Advances and Applications in Mathematical Sciences
Volume 21, Issue 5, March 2022, Pages 2437-2457

A COMPARATIVE RETROSPECTION OF STATISTICAL CLASSIFICATION TECHNIQUES ON FISHER'S IRIS DATA

BODHODITYA BARMA and SARAN ISHIKA MAITI

Department of Statistics
Visva-Bharati University
Santiniketan, India

Abstract

In multivariate data analysis, Fisher's linear discriminant analysis is pretty common tool for classification of late, with the rise of machine learning techniques, advanced nonlinear classification tools like kernel Fisher discriminant Analysis, quadratic discriminant analysis etc. have also been introduced. In this article we try to revisit Fisher's trailblazing experiment of statistical classification on Irish data (1936), where measurement are recorded on four distinct variables of three species of Iris genus. We take a discourse to this classical experiment through several linear/nonlinear classification techniques, along with a comparative documentation on misclassification errors. Also we venture on the segregation of species via four variables, bunched in possible combinations, viz., taking single variable/two variables/three variables/four variables in construction of classification rule.

1. Introduction

Statistical Classification is a methodology where we endeavour to categorize a random observation under a specified class out of given number of classes. The main goal of a classification problem is to identify a category or class to which a new observation can be assigned to. An algorithm that enacts classification, especially in a concrete implementation, is known as a classifier. The term classifier refers to the mathematical function, validated by a classification algorithm that maps the input data to a prefixed category. The entire idea of classification theory originates from statistical learning mechanism where the system receives data (observations) as input and

2020 Mathematics Subject Classification: 62H30, 62P10.
Keywords: Fisher's linear discriminant function; Generalized discriminant analysis; Kernel function; Kernel discriminant analysis; Statistical classification.
Received June 16, 2021; Accepted September 8, 2021

outputs a function that can be used to predict some features of future data. So borrowing upon the idea of statistical learning theory, given, say a data set with two classes, the quest of classification is to label the best set of features of the classes in order to discriminating between the two classes.

Statistically speaking, in classification theory we would search for a subspace (or sub-manifold) which separates the classes as much as possible while the data become as spread as possible. In computer sciences, support vector machine (SVM) (Vapnik [13]) works on almost same mechanism as statistical learning theory. In SVM, an optimal separating hyper plane is searched one subset of training samples, namely, the support vectors. However in particular case, this separating plane may be considered as a single dimension separating line, viz., $f(\boldsymbol{x}) = \mathbf{w}'\boldsymbol{x} + \boldsymbol{b}$. The input vector $\boldsymbol{x}$ classifies to the first class if $f(\boldsymbol{x}) > 0$ and $\boldsymbol{x}$ falls to second class if $f(\boldsymbol{x}) < 0$. In statistics such a function is called a linear discriminant function (LDA). But it should be kept in mind that linear discriminant analysis produces the optimal discriminant linear functions when each of the labeled class data are assumed to be distributed normally. On the other hand, in SVM there is no such distributional assumption on the data. The seminal work in statistical learning theory is the inception of linear discriminant analysis which was first disseminated by Sir. Ronald Aylmer Fisher.

In order to cater an objective of linear separation of two classes, Fisher (1890-1962) who adopted the idea of between class variance maximization. One can extend the two class Fisher's discriminant function in categorizing k classes as well. But this traditional Fisher's method fails in case of a nonlinear class separability. To extract the nonlinear discriminant features, Mika et al. (1999) used kernel function as a class separable tool and hence introduced a new area termed kernel discriminant analysis. A little later, taking the cue from SVM technique, Baudat et al. [4] improvised kernel method to Generalized Dicriminant Analysis (GDA) which provides a mapping of the input vectors into high dimensional feature space. Although the idea of discriminant analysis was set forth by R. A. Fisher the development of the same has been burgeoned by the computer scientists.

Fisher's path breaking paper, on linear discriminant analysis [7] introduced the well-known Iris flower data set to the audience. Fisher's work

concentrates on the supervised classification methods for different morphologic variations of three related species of Iris flowering plants, namely, Iris setosa, Iris versicolor and Iris virginica, where the separation of the species have been done through four morphological characters (variables). Even though, Fisher's formative article harboured a novel statistical idea, stemmed from Iris flower set, till now no significant works have been invoked towards the further extensive classification study on the said data set.

This article delivers a comparative testimony of the several statistical classification techniques exerted on Iris data. Apart from the usual traditional linear discriminant analysis here we include quadratic discriminant analysis from parametric classification methods. Additionally, as an advanced nonparametric classification tools, we execute Kernel Fisher's Discriminant Analysis (KFDA) and Kernel Discriminant Analysis (KDA) on the data. Not only the classification method employed on all variables (as done in Fisher's well acclaimed paper) but also we perform separation, bunching out every possible combination of variables, e.g., taking two variables one at a time, three variables one at a time and finally the entire group of variables. In each separative method, done by the virtue of several combinations of variables, we report misclassification errors as well.

We believe that this way of critically re-excavating Iris data from the standpoint of several discriminant analysis schemes would train unlabeled information, (hidden in the data), to labeled classes which helps to understand the inter relationship of the flowering variable in identifying the species. Also, this effort would inspire budding researchers to introspect any data oriented classification problem through the traditional statistical techniques rather not hurling ideas from machine learning tools.

