | |
|---|---|---|---|---|---|---|---|---|
| | mean | se | mean | se | mean | se | mean | se |
| KC.M | 0.891 | 0.00166 | 0.895 | 0.00230 | 0.782 | 0.00334 | 0.898 | 0.00165 |
| KC.ML | 0.804 | 0.00171 | 0.822 | 0.00256 | 0.606 | 0.00348 | 0.814 | 0.00169 |
| KC.NL | 0.830 | 0.00176 | 0.830 | 0.00247 | 0.658 | 0.00359 | 0.843 | 0.00163 |
| LDA | 0.830 | 0.00176 | 0.830 | 0.00247 | 0.658 | 0.00359 | 0.843 | 0.00163 |
| Logistic | 0.821 | 0.00178 | 0.827 | 0.00251 | 0.640 | 0.00364 | 0.833 | 0.00170 |
| QDA | 0.820 | 0.00176 | 0.827 | 0.00251 | 0.638 | 0.00360 | 0.832 | 0.00167 |

*Figure 5: Bootstrapped performance indices of the 7 models*

Figure 5. shows the bootstrapped performance indices (mean and standard error) of the models. The result is in support of the initial study (figure 4.), where we see the kernel classifier using misclassification bandwidth on the unreduced data coming on top in all 4 performance indices, achieving an average accuracy of 0.891, PPV of 0.895, MCC of 0.782, and F1 score of 0.898, significantly higher than the remaining models. In addition, the kernel classifier using normal reference bandwidth on the reduced data display performance comparative to parametric models, meaning it is also a very viable option in this setting. It is worth noting that to our surprise the kernel classifier using misclassification bandwidth on the reduced data ended up with the worst performance indices amongst all 7 models, which was unexpected since dimension reduction was argued to be beneficial for kernel classifiers [1]. From this study, we have not seen the impact of the curse of dimensionality as discussed in other works. We hypothesize that this could have been a result of bandwidth selection, since for the dimension reduced model we were unable to identify a region for the global maximum from the contour, and hence had to choose the bandwidths randomly from existing maximums.

## 6. Conslusion

To sum up, in this report we started with an examination of the problem of binary classification, and discussed the limitations of existing parametric models. From there, we introduced the kernel classifier that is based on a combination of the theory posterior probability, and multivariate kernel density estimator. The resulting classifier is one that is free of underlying distributional assumptions, can be generalized to multi-class settings, and possesses the property of consistency. Like any other non-parametric models, the choice of the smoothing parameter is also essential in this context, and thus we provided a detailed overview of 3 of the well-known procedures, discussed their strength and weaknesses; Specifically, we examined the MISE minimization criterion (choosing bandwidths independently , and minimizing the MISE of each densities), between group density cross-validation (minimizes the $L^2$ distance between the true boundary function and estimated boundary function), and misclassification rate cross-validation (minimizes the misclassification rate calculated via leave-one-out cross-validation).

Finally, we applied the kernel classifier to the sonar data from the UCI machine learning repository, considering 2 bandwidth selection procedures and dimension reduction, and compared it to traditional parametric models. From the accuracy contours, we encountered the very practical problem of multiple maximums as discussed in [3]. During model testing, performance indices suggested that amongst the 3 variants of kernel classifiers, 1 performed significantly better than all parametric models, 1 performed comparative to parametric models, and 1 performed slightly worse than all parametric models, possibly due to a sub-optimal choice of bandwidths. Hence, we can conclude that the kernel classifier, due to its flexibility and performance, is indeed a contender in the realm of classification (at least for binary classification), but just like any other non-parametric model, it is plagued with the problem of efficient and optimal bandwidth selection.

## References


[1] CRAIG A. COOLEY and STEVEN N. MACEACHERN. "Classification via kernel product estimators". In: *Biometrika* 85.4 (Dec. 1998), pp. 823–833. ISSN: 0006-3444. DOI: 10.1093/biomet/85.4.823. eprint: http://oup.prod.sis.lan/biomet/article-pdf/85/4/823/698620/85-4-823.pdf. URL: https://doi.org/10.1093/biomet/85.4.823.

[2] Jarek Duda. "Gaussian Auto-Encoder". In: *CoRR* abs/1811.04751 (2018). arXiv: 1811.04751. URL: http://arxiv.org/abs/1811.04751.

[3] Anil K Ghosh, Probal Chaudhuri, and Debasis Sengupta. "Classification Using Kernel Density Estimates". In: *Technometrics* 48.1 (2006), pp. 120–132. DOI: 10.1198/004017005000000391. eprint: https://doi.org/10.1198/004017005000000391. URL: https://doi.org/10.1198/004017005000000391.

[4] Anil Ghosh and Probal Chaudhuri. "Optimal smoothing in kernel discriminant analysis". In: *Statistica Sinica* 14 (Apr. 2004), pp. 457–483.

[5] Peter Hall and Matthew P. Wand. "On Nonparametric Discrimination Using Density Differences". In: *Biometrika* 75.3 (1988), p. 541. DOI: 10.2307/2336605.

[6] Trevor J Hastie, Jerome H Friedman, and Robert J Tibshirani. *The elements of statistical learning*. Springer, 2017, Chapter 6, Chapter 6.

[7] Charles Taylor. "Classification and kernel density estimation". In: *Vistas in Astronomy* 41.3 (1997). From Information Fusion to Data Mining, pp. 411–417. ISSN: 0083-6656. DOI: https://doi.org/10.1016/S0083-6656(97)00046-9. URL: http://www.sciencedirect.com/science/article/pii/S0083665697000469.