Comparison Study of Generative and Discriminative Models for Classification of Classifiers.

[1]Anthony Rotimi Hassan, [2*]Rasaki Olawale Olanrewaju, [3]Queensley C. Chukwudum, [4]Sodiq Adejare Olanrewaju, [5]S. E. Fadugba

[*]Pan Africa University Institute for Basic Sciences, Technology, and Innovation
P.O. Box 62000-00200, Nairobi
Kenya

Abstract— In classification of classifier analysis, researchers have been worried about the classifier of existing generative and discriminative models in practice for analyzing attributes data. This makes it necessary to give an in-depth, systematic, interrelated, interconnected, and classification of classifier of generative and discriminative models. Generative models of Logistic and Multinomial Logistic regression models and discriminative models of Linear Discriminant Analysis (LDA) (for attribute P=1 and P>1), Quadratic Discriminant Analysis (QDA) and Naïve Bayes were thoroughly dealt with analytically and mathematically. A step-by-step empirical analysis of the mentioned models were carried-out via chemical analysis of wines grown in a region in Italy that was derived from three different cultivars (The three types of wines that constituted the three different cultivars or three classifiers). Naïve Bayes Classifier set the pace via leading a-prior probabilities.

It is noted that classifier of classification can be extended to K-Nearest Neighbors (KNN) model such that a value for "K", the number of nearest neighbors can be used as the classifier.

Keywords— Classification of Classifiers, Discriminative models, Generative models, Naïve Bayes, Regression Models.

I. Introduction

Rough categorizations of machine learning models are generative and discriminative algorithms. Generative models such as the Naïve Bayes often model $P(z_i \mid y)$ and $P(y)$ separately while discriminative models such as logistic regression model $P(y \mid z_i)$. In other words, the latter correspond the image samples $"z"$ to the class labels $"Y"$ (image classification) as opposed to image reconstruction, which is a characteristic of the former category. This implies that the generative models usually define how the data is generated while the discriminative model does not. A detailed study of these two categories was accomplished by [4].

.

Over the years, scholars have highlighted varying issues as it relates to the classification of classifier analysis of attributes data in practice. Ref. [6] studied how best algorithmic bias can be detected and mitigated when applying machine-learning algorithms in the making of both simple and complex decision processes. The effectiveness of various machines learning classification models have also been assessed by various authors [9], [2], [10], [3]. Ref. [8] specifically points out that the effectiveness of a machine learning solution is directly linked to the data's characteristic and nature as well as the learning algorithm's performance. This was the reasoning behind the author's study, that is, to provide the reader with an in-depth understanding of the principles behind the different

machine learning algorithms and how they can be applied practically in the different facets of life.

This is indicative of the fact that there exist very few studies that are geared towards enlightening users on how, why and when the different learning models can be applied. Our study is meant to contribute to literature from this perspective. Through the lens of generative and discriminative machine learning models, we provide an in-depth, systematic, interrelated, interconnected and classification of classifier of five (5) learning algorithms namely, logistic and multinomial regression models, Linear Discriminant Analysis (LDA) (for attribute P=1 and P>1), Quadratic Discriminant Analysis (QDA) and Naïve Bayes. This research is meant to serve as a learning tool and reference guide for both researchers in academia as well as industry professionals in the areas of risk management, cyber security, business and health to mention a few.

In the following sections of III, IV, V, VI, VII, VIII, IX and X; we critically look at the identified machine learning models, their limitations, possible interpretations of the parameters and the ways in which the parameters can be estimated. In section 8, a comprehensive application is undertaken in a systematic manner to display the applicability of the learning techniques in the field of (wine) business.

II. Methods

A Background

This research work aimed to mathematically, analytically and interpretatively visited generative models of Logistic and Multinomial Logistic regression models and discriminative models of Linear Discriminant Analysis (LDA) (for attribute P=1 and P>1), Quadratic Discriminant Analysis (QDA) and Naïve Bayes. The generative and discriminative models be juxtaposed via their aims, designs, descriptions, classifications, classification of the classifiers' methods, attributes, restrained boundaries, decision boundaries, special cases, dimensional functions and limitations. A classification of three classifier of a response variable will be used to carry out numerical analysis.

III. Logistic Regression Model

Regression models are models designed to ascertain covariates that contributed to a certain response variable [7]. There are variants of regression models that range from the linear to non-linear types. In scenarios where the response variable can be categorized as dichotomous or binary in nature (that is, one or two categories, say Yes or No) for K=2 classes of classification, logistic regression has proved to be the idea generalization as propounded by [5] and [11]. Instead of modeling the dependent variable, say, "Y" directly, the non-linear logistic regression model will be the appropriate generalization in order to estimate the probability that "Y" belongs to a particular paradigm as claimed by [12] and [1]. For example, if in a regression set-up of tossing a coin data for n^{th} trials with some deterministic measurements of exogenous or independent variable(s) (be it discrete or continuous), say "Z". The category here is of either a Head or Tail. If the interest is to find the logistic regression model that the probability belongs to head, then the probability of "Head" given "Z" can be written as