This short article is organized as follows. A primary description on Fisher's Irish data is furnished in section 2. Section 3 briefly discusses the theories of separation techniques which are used. Section 4 unravels the analysis and exploration on Iris data with the corresponding misclassification errors. Finally, section 5 concludes this article with few directions to future studies.

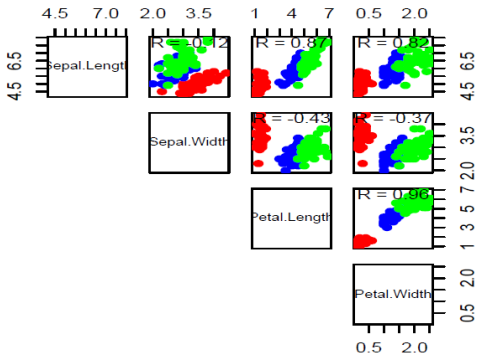
2. Brief Description of Iris Data

Iris, named after the Greek Goddess, is easy-to-grow perennial plants

with elegant, colorful flowers. It grows during early summer in northern hemisphere zones, spreading across Europe, northern Asia and northern America, specially where the climate is cold, dry and semi-desert. There are some 300 species in the genus Iris. These distinctive, six-petaled flowers have three outer hanging petals (called "falls") and three inner upright petals (called "standards").

Fisher, in his pioneering paper," The use of multiple measurements in taxonomic problems" (1936), considered measurements of the flowers of fifty plants on each of the three species I. setosa, I. versicolor and I. virginica, found growing together in the same colony. For each of the flower, four flowering measurements, viz., sepal length (X_1), sepal width (X_2), petal length (X_3) and petal width (X_4) are furnished in the Fisher's data. All measurements are given in centimeters.

As a beginner's stepping stone, first we furnish pairs plot which allows to visualize the distribution of single variable as well as relationship between two variables through the matrix of association. Three colors (red, green and blue) are used for setosa, versicolor, and virginica respectively. The figure below displays possible two-dimensional projections of multidimensional data (in this case, four dimensional). The plot projects the distinctiveness of each feature across the three species so that most of the variables could be used to predict the species. It is pretty clear that I. setosa (cluster of red dots) distinguishes itself from the other two by dint of those four variables. Also I. versicolor (blue) and I. virginica (green) are to some extent alike on the basis of the same.



Also Person's product moment correlation for the pairs of four variables

are recorded on the top of each cell in pair matrix plot. The correlation exhibit the degree of association among the variables e.g., petal width (x_4) and petal length (x_3) are strongly positively associated whereas weak negative correlation is present between sepal length (x_1) and sepal width (x_2).

Since the scatter matrix (pairs plot) reveals internal association within the variables whether weak or strong, an intuitive guess may be developed to introspect on the separation of species by means of two variables, three variables and finally by the set of four variables.

3. Descriptions of the Techniques Adopted

In effort to classify the species of Iris data through the successive bunching of four variables we employ four different types of techniques- Linear Discriminant Analysis (LDA), Kernel Fisher's Discriminant Analysis (KFDA) or Generalized Discriminant Analysis (GDA) and Kernel Discriminant Analysis (KDA) among which LDA and QDA methods are parametric while KFDA / GDA and KDA methods are the nonparametric methods of classification. For each method, to gauge errors committed we report misclassification rate (probability).

Definition 1. Misclassification Rate. In statistical terminology, 'misclassified' explains that one object coming from one class gets assigned to a different class by means of a classification technique. Misclassification rate is calculated using the following formula.

$$\text{Misclassification rate } = \frac{b}{n} = 1 - \frac{a}{n}$$

where $n =$ total number of object to be classified, $a =$ number of object classified correctly and $b =$ number of objects. Quite intuitively lower the misclassified error better is the classification rule.

The following subsections sketch a brief theoretical discussion on each methods of classification, adopted here.

Linear Discriminant Analysis (LDA) 3.1. Suppose we have N classes and $i-th$ class denoted by the class level π_i where $i = 1, 2, \ldots, N$. The objects are classified on the basis of q associated random variables

$\boldsymbol{X}' = [X_1, X_2, \ldots, X_q]$. The observed values of $\boldsymbol{X}$ differ from one class to another f_i and p_i, $i = 1, 2, \ldots, N$ denoted the probability density function and the prior probability respectively of the i^{th} class. In case of nonavailability of prior probabilities, p_i's are considered equi-probable for each class. Let $\boldsymbol{X}$ be the observation to be assigned among any of N classes. In LDA the parametric assumption to be undertaken is that probability density function for any i^{th} family is multivariate Gaussian, i.e. $f_i(x) \sim N(\mu_i, \Sigma_i)$, $i = 1, 2, \ldots, N$ where $\boldsymbol{\mu}_i$ and Σ_i are the mean vector and covariance matrix correspond to i^{th} class. Consider that, the covariance matrices are equal for all class which is Σ, $\Sigma = \Sigma_1 = \Sigma_2 = \ldots = \Sigma_n$. The linear discriminant scores are thereby calculated through the following formula

$$d_i(\boldsymbol{x}) = \hat{\boldsymbol{\mu}}_{ii}' \Sigma^{-1} \boldsymbol{x} - \frac{1}{2} \hat{\boldsymbol{\mu}}_{ii}' \Sigma^{-1} \hat{\boldsymbol{\mu}}_i' + \ln p_i, \quad i = 1, 2, \ldots, N.$$

A specific observation $\boldsymbol{x}$ will be allocated to the π_g if the linear score $d_g(\boldsymbol{x}) = \text{Largest}(d_1(\boldsymbol{x}), d_2(\boldsymbol{x}), \ldots, d_N(\boldsymbol{x}))$.

In general the mean vector $\boldsymbol{\mu}_i$ and covariance matrix Σ_i are unknown. For classification of training data set, the mean vectors and covariance matrices for each class are estimated from training sample date. $\hat{\mu}_i$ and $\hat{\Sigma}_i$ are the sample mean vector and sample covariance matrix of the i^{th} class and Σ is replaced by the pooled estimate Σ_{pooled}.

$$\Sigma_{pooled} = \frac{1}{n_1 + n_2 + \ldots n_N - N}((n_1 - 1)\hat{\Sigma}_2 + \ldots + (n_N - 1)\hat{\Sigma}_N).$$

p_i is replaced by the sample proportion $\hat{p}_i = \frac{n_i}{n}$, where $n = \sum_{i=1}^{N} n_i$. The estimated linear discriminant score, then $\hat{d}_i(\boldsymbol{x})$ is given by

$$\hat{d}_i(\boldsymbol{x}) = \hat{\boldsymbol{\mu}}_i' \Sigma^{-1} \boldsymbol{x} - \frac{1}{2} \hat{\boldsymbol{\mu}}_i' \Sigma_{pooled}^{-1} \hat{\boldsymbol{\mu}}_i' + \ln p_i, \quad i = 1, 2, N.$$

x will be allocated to the π_g if the linear score $\hat{d}_g(\boldsymbol{x}) = \text{Largest}$ $(\hat{d}_1(\boldsymbol{x}), \hat{d}_2(\boldsymbol{x}), \ldots, \hat{d}_N(\boldsymbol{x}))$. Fisher proposed an extension of LDA where

discrimination of the population can be done by taking few linear combination of the observed variables, viz., $\boldsymbol{\alpha}_1'\boldsymbol{X}, \boldsymbol{\alpha}_2'\boldsymbol{X}, \boldsymbol{\alpha}_3'\boldsymbol{X}, \ldots$ The advantages of this method are dimension reduction, visual inspection of the population groups and also helping to find any abnormalities in the data by plotting against the first two discriminants.

The Fisher's criterion is the maximization of the following ratio with respect to $\boldsymbol{\alpha}$,

$$J(\boldsymbol{\alpha}) = \frac{\boldsymbol{\alpha}' B \boldsymbol{\alpha}}{\boldsymbol{\alpha}' V \boldsymbol{\alpha}'},$$

where B is the between class sum of square, $B = \sum_{i=1}^{N} (\boldsymbol{\mu}_i - \bar{\boldsymbol{\mu}})(\boldsymbol{\mu}_i - \bar{\boldsymbol{\mu}})'$ where, $\bar{\boldsymbol{\mu}} = \sum_{i=1}^{N} \boldsymbol{\mu}_i$ and V is the within class sum of square, $V = \sum_{i=1}^{N} (n_i - 1) \sum_i$. $\boldsymbol{\alpha}$'s are taken as the eigenvectors of the $V^{-1}B$ matrix which maximize the ratio $J(\boldsymbol{\alpha}) \cdot \lambda_1, \lambda_2, \ldots, \lambda_s$, where $s = \min(N-1, p)$, p is the number of variables, are the eigenvalues $V^{-1}B$. then $\boldsymbol{\alpha}_1'\boldsymbol{x}$ is the first discriminant, whereas $\boldsymbol{\alpha}_2'\boldsymbol{x}$ being the second discriminant and so on. In linear discriminant plot, LD1 and LD2 denote the first and second discriminant respectively.

Quadratic Discriminant Analysis (QDA) 3.2. Similar to LDA, in Quadratic discriminant analysis method, probability density function i^{th} class is also considered as the normal density.

$$f_i(\boldsymbol{x}) \sim N(\boldsymbol{\mu}_i, \sum_i), \quad i = 1, 2, \ldots, N,$$

where $\boldsymbol{\mu}_i$ and $\sum_i$ are the mean vector and covariance matrix correspond to i^{th} class. But, in this method $\sum_i' s$, $i = 1, 2, \ldots, N$ are not assumed to be equal for all classes. The quadratic discriminant scores are calculated by the following formula

$$d_i^Q(\boldsymbol{x}) = -\frac{1}{2} \ln |\sum_i| - \frac{1}{2} (\boldsymbol{x} - \boldsymbol{\mu}_i)' \sum_i^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_i) + \ln p_i \quad i = 1, 2, \ldots, N.$$

$\boldsymbol{x}$ will be allocated to the π_g if the quadratic score $d_g^Q(\boldsymbol{x})$ = Largest of $d_1^Q(\boldsymbol{x}), d_2^Q(\boldsymbol{x}), d_N^Q(\boldsymbol{x})$. Similar to LDA, the mean vector and covariance matrix for i^{th} are unknown and need to be estimated from train sample data set. In

case of nonavailability of true discriminant score, the estimated quadratic discriminant score, $\hat{d}_i^Q(x)$ might be given by

$$\hat{d}_i^Q(\boldsymbol{x}) = -\frac{1}{2}\ln|\Sigma_i| - \frac{1}{2}(\boldsymbol{x}-\hat{\boldsymbol{\mu}}_i)'\hat{\Sigma}_i^{-1}(\boldsymbol{x}-\hat{\boldsymbol{\mu}}_i) + \ln\hat{p}_i \quad i = 1, 2, \ldots, N,$$

where $\hat{\boldsymbol{\mu}}_i$, $\hat{\Sigma}_i$ and the estimates.