$$P\left(Y = Coin \,|\, \text{Head}\right) \tag{1}$$

The values of P(Y = coin|head), that can be acronym as P(Head) will range between 0 and 1, that is, $0 \le P\left(Head\right) \le 1$. The category of "Head" can be assigned "1", while "Tail" can be assigned "0". This means a generic code of 0/1 is used for the response variable. The question is how should the relationship between P(Z) = P(Y = 1|Z) and "Z" be modeled? Considering using a simple (Single Predictor of "Z") linear regression to typify these probabilities:

$$\mathrm{P}(Z) = \alpha_0 + \alpha_1 Z \tag{2}$$

This makes it easy to predict Y=Coin-using Head, the solution model for prediction is as shown in equation (3) below. The main problem here is the required technique for estimating the prediction: for Heads approaching zero estimates, negative probability of coin would be predicted. In case the prediction for Heads is very large, one could get values greater than 1. These predictions are not reasonable because ideally the true probability of coin, regardless of "Head" or "Tail" surfaces must lie between 0 and 1 inclusively. The lacuna is not peculiar to "Head" or "Tail" outcome of tossing of a coin only. Often time a straight line is fitted to a dichotomous dependent variable with generic code of 0/1, the governor bedrock is always P(Z)<0 for some values of "Z" and P(Z) > 1 for others that were not captured in P(Z)<0. To overcome this lacuna, P(Z) must be modeled using a function that enables outputs of "Z" to lie between 0 and 1 inclusively, that is, $0 \le Z \le 1 \quad \forall Zs$. A few functions has been formulated to meet-up this delineation, but the noted one is the logistic function of equation (3) below,

$$P\left(Z\right) = \frac{e^{\alpha_0 + \alpha_1 Z}}{1 + e^{\alpha_0 + \alpha_1 Z}} \tag{3}$$

Fitting equation (3) requires methods like Maximum Likelihood (ML), Reweighted Iterative Procedure etc., but the ML method would be employed in this work. Considering equation (3) with coin data, it is to be noted that few "Heads" in n^{th} trail of tossing a coin, the prediction of "Y" would be very close to zero, but can never go below zero. Similarly, many "Heads" in n^{th} trail of tossing a coin, the prediction of

"Y" will yield probability one, but never exceeded one. The S-shaped curve is the product of the logistic function of equation (3) for single variable "Z", irrespective of the value of "Z". Manipulatively, equation (3) can be written as:

$$\frac{P(Z)}{1-P(Z)} = e^{\alpha_0+\alpha_1 Z} \tag{4}$$

The logistic regression model is known to be able to capture range of probabilities in comparison to just linear regression. The magnitude $\frac{P(Z)}{1-P(Z)}$ is usually referred to as the odds, such that is takes values $0 < \frac{P(Z)}{1-P(Z)} < \infty$. Odd values close to the boundary of zero and infinity indicate low and high probabilities of the coin respectively. Odds are customarily used instead of probabilities. Taking the logarithm of equation (4) gives

$$\log\left(\frac{P(Z)}{1-P(Z)}\right) = \alpha_0 + \alpha_1 Z \tag{5}$$

Moreover, is referred to as logarithms of odds or logit (that is, logistic regression model has a logit that is linear in "Z").

B Interpretation

It connotes that an increment in "Z" by one-unit alters the logarithms of odds in equation (5). In equal manner, it manifolds the odds by e^{α_1} in equation (4). Since the relationship between P(Z) and "Z" is always non-linear, α_1 does not jibe to the alternation in P(Z) that associate to one-unit increment in "Z". The magnitude that P(Z) alters due to a unit alternation in "Z" lies solely on the current value of "Z". Irrespective of the magnitude of "Z", if α_1 is positive, then increasing "Z" will simply imply an increment in P(Z). Similarly, if α_1 is negative, then decreasing "Z" will simply connote a decrement in P(Z). In a nut shell, the rate of alternation in P(Z) per unit alternation in "Z" lies solely on the current value of "Z".

C Estimating the Logistic Regression Coefficients

The coefficients α_0 and α_1 in (5) are unknown and can be estimated based on the available training of the data via

$$\ell(\alpha_0,\alpha_1) = \prod_{if,y=0}\left(1-p(\mathrm{z}_i)\right)\prod_{if,y=1} p(\mathrm{z}_i) \tag{6}$$

The estimates α_0 and α_1 are chosen to maximize this likelihood function.

IV. MULTINOMIAL LOGISTIC REGRESSION

Logistic regression model with K = 2 classes of classification for dependent variable "Y". In situation where predicting response uses multiple predictors coupled with classes of classification is strictly greater than two (generalization in terms of multiple predictors and more than two classes of classification), that is setting of K>2 classes. This kind of extension is referred to as multinomial logistic regression. In this kind of setting, a single class would be firstly selected as benchmark such that no loss of particularity about the selected K^{th} class.