Kernel Fisher's Discriminant Analysis (KFDA) 3.3. Kernel Fisher's discriminant Analysis is a kernelized version of Linear discriminant Analysis where kernel function is taken as Gaussian kernel in order to perform nonlinear mapping on input data set to the high dimensional feature space with linear properties, i.e.,

$$\phi : R^n \to F \Rightarrow x \to \phi(x)\ \forall x,$$

where ϕ being the mapping function. In the feature space classes are emerged as linearly separable classes (Baudat et al. [4]) Note that the mapped observations are centered in the feature space (Schölkopf et al. [11]). According to the Fisher's classification criterion, maximizing the intra-class inertia and minimizes the within-class inertia would produce the following ratio measuring the variability between groups values to common variability within group values in feature space,

$$\frac{\boldsymbol{v}'V\boldsymbol{v}}{\boldsymbol{v}'B\boldsymbol{v}}, \tag{3.1}$$

where V and B are the following intra-classes inertia and inter-classes inertia in the feature space. We can select a $\boldsymbol{v}$ to maximize the ratio. The eigenvector of the largest eigenvalue of $B^{-1}V$ gives the maximum of the above ratio. As because the eigenvectors are linear combinations of feature elements, there exist coefficients $\alpha_{pq}, (p = 1, 2, \ldots N, q = 1, 2, \ldots, n_p)$ for which

$$\boldsymbol{v} = \sum_{p=1}^{N}\sum_{q=1}^{n_p}\alpha_{pq}\phi(x_{pq}) \tag{3.2}$$

Due to the high dimensionality structure directly solving (3.1) is difficult.

In order to address this, dot product kernel $k(x_i, x_j)$ is used on Hilbert space (Aizerman et al. [2], Boser et al. [3]) which can be proposed as follows.

$$k(x_i, x_j) = k_{ij} = \Phi'(x_i)\Phi(x_j) \tag{3.3}$$

Dot product kernel requires normalization of the data required, i.e., mean zero and equal variance. In terms of the dot product kernel, (3.1) can be written as,

$$\lambda = \frac{\boldsymbol{\alpha}'KWK\boldsymbol{\alpha}}{\boldsymbol{\alpha}'KK\boldsymbol{\alpha}}. \tag{3.4}$$

Where K is the kernel matrix, $K = (K_{pq})_{p=1, 2, \ldots, N, q=1, 2, N}$ where $K_{pq} = (K_{ij})_{i=1, 2, \ldots, n_p, j=1, 2, \ldots, n_q}$ and W is the block diagonal matrix, $W = (W_l)_{l=1, 2, \ldots, N}$ where W_l the $(n_l \times n_l)$ matrix whose all terms are equal to $\frac{1}{n_l}$. Using eigenvalue decomposition of kernel matrix, $\boldsymbol{\alpha}$ can be derived maximizing (3.4). For detailed derivation of the process, the readers are recommended to see Boudat et al. [4]. The projection of the test observation $\boldsymbol{z}$ on feature space can be expressed as

$$\boldsymbol{v}'\phi(z) = \sum_{p=1}^{N}\sum_{q=1}^{n_p}\alpha_{pq}k(x_{pq}, \boldsymbol{z}). \tag{3.5}$$

Kernel Discriminant Analysis (KDA) 3.4. Kernel discriminant analysis is a broader concept which is an extension of Bayesian classification rule. Suppose we have N populations (classes) and the class levels are denoted by π_i. Each population is associated with the probability density function f_i and the prior probability of the i^{th} class is p_i. We have an unknown test point x and we are to allocate an unknown test point $\boldsymbol{x}$ to one of those populations. The Bayes discriminant rule is to allocate $\boldsymbol{x}$ to the π_j^0 class if $\pi_j^0 = \arg\max_{i\in\{1, 2, \ldots N\}} p_i f_i(\boldsymbol{x})$.

The Kernel Discriminant Rule (KDR) is the Bayes discriminant rule just via replacement of f_i by its kernel density estimates

$$\hat{f}_i(x; H_i) = n_i^{-1} \sum_{j=1}^{n_i} K_{H_i}(x - X_{ij})$$

and p_i is replaced by the sample proportion $\frac{n_i}{n}$, where $n = \sum_{i=1}^{N} n_i$. The kernel function $K_H(x)$ is the probability density function and H is the chosen bandwidth matrix which is 7 symmetric and positive definite. The commonly used kernel functions are Gaussian, uniform, triangular, Epanechnikov, etc. Evidently the choice of bandwidth H plays crucial role for the performance of kernel density estimates. Optimal bandwidth can be selected by using the Mean Integrated Squared Error (MISE) criterion

$$MISE(H) = E \int_{\Re^d} [\hat{f}(\boldsymbol{x}; H) - f(\boldsymbol{x})]^2 d\boldsymbol{x}. \tag{3.6}$$

The kernel discriminant rule of classification is hence proposed as to allocate $\boldsymbol{x}$ to the π_j^0 class if $\pi_j^0 = \arg\max_{i \in \{1, 2, \ldots, \}} \hat{p}_i \hat{f}_i(\boldsymbol{x}, H_i)$.

Usually, standard normal kernel function is employed for estimating density function f_i, where

$$K(x) = \frac{1}{(2\pi)^{\frac{d}{2}}} \exp\left(-\frac{1}{2} x^T x\right),$$

d being the dimension of the variable $\boldsymbol{x}$.

The bandwidth matrix is selected by using the method of Smoothed Cross Validation (SCV) (Hall, Marron and Park [9]). The SCV is defined by starting with modified version of least square cross validation. HSCV is the minimizer of

$$SCV_g(h) = (nh)^{-1} \int_{\Re} K^2 + \hat{B}_g(h),$$

where

$$\hat{B}_g(h) = \frac{1}{n(n-1)} \sum \sum_{i<j} \{(K_h * K_h - 2K_h + K_0) * K_g * K_g\}(X_i - X_j).$$

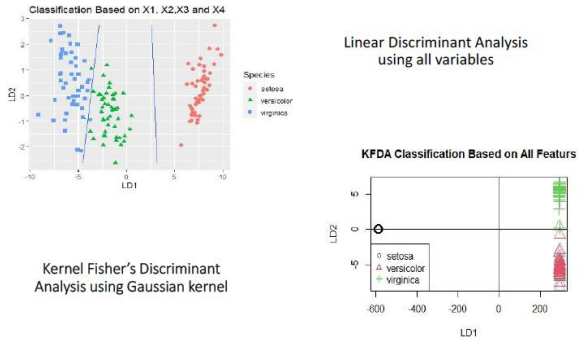
K_0 is the Dirac delta function, K_h kernel function with bandwidth h and K_g the possibly different kernel functions with bandwidth g.

4. Analysis and Exploration on Iris Data Set

Linear Discriminant Analysis (LDA), Kernel Fisher's Discriminant Analysis (KFDA), Quadratic Discriminant Analysis (QDA), and Kernel Discriminant Analysis (KDA) methods are performed on the entire Iris data set by means of four morphological variables X_1, X_2, X_3 and X_4. For each case, misclassification rate tables are reported along with the classification plots.

- classification by taking **each** of the variable $\{X_1, X_2, X_3, X_4\}$.
- classification by taking **pair** of variables $\{X_1X_2, X_1X_3, X_1X_4, X_2X_3, X_2X_4, X_3X_4\}$.
- classification by taking **three** variables one at the time $\{X_1X_2X_3, X_1X_2X_4, X_2X_3X_4, X_1X_3X_4\}$.
- finally classification by taking **four** variables together $\{X_1, X_2, X_3, X_4\}$.

Misclassification errors are given in the format of fraction. They may be otherwise, re-expressed in percentage format. Moreover, which classifying rule works best is also enlisted in the misclassification table. Clearly, each classification method would provide $\binom{4}{1}+\binom{4}{2}+\binom{4}{3}+\binom{4}{4} = 4 + 6 + 4 + 1 = 15$ plots. Therefore, for the four aforementioned techniques one can expect $15 \times 4 = 60$ classification plots altogether. Due to space limitation each and every plot is not furnished in this article. As a thrusting spirit, we present only the separating plot of Linear Discriminant (LDA) function as well as Fisher's Kernel Discriminant Analysis (KDA) taking **all** features (variables) in formulating classification rule. In contrast to LDA by four variables, KFDA using Gaussian kernel successfully separates out three species where I. setosa is way apart from the other two species. To be more specific, I. versicolor and I. virginica behave almost alike to each other but still, KFDA analysis skillfully manages to figure out a separating line in between them.



To keep the visual acceleration up, all of the plots are placed in Appendix. Note that, since KDA involves estimating density functions we can portray KDA separative plot up to three dimensional structure. Beyond three dimension drawing plot is impossible.

Few Technical Details 4.1. The entire exercise is executed by *R* version 4:1:0. In LDA and QDA analysis we use the package 'MASS' while for KFDA and KDA, the package 'KFDA' and 'KS' are used respectively. In KFDA and KDA we use Gaussian kernel, $k(x_i, x_j) = \exp\left(-\frac{\| x_i - x_j \|^2}{\sigma}\right)$, where the scale parameter (σ) is chosen as 07. In contrast to KFDA, in KDA we consider sample proportions as the estimates of prior probabilities while bandwidth is selected by smoothed cross validation technique. For calculating misclassification error under KDA we use compare.kda.cv function from the package KDA. R codes used for the analysis are available in Github repository (https://github.com/Bodhoditya/Iris-R-Codes.git).

Analysis and Discussion 4.2.

4.2.1. Case 1

The classification of four species is done by each of $\{X_1, X_2, X_3, X_4\}$. Misclassification errors rates are recorded in the following table. For each row(variable), minimum entry is marked by an asterisk $(*)$.

Table 1. Misclassification rate taking single variable for classification.