Then equation (3) can be replaced by

$$P\left(Y=k/\mathrm{Z}=z\right) = \frac{e^{\alpha_{k0}+\alpha_{k1}z_1+\cdots+\alpha_{kp}z_p}}{1+\sum_{i=1}^{K-1} e^{\alpha_{k0}+\alpha_{k1}z_1+\cdots+\alpha_{kp}z_p}} \quad \forall\ k=1,\cdots,K-1 \tag{7}$$

$$P\left(Y=k/\mathrm{Z}=z\right) = \frac{1}{1+\sum_{i=1}^{K-1} e^{\alpha_{k0}+\alpha_{k1}z_1+\cdots+\alpha_{kp}z_p}} \tag{8}$$

The logarithm of odds or logit for multinomial logistic regression is:

$$\log\left(\frac{P\left(Y=k/\mathrm{Z}=z\right)}{P\left(Y=K/\mathrm{Z}=z\right)}\right) = \alpha_{k0}+\alpha_{k1}z_1+\cdots+\alpha_{kp}z_p \tag{9}$$

So,

$$\log\left(\frac{p(Z)}{1-p(Z)}\right) = \alpha_{k0}+\alpha_{k1}z_1+\cdots+\alpha_{kp}z_p \tag{10}$$

Where $Z=\left(Z_1,\cdots,Z_p\right)$ for "p" predictors or independent variables. We have described how to model a logistic and multinomial logistic regression using logistic functions of equation (5) and equation (10) via a direct approach of $P\left(Y=k\middle|Z=z\right)$. Another approach will be needed for both the logistic and multinomial logistic regression when there is considerable large margin between the two classes (for logistic regression) or among K-classes for multinomial logistic regression, which usually lead to their coefficients, α_s been shockingly unstable. Additionally, when the distribution of "Z" within each class of "K" is known to be normally distributed and the sample size is small, the logistic and multinomial regressions are no accurately reliable.

Considering an alternative and not crooked approach for estimating the probabilities of logistic and multinomial regressions, such that the distribution of the independent variables "Z" are modeled separately in each classes of the response variable "Z". The Bayes' theorem can be used to tumble into estimating $P\left(Y=k\,|\,Z=z\right)$ especially when the

distribution of "Z" within each class of "K" is known to be normally distributed. Assuming classifying measurements into one of K-classes, such that $K \geq 2$. In a simplified term, the qualitative measurement of "Z" can assume values of K-possible non-identically and irregular measurements. Let η_k denote the total or prior probability that a "Z" measurement emanates from the K^{th} class. Furthermore, let $g_k(Z) = P(Z \mid Y = k)$ denote the density function of "Z" of measurement emanates from the K^{th} class. Alternatively, $g_k(Z)$ is referencing large if there is high chance that a measurement in the K^{th} class has $Z = z$ and $g_k(Z)$ is small if there is no chance that a measurement in the K^{th} class has $Z = z$. Then the Bayes' theorem mathematically,

$$P(Y = k \mid \mathrm{Z} = z) = \frac{\eta_k g_k(\mathrm{z})}{\sum_{i=1}^{K} \eta_i g_i(\mathrm{z})} \tag{11}$$

The notation $P_k(Z) = P(Y = k \mid Z = z)$ stands for the posterior probability that a measurement $Z = z$ falls to the K^{th} class. In other words, it is the probability that a measurement belongs to the K^{th} class, given the value of the independent variable "Z". Equation (11) postulates an alternative direct computation of posterior probability $P_k(\mathrm{z})$ of logistic and multinomial regression. Computational wise, estimation of η_k can be done via a random sample from population as fraction of trained observations that falls into K^{th} class. However, it might be difficult to estimate $g_k(\mathrm{z})$, expect when some modifying and simplifying assumptions are made. The Bayes classifier can now be used to estimate $g_k(\mathrm{z})$ via classification of an observation say "z" to the category for which $g_k(\mathrm{z})$ is of high magnitude, possessed the lowest possible error rate out of all the classifiers. This can only be true and possible via correct specification of terms in equation (11). However, $g_k(\mathrm{z})$ can be estimated via some classifiers, then plug into equation (11) for proper approximation of the Bayes classifier.

Three different classifiers that use different estimates of $g_k(\mathrm{z})$ of equation (11) will be used as correctness for the Bayes classifier: Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), and Naïve Bayes.

V. Linear Discriminant Analysis (LDA) When Predictor is One (P=1)

.

Assuming the predictor is just one, that is, $p = 1$. Our focus is to estimate $g_k(\mathrm{z})$ then plug it into equation (12) in order to estimate $P_k(\mathrm{z})$. The goal is to classify a predictive measurement to the class for which of $P_k(\mathrm{z})$ is largest. To estimate $g_k(\mathrm{z})$, assumptions about its form need to be made. Assuming $g_k(\mathrm{z})$ is Gaussian of one-normal Gaussian dimensional; the Gaussian density takes the form

$$g_k(\mathrm{z}) = \frac{1}{\sqrt{2\pi\sigma_k}} \exp\left(-\frac{(\mathrm{z}-\mu_k)}{2\sigma_k^2}\right) \tag{12}$$

Where μ_k and σ_k^2 are the mean and variance parameters of the K^{th} class. Assume further that $\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_K^2$, that is the same-shared variance across all K^{th} classes. Inserting equation (12) into equation (11) gives

$$p_k(\mathrm{z}) = \frac{\eta_k \frac{1}{\sqrt{2\pi\sigma}} \exp\left(-\frac{(\mathrm{z}-\mu_k)^2}{2\sigma^2}\right)}{\sum_{i=1}^{K} \eta_i \frac{1}{\sqrt{2\pi\sigma}} \exp\left(-\frac{(\mathrm{z}-\mu_k)^2}{2\sigma_i}\right)} \tag{13}$$

η_k stands for the prior probability that a predictive measurement belongs to the K^{th} class. The Bayes classifier is concern about assigning a predictive measurement $Z = z$ to the class for which equation (13) has greatest magnitude. Taking the logarithm of equation (13) and rearranging the terms gives

$$\varsigma_k(\mathrm{z}) = z.\frac{\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} + \log(\eta_k) \tag{14}$$

is largest. If K = 2, then $\eta_1 = \eta_2$, then the Bayes classifier apportions a predictive measurement of first class if $2z(\mu_1 - \mu_2) > \mu_1^2 - \mu_2^2$ and to second class otherwise. The Bayes decision boundary is the quantity, $\varsigma_1(\mathrm{z}) = \varsigma_2(\mathrm{z})$, this is tantamount to

$$z = \frac{\mu_1^2 - \mu_2^2}{2(\mu_1 - \mu_2)} = \frac{(\mu_1 - \mu_2)(\mu_1 + \mu_2)}{2(\mu_1 - \mu_2)} = \frac{\mu_2 + \mu_1}{2} \tag{15}$$

At times, when we are sure of the assertion that "Z" is sampled from a Gaussian distribution within each K^{th} class, the parameters $\mu_1, \cdots, \mu_k$, $\eta_1, \cdots, \eta_k$ and σ^2 still need to be estimated when applying the Bayes classifier. The Linear Discriminant Analysis (LDA) technique will correct the Linear Discriminant Analysis Bayes Classifier (LDABC) by inserting the estimates of μ_k, σ^2 and η_k into equation (16) and the estimates are

$$\mu_k = \frac{1}{n_k} \sum_{i:y_i = k} z_i \tag{16}$$

$$\sigma^2 = \frac{1}{n-k}\sum_{k=1}^{K}\sum_{i:y_i=k}\left(z_i - \mu_k\right)^2 \quad (17)$$

"n" is the totality of trained predictive measurements; "n_k" is the totality of trained predictive measurements in the K^{th} class. μ_k is the mean of all trained predictive measurements of K^{th} class; σ^2 is the weighted average of trained variances for each K^{th} classes. The membership probabilities of $\eta_1,\cdots,\eta_k$ at sometimes known and can be used directly. In cases where the probabilities have any additional information, LDA can estimate η_k using the apportion trained predictive measurements that belong to K^{th} class, that is,

$$\eta_k = {}^{n_k}\!/_{n} \quad (18)$$

For the LDA classifier, we insert estimates of equation (17) and (18) into equation (14) and allot predictive measurement Z=z to the class for which

$$\varsigma_k(\mathrm{z}) = \mathrm{z}.\frac{\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} + \log(\eta_k) \quad (19)$$

is largest quantity. The discriminant functions $\varsigma_k(\mathrm{z})$ in (19) are linear functions of "z".

VI. Linear Discriminant Analysis (LDA) when Predictor is Greater than One (P=1)

Extending the LDA classifier to case of multiple predictors, that is, $p>1$. By doing so, assuming $Z=\left(Z_1,Z_2,\cdots,Z_p\right)$ is sampled from a multivariate normal distribution with a class-specific multivariate mean vector and a common covariance matrix. The multivariate normal distribution makes it possible for each of the predictor "Z" with correlation between each pair to follow the defined one-dimensional Gaussian distribution specified in equation (12).

For p-dimensional random variable of "Z" to follow a multivariate Gaussian distribution. It is denoted by $Z \square N\left(\mu,\Phi\right)$, such that $\mathrm{E}(Z)=\mu$ is the mean vector of "Z" with p-components, $Cov(Z)=\Phi$ stands for the $p\times p$ variance-covariance matrix of "Z". The p-dimensional multivariate Gaussian density can be mathematically written as,

$$\mathrm{g(z)} = \frac{1}{(2\pi)^{\frac{p}{2}}\left|\Phi\right|^{\frac{1}{2}}}\exp\left(-\frac{1}{2}(\mathrm{z}-\mu_k)^T\Sigma^{-1}(\mathrm{z}-\mu_k)\right) \quad (20)$$

It implies that the LDA classifier for $p>1$ presumes predictive measurements in the K^{th} class are sampled from a p-dimensional multivariate Gaussian density with $N\left(\mu,\Phi\right)a$, where μ_k is the class-specified mean vector and Φ is the $p\times p$ variance-covariance matrix of "Z" that is common to all K^{th} classes. Putting the p-dimensional multivariate Gaussian density for the K^{th} class, $g_k(\mathrm{z})$, into equation (12) and after working-out the solution that the Bayes classifier presumes a predictive measurement Z=z to the K^{th} class for which

$$\varsigma_k(\mathrm{z}) = z^T\Phi^{-1}\mu_k - \frac{1}{2}\mu_k^T\Phi^{-1}\mu_k + \log\eta_k \quad (21)$$

is the largest quantity. The vector is the updated version of equation (14). The Bayes decision boundary is the quantity, $\varsigma_k(\mathrm{z})=\varsigma_t(\mathrm{z})$, this is tantamount to

$$z^T\Phi^{-1}\mu_k - \frac{1}{2}\mu_k^T\Phi^{-1}\mu_k = z^T\Phi^{-1}\mu_t - \frac{1}{2}\mu_t^T\Phi^{-1}\mu_t \quad (22)$$

For $k\neq t$. It is to be noted that the quantity $\log\eta_k$ in equation (22) has vanished because each of the three classes has the same number of trained predictive measurements, that is η_k is the same in each class.