Variable taken as classifier	LDA	KFDA	QDA	KDA
X_1	0.2533*	-	0.2733	0.2533*
X_2	0.4467	-	0.4467	0.4133*
X_3	0.0533	-	0.4467*	0.0467*
X_4	0.0400*	-	0.0400*	0.0400*

For single variable classification KFDA is not possible. The other three methods run successfully. Misclassification rate is maximum uniformly in every method while separation is done by sepal width (X_2) of Iris flower while least misclassification error occurs uniformly in classification by petal width (X4). KDA appears as the best method retaining minimum classification error in all categories over LDA and QDA.

4.2.2. Case 2

Next we club any two variables as the key to form a classification rule. The possible combinations are

$$\{X_1X_2, X_1X_3, X_1X_4, X_2X_3, X_2X_4, X_3X_4\}.$$

The misclassification error rate table is given below. Asterisk $(*)$ means for the minimum entry.

Table 2. Misclassification rate taking pair of variables for classification.

Variable taken as classifier	LDA	KFDA	QDA	KDA
X_1, X_2	0.2000	0.1867*	0.2000	0.2067
X_1, X_3	0.0	0.0	0.0	0.

	330*	333*	400	0400
X_1, X_4	0.0 400	0.0 330*	0.0 330*	0. 0400
X_2, X_3	0.0 467	0.0 400*	0.0 467	0. 0533
X_2, X_4	0.0 330*	0.0 467	0.0 467	0. 0600
X_3, X_4	0.0 400	0.0 330	0.0 200*	0. 0267

The misclassification rate is minimum for LDA when sepal length (X_1) and sepal width (X_2) are combined and the other methods are performing neck by neck.

4.2.3. Case 3

Now we are considering three features at a time, and the possible combinations are $\{X_1X_2X_3, X_1X_2X_4, X_2X_3X_4, X_1X_3X_4\}$. The misclassification error rate table given below.

Table 3. Misclassification rate taking three variables together for classification.

Variable taken as classifier	LDA	KFDA	QDA	KDA
X_1, X_2, X_3	0.0 333	0.0 267*	0.0 530	0.0533
X_1, X_2, X_4	0.0 400	0.0 200*	0.0 400	0.0733
X_1, X_3, X_4	0.0 267	0.0 200*	0.0 200*	0.0267
X_2, X_3, X_4	0.0 200	0.0 133*	0.0 267	0.0400

Prominently, as the number of variables involved in classifier goes up the performance of KFDA grows better in terms of misclassification error over

the other techniques.

4.2.4. Case 4

Ultimately, we involve all four features together in shaping up a classification rule for separating the three Iris species. The misclassification error rate table is given below.

Table 4. Misclassification rate taking four variables together for classification.

Variable	LDA	KFDA	QDA	KDA
$X_1, X_2, X_3,$	0.0200	0.0067*	0.0200	0.0330

In terms of misclassification error KFDA secures the lowest value, only a mere .0067, even though LDA hits closest to KFDA. This is quite persuasive as the underlying distribution of Iris data is more or less predominated by multivariate Gaussian distribution. Note that, the multivariate Shapiro-Wilk test on Iris data yields p-value 0.023(R package *mvnormtest*). Also, Mardia's Multivariate normal test based on multivariate skewness and kurtosis ensures the multivariate normality of iris data at both 1% and 5% significance level. Furthermore, KFDA, defined via Gaussian kernel function, would most efficiently handle the non-linearly separable variables in input space by having a transformation to the high dimensional, linearly separable feature space.

5. Conclusion

In this article, we effort to dissect the 'famous' Iris data through the light of few potential statistical classification techniques. The motivation of this discourse is to introspect the alikeness and incongruity of three species of Iris flower, captured by dint of four distinguished variables, chosen in single, in pairs, in threes and ultimately taking all four. Although LDA is computationally easier it fails to discriminate three species exclusively by the linear combination made on two variables, three variable and four variables cases. Specifically, it manages to part I. setosa from the other two in each case but falls short in separating I. versicolor and I. virginica. Partition plot, done on QDA, affirms below average effectiveness in separative analysis. In

contrast, kernel density analysis defined through Gaussian weight function, displays fairly distinct separation among three species. Separation by KDA is visually reflected by 2D-contour plots and 3D-perspective cubes (vide Appendix). Classification plot based on KDA can only be possible for one variable case (1-D), two variables case (2-D) and three variables case (3-D). Beyond three variables, drawing plot is not possible. Kernel Fisher's Discriminant Analysis (KFDA) with Gaussian kernel emerges as the most efficient classifier for Iris data set. The more the inclusion of variables, better is the performance of classification of KFDA. For instance, KFDA plot on three variables projects more arrayed arrangement of points than the KFDA plot by two variables. In fact in binding three variables, viz., sepal width, petal length and petal width, KFDA shows up with minimum misclassification error .0133 among all other possible combination of variables. Also, while using all four variables in classification KFDA retains its superiority with reference to lowest misclassification error. Additionally, it is justified from misclassification error tables that for Iris data, sepal width, petal length and petal width are the essential morphological traits, inclusion of which enhances the effectivity of separating rule. In fact, classification among the species would be equally effective if we drop out sepal width from 3-variable bunching and include sepal length instead (See diagram in Annexure 9 for understanding).

In Iris data set, the association between sepal length and sepal width is low just –0.12. This near non-association boosts a reasoning on the non-importance of both of the sepal measurements (length/width) during analysis. In fact a statistician's take-home message may be framed as for Iris species separation, petal length and petal width play much influential role in comparison with the sepal measurements.