VII. Quadratic Discriminant Analysis (QDA)

LDA presumes that the p-dimensional multivariate Gaussian density with $N\left(\mu,\Phi\right)$, where μ_k is the class-specified mean vector and Φ is the $p\times p$ variance-covariance matrix of "Z" that is common to all K^{th} classes predictive measurements within each class that are sampled from p-dimensional Gaussian density. Quadratic Discriminant Analysis (QDA) presumes an alternative discriminant analysis approach via assuming that the predictive measurements from each class emanated from Normal distribution and putting it into the Bayes' theorem of equation (11) in order to aid prediction. Unlike LDA, QDA presumes that each class component of "K" has its own variance-covariance matrix. In other words, it suggested that a predictive measurement from the K^{th} class is of the form $Z \square N\left(\mu_k,\Phi_k\right)$, such that Φ_k is the covariance matrix for the K^{th} class. Under this assumption, the Bayes classifier presumes a predictive measurement Z = z to the class for which

$$\varsigma_k(\mathrm{z}) = -\frac{1}{2}(\mathrm{z}-\mu_k)^T\Phi_k^{-1}(\mathrm{z}-\mu_k) - \frac{1}{2}\log\left|\Phi_k\right| + \log\eta_k$$

$$= -\frac{1}{2}z^T\Phi_k^{-1}z + z^T\Phi_k^{-1}\mu_k - \frac{1}{2}\mu_k^T\Phi_k^{-1}\mu_k - \frac{1}{2}\log\left|\Phi_k\right| + \log\eta_k \quad (23)$$

is the largest quantity. QDA entangles putting the estimates of Φ_k, η_k and μ_k into equation (23), and thereafter assign predictive measurement Z=z to the class to which this quantity is largest. Unlike equation (22), the magnitude "z" becomes

visible to be a quadratic function in equation (23). In fact, this where Quadratic Discriminant Analysis (QDA) extracts its name. Does it meaningful whether or not it is possible to presume K^{th} classes share a common variance-covariance vector?

VIII. NAÏVE BAYES

In this section, we shall be using Bayes' theorem of equation (11) to evolve the QDA and LDA classifiers. The motive here is to use Bayes' theorem to popularize the Naïve Bayes classifier. From equation (11) of the Bayes' theorem, the mathematical expression provides the reflection of the posterior probability $P_k(z) = P(Y = k \mid Z = z)$ in quantity of $\eta_1, \cdots, \eta_K$ and $g_1(z), \cdots, g_K(z)$. η_k can be estimated via the proportion of trained predictive measurements that belong to K^{th} class, for $k = 1, \ldots, \mathrm{K}$. It can be impalpable in estimating $g_1(z), \cdots, g_K(z)$. The Naïve Bayes would tackle estimating of $g_1(z), \cdots, g_K(z)$ from another angle. Instead of assuming that $g_1(z), \cdots, g_K(z)$ belong to multivariate normal distribution, we just assume a single assumption of

$$g_k(z) = g_{k1}(z_1) \times g_{k2}(z_2) \times \cdots \times g_{kp}(z_p) \qquad k = 1, \ldots, K \qquad (24)$$

Where g_{kj} is the distribution function of the j[th] of the independent among the predictive measurements in the k[th] class. This is somehow refers to as independency assumption. This assumption is very cognate because it usually engaging both estimating the marginal distribution of each independent variable and their joint distribution for a p-dimensional density functions. Naïve Bayes is a good choice since a huge amount of data is always require when estimating joint distribution. However, the independency of Naïve Bayes might introduce some sort of biasedness, but noted for variance reduction. Establishing the Naïve Bayes assumption makes it realistic to insert equation (24) into equation (11) to obtain a posterior probability of:

$$P(Y = k \mid Z = z) = \frac{\eta_k \times g_{k1}(z_1) \times g_{k2}(z_2) \times \cdots \times g_{kp}(z_p)}{\sum_{\iota=1}^{K} \eta_\iota \times g_{\iota 1}(z_1) \times g_{\iota 2}(z_2) \times \cdots \times g_{\iota p}(z_p)}$$

$$k = 1, \cdots, K \qquad (25)$$

To estimate the one-dimensional density function g_{kp} using trained data of $z_{1j}, \cdots, z_{np}$, we can still have some alternatives:

(i) In case Z_j is quantitative, $Z_j \mid Y = k \sim N(\mu_{jk}, \sigma_{jk}^2)$. In case it is presume that within each class, the j^{th} independent variables sample from a univariate Gaussian distribution. This might sound like QDA, but there is a cognate difference of independent variables assume to be independent, which tantamount to a supplementary assumption that the each K[th] class-specific variance-covariance vector is diagonal.

(ii) Another viable option is non-parametric estimation of g_{kp} when Z_j is quantitative. This can be done via bins of histogram for predictive measurements for j^{th} independent variables within each class. $g_{kp}(z_p)$ can be estimated as a fraction of trained predictive measurements in the K[th] class that falls into the same bin of histogram as z_p. Similarly, Kernel Density Estimator (KDE) to estimate g_{kp}; KDE is an essential smoothed version of histogram via bandwidths.