In this analysis we employ Gaussian kernel (symmetric) both in KFDA and KDA. But for the interested readers, choice of other types of kernels, (symmetric / asymmetric) would always be open. However, the possibility to use any desired kernels allows generalizing the method of separation which might involve complex algorithm as well as lengthy time. Also one might be inquisitive to look forward the best classifier over KFDA. Probabilistic neural network (PNN) which is a mapping operator built on a set of input-output observations might be a potential recipe. In fact KFDA defines a hyperplane

separation globally while PNN surfaces to more specific locally fitted, separative hyperplane.

References

[1] T.W. Anderson, An Introduction to Multivariate Statistical Analysis, New York, Wiley, Third Edition, (1984).

[2] M. A. Aizerman, E. M. Braverman and L. I. Rozonoer, Theoretical foundations of the potential function method in pattern recognition learning, Automation and Remote Control 25 (1964), 821-837.

[3] B. E. Boser, I. M. Guyon and V. N. Vapnik, A training algorithm for optimal margin classifiers. In D. Haussler edited vol. on 5th Annual ACM Workshop on COLT Pittsburgh, PA. (1992), 144-152.

[4] G. Baudat, and F. Anouar, Generalized discriminant analysis using a kernel approach, Neural Computation 12(10) (2000), 2385-2404.

[5] Donghwan Kim kfda, Kernel Fisher Discriminant Analysis, R package version >= 3.0.0, (2017). URL: https://github.com/ainsuotain/kfda.

[6] T. Duong, Kernel density estimation and kernel discriminant analysis for multivariate data in R, J. Stat. Softw. 21 (2007), 1-16.

[7] Ronald A. Fisher, The use of multiple measurements in taxonomic problems, Annals of eugenics 7(2) (1936), 179-188.

[8] P. Hall, and M. P. Wand, Minimizing L1 distance in nonparametric density estimation, Journal of Multivariate Analysis 26 (1988), 59-88.

[9] P. Hall, J. S. Marron and B. U. Park, Smoothed Cross-Validation, Probability Theory and Related Fields 92 (1992), 1-20.

[10] R. A. Johnson, and D. W. Wichern, Applied Multivariate Statistical Analysis, Prentice Hall India Learning Private Limited; 6th edition, (1982).

[11] B. Scholkopf, A. Smola and K. R. Muller, Nonlinear component analysis as a kernel eigenvalue problem, Neural Computation 10 (1998), 1299-1319.

[12] J. S. Simonoff, Smoothing Methods in Statistics, Springer-Verlag. New York, (1996).

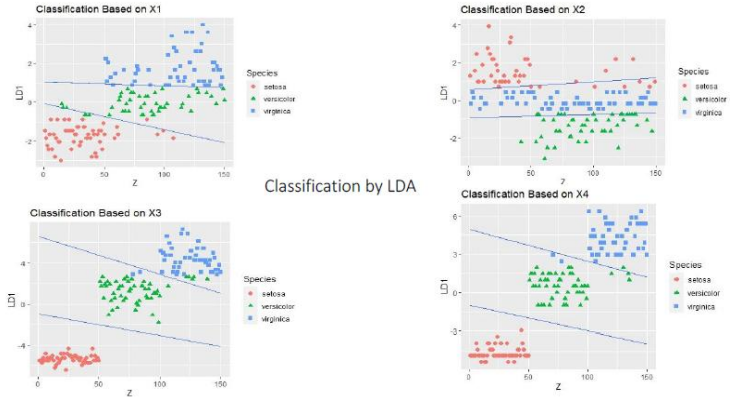
[13] V. Vapnik and C. Cortess, Support-vector networks, Machine Language 20(3) (1995), 273-29.

[14] J. Yang, Z. Jin, J. Y. Yang, D. Zhang, and A. F. Frangi, Essence of kernel Fisher discriminant, KPCA plus LDA, Pattern Recognition 37(10) (2004), 2097-2100.

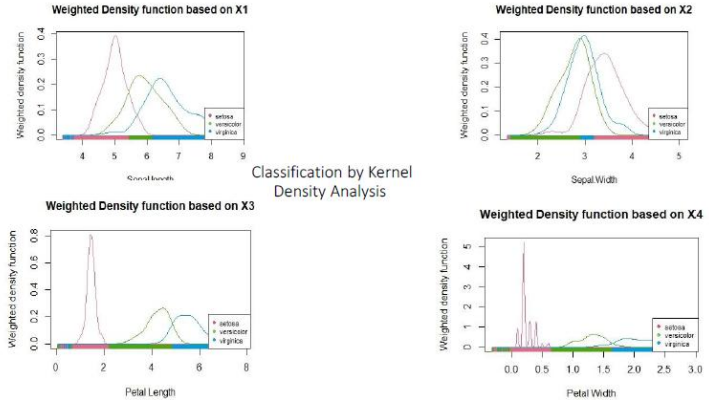
[15] Mika et al., Fisher discriminant analysis with kernels, Neural Networks For Signal Processing Ix: Proceedings of the IEEE Signal Processing Society Workshop, (1999).

Annexure

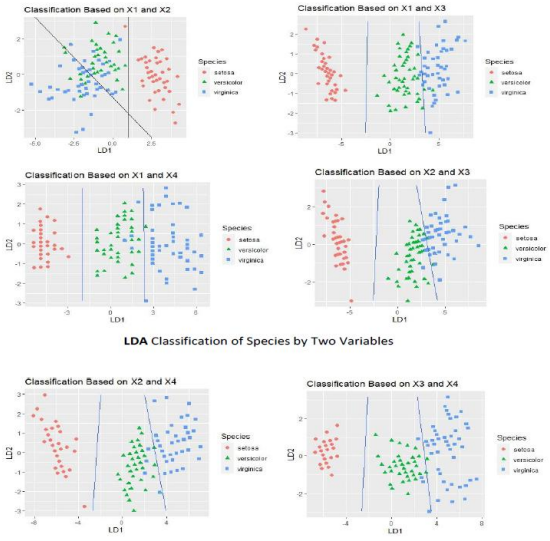
1. Classification based on a single variable



2. Classification by Kernel Density Analysis (KDA)

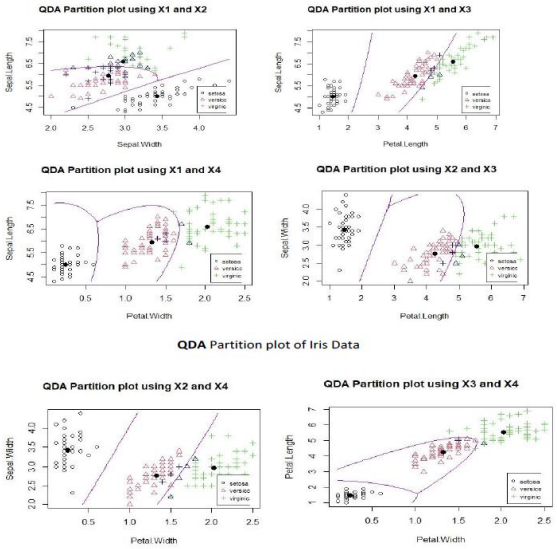


3. LDA Classification based on two variables

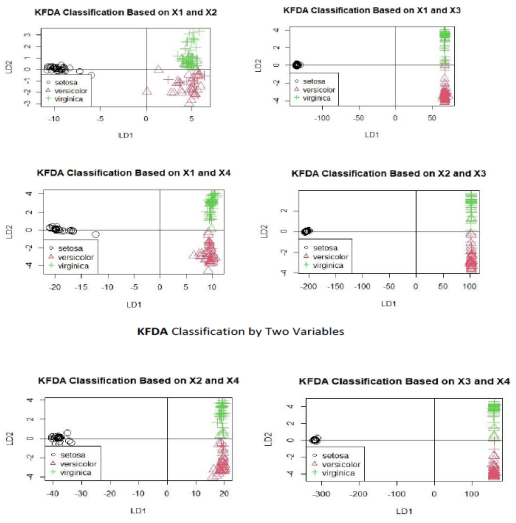


4. QDA plot taking two variables as classifying rule

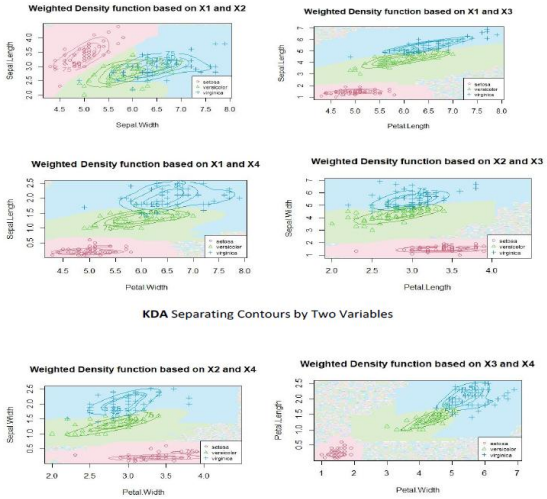
Here, each classification plot is divided by three colors indicating three species. Approximate error is also mentioned on top of each graph.



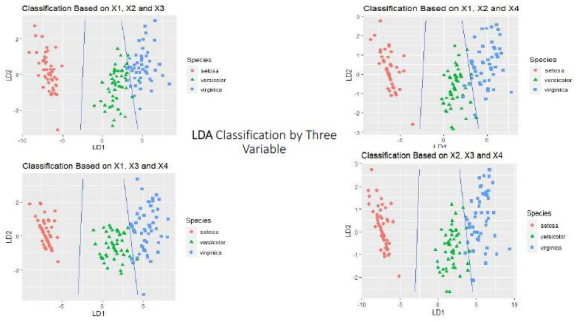
5. KFDA plot taking two variables



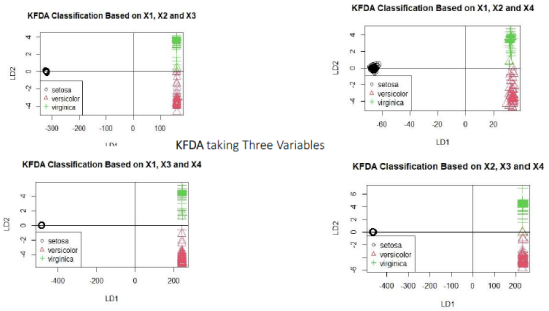
6. KDA plot taking two variables



7. LDA plot taking three variables as classifying rule



8. KFDA plot taking three variables as classifying rule



9. KDA plot taking three variables as classifying rule