(iii) In case Z_j is qualitative, a simply count of the proportion of trained predictive measurements for the J^{th} independent variable corresponding to each class. For example, if $Z_p = \{5, 6, 7\}$, and 200 predictive measurements are in the K[th] class. Furthermore, if J^{th} independent variable takes on values of 5, 6, and 7 in 12, 24, and 36 of those predictive measurements, then g_{kp} can be estimated as,

$$g_{kp}(z_p) = \begin{cases} 0.025, \text{ if } z_p = 5 \\ 0.03, \text{ if } z_p = 6 \\ 0.036, \text{ if } z_p = 7 \end{cases} \qquad (26)$$

The illustrated Naïve Bayes classifier is with P = 3 independent variables and K = 2 classes. The first two independent variables are quantitative, while the third independent variable is qualitative with three levels.

IX. COMPARISON OF CLASSIFICATION OF THE CLASSIFIERS' METHODS

Performing an analytical comparison of logistic and multinomial regression; LDA, QDA and Naïve Bayes. Considering the approaches of these mentioned methods in a setting of K-classes and assigning a predictive measurement to that minimizes $P(Y = k \mid Z = z)$. Significantly, setting K as

baseline class and apportion a predictive measurement to the class that minimizes gives

$$\log\left(\frac{P(Y=k\mid \mathrm{Z}=z)}{P(Y=K\mid \mathrm{Z}=z)}\right) k=1,\cdots,K \tag{27}$$

Specifying the form of equation (27) for each method of LDA, QDA, and Naïve Bayes will furnish us with a clear understanding of their differences and similarities. However, for LDA, the Bayes' theorem of (11) coupled with the assumption that the independent variables within each class are sampled from a multivariate Gaussian density of equation (20) with K^{th}-class-specific mean and shared variance-covariance vector as shown below:

$$\log\left(\frac{P(Y=k\mid \mathrm{Z}=z)}{P(Y=K\mid \mathrm{Z}=z)}\right)=\log\left(\frac{\eta_k g_k(\mathrm{z})}{\eta_K g_K(\mathrm{z})}\right) \tag{28}$$

$$=\log\left(\frac{\eta_k \exp\left(\frac{-\left((\mathrm{z}-\mu_k)^T\Phi^{-1}(\mathrm{z}-\mu_k)\right)}{2}\right)}{\eta_K \exp\left(\frac{-\left((\mathrm{z}-\mu_K)^T\Phi^{-1}(\mathrm{z}-\mu_K)\right)}{2}\right)}\right) \tag{29}$$

$$=\log\left(\frac{\eta_k}{\eta_K}\right)-\frac{(\mathrm{z}-\mu_k)^T\Phi^{-1}(\mathrm{z}-\mu_k)}{2}+\frac{(\mathrm{z}-\mu_K)^T\Phi^{-1}(\mathrm{z}-\mu_K)}{2} \tag{30}$$

$$=\log\left(\frac{\eta_k}{\eta_K}\right)-\frac{(\mu_K+\mu_k)^T\Phi^{-1}(\mu_K-\mu_k)}{2}+\frac{z^T\Phi^{-1}(\mu_k-\mu_K)}{2} \tag{31}$$

$$=c_k+\sum_{j=1}^{p}d_{kj}z_j \tag{32}$$

Where $c_k=\log\left(\frac{\eta_k}{\eta_K}\right)-\frac{1}{2}(\mu_k+\mu_K)^T\Phi^{-1}(\mu_k-\mu_K)$. d_{kj} is the J^{th} factor of $\Sigma^{-1}(\mu_k-\mu_K)$.

In a similar vein like that of logistic regression, LDA presumes log-odds of the posterior probabilities to be linear in "z". Similarly, to the calculations of equation (32), QDA becomes

$$\log\left(\frac{P(Y=k\mid \mathrm{Z}=z)}{P(Y=K\mid \mathrm{Z}=z)}\right)=c_k+\sum_{j=1}^{p}d_{kj}z_j+\sum_{j=1}^{p}\sum_{l=1}^{p}e_{kj}z_jz_l \tag{33}$$

where c_k, d_{kj} and e_{kjl} are functions of η_k, η_K, μ_k, μ_K, Φ_k and Φ_K. QDA presumes log-odds of the posterior probabilities to be quadratic in z. Lastly, examining equation (27) in Naïve Bayes setting; It is to be recollected that $g_k(z)$ is modeled as a multiplication of p one-dimensional functions $g_{kj}(z_j)$ $j=1,\ldots,p$. Henceforth,

$$\log\left(\frac{P(Y=k\mid \mathrm{Z}=z)}{P(Y=K\mid \mathrm{Z}=z)}\right)=\log\left(\frac{\eta_k g_k(\mathrm{z})}{\eta_K g_K(\mathrm{z})}\right) \tag{34}$$

$$=\log\left(\frac{\eta_k\prod_{j=1}^{p}g_{kj}(\mathrm{z}_j)}{\eta_K\prod_{j=1}^{p}g_{Kj}(\mathrm{z}_j)}\right) \tag{35}$$

$$=\log\left(\frac{\eta_k}{\eta_K}\right)+\sum_{j=1}^{p}\log\left(\frac{g_{kj}(\mathrm{z}_j)}{g_{Kj}(\mathrm{z}_j)}\right) \tag{36}$$

$$=c_k+\sum_{j=1}^{p}h_{kj}(\mathrm{z}_j) \tag{37}$$

$c_k=\log\left(\frac{\eta_k}{\eta_K}\right)$ and $h_{kj}(\mathrm{z}_j)=\log\left(\frac{g_{kj}(\mathrm{z}_j)}{g_{Kj}(\mathrm{z}_j)}\right)$. It is to be noted that equation (37) takes the form of a Generalized Additive Model (GAM). After proper scrutiny and investigation of equation (32), (33), and (37) payoff the following reflections about LDA, QDA, and Naïve Bayes:

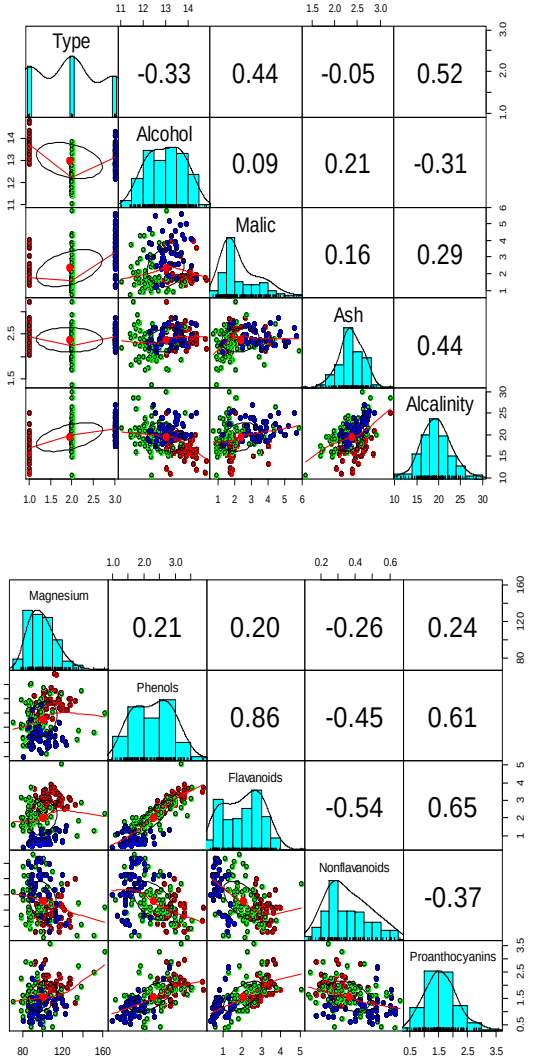
- QDA is tantamount to LDA provided $e_{kjl}=0$ $\forall$ $j=1,\ldots,p$ and $k=1,\ldots,\mathrm{K}$.
- LDA is a restrained bound of QDA version if and only if $\Phi_1=\Phi_2=\cdots=\Phi_K=\Phi$.
- Classifiers with linear decision boundaries are special case of Naïve Bayes with $h_{kj}(z_j)=d_{kj}z_j$. This connote that LDA is a special case of Naïve Bayes.
- If $N\left(\mu_{kj},\sigma_j^2\right)$ is the distribution of the one-dimensional Gaussian distribution modeled via Naïve Bayes of $g_{kj}(z_j)$, then $g_{kj}(z_j)=d_{kj}z_j$, such that $d_{kj}=\frac{\left(\mu_{kj}-\mu_{Kj}\right)}{\sigma_j^2}$. It make-up that Naïve Bayes is a special case of LDA with Φ restrained bound to a diagonal vector with J^{th} diagonal element that equal to σ_j^2.

The classification of classifiers can be extended to K-Nearest Neighbors (KNN) that is completely a non-parametric technique.

X. RESULTS

The dataset used to validate comparison of the model is a chemical analysis of wines grown in a region in Italy that was

derived from three different cultivars (The three types of wines constitute the three different cultivars) obtained in 1991. Thirteen (13) distinct synthetics comprises of the chemical analysis: these are malic_acid, alcalinity of ash, alcohol, magnesium, total phenols, flavanoids etc. that contains 178 occurrences. The whole thirteen (13) variables are regarded as attributes and all continuous. The three types of wines are classifier attribute with three identification (1-3). In this classification context, the classifier structures posed a well-behaved good dataset of classifier.



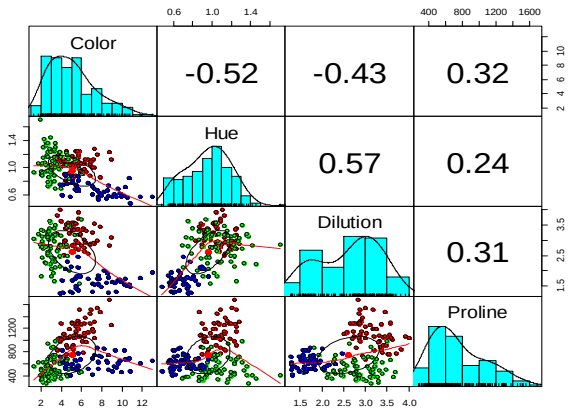


Fig. 1 Pair Panel Plot for the Thirteen (13) distinct synthetics chemicals of the Wine.

Having known that the pair panel plot not only reveals the correlation coefficient between each pairwise combination of variables as well as a density plot for each individual variable. It also shows the histogram for each variable of concern. The classifier variable itself "Type" made of three classes "1, 2, 3" with the second type of the wine being the apex. It has a positive relationship with distinct synthetics of malic, alcalinity, magnesium, phenols, flavanoids, Proanthocyanins, and proline. Whereas it possessed a negative relationship with other distinct synthetics. It can be deduced that distinct synthetics of alcohol, malic, ash, alkalinity, magnesium, proanthocyanins, proline, and hue are somewhat perfectly approximately normal (Gaussian) around there means, while the likes of Dilution, Non-flavanoid, flavonoids, and phenols approximately normal (Gaussian) around there means, but in a multimodal manner.

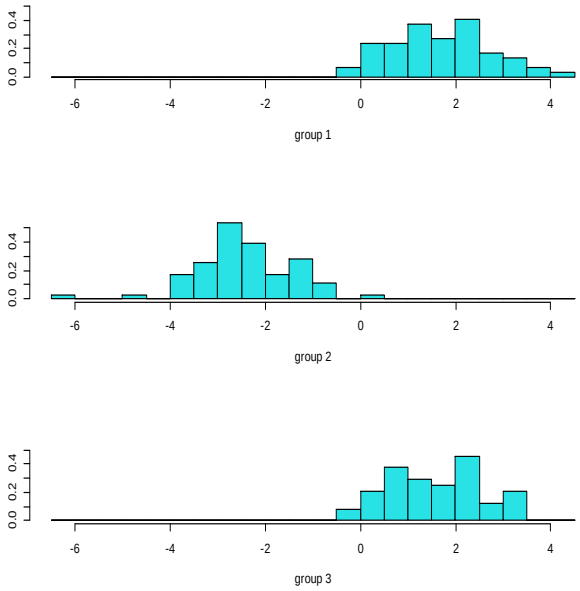


Fig. 2 Histogram Plot of the three Classifier of the Wine Types

The first and third group/type of the wine are asymmetrically left-skewed distributed. Meaning that a natural limit prevents outcomes on one side and center. In other words, the distributions' peak are centered-off towards limits and tail stretches away from them, literally saying that the "wine" cannot be more than 100 percent pure, or one type. The second group/type of the wine possessed a distanced and semi-hidden plateau (multimodal) normal distribution processes that are rightly skewed because there are like three (semi hidden and clearly seen) peaks closed together. This justified the three-classifier possession of the wine datasets.

Table 1. Multinomial Logistic Regression

	Intercept	Alcohol	Malic	Ash	Alcalinity	Magnesium	Phenols	Flavanoids	Nonflavanoids	Proanthocyanins
2	848.475 (0.018)	-50.420 (0.194)	-16.83 (0.022)	-376.293 (0.067)	30.780 (0.603)	-0.319 (2.334)	115.4110 (0.078)	-93.036 (0.150)	358.949 (0.007)	100.583 (0.046)
3	-149.965 (0.013)	61.762 (0.167)	14.006 (0.042)	-93.303 (0.033)	28.102 (0.286)	-4.288 (1.377)	267.941 (0.021)	-256.984 (0.008)	-389.821 (0.008)	151.492 (0.012)

Color	Hue	Dilution	Proline	Sum of Probabilities	Mean of Probabilities
-24.093 (0.115)	442.306 (0.017)	-26.299 (0.087)	-0.558 (7.118)	178	0.3333333
37.772 (0.073)	-220.33 (0.011)	-166.23 (0.027)	-0.411 (7.408)		

Table 2. Linear Discriminant Analysis (LDA)

Group Means	Alcohol	Malic	Ash	Alcalinity	Magnesium	Phenols	Flavanoids	Nonflavanoids	Proanthocyanins
1	13.745	2.011	2.456	17.037	106.3390	2.840	2.982	0.290	1.899
2	12.279	1.933	2.245	20.238	94.549	2.259	2.081	0.364	1.630
3	13.154	3.334	2.437	21.417	99.313	1.679	0.782	0.448	1.154

Color	Hue	Dilution	Proline	Sum of Posterior	Mean of Posterior	Prior probabilities of groups:
5.528	1.062	3.158	1115.712	178	0.335	0.332
3.087	1.056	2.785	519.507			0.399
7.396	0.683	1.684	629.896			0.270

Table 3. Coefficients of Linear Discriminants Analysis

Covariates	LD1	LD2	Proportion of trace (LD1)	Proportion of trace (LD2)
Alcohol	-0.404	0.872	0.6875	0.3125
Malic	0.165	0.305		
Ash	-0.369	2.346		
Alcalinity	0.155	-0.146		
Magnesium	-0.002	-0.0005		
Phenols	0.618	-0.032		

Flavanoids	-1.661	-0.492	
Nonflavanoids	-1.496	-1.631	
Proanthocyanins	0.134	-0.307	
Color	0.355	0.253	
Hue	-0.818	-1.516	
Dilution	-1.158	0.051	
Proline	-0.003	0.003	

Table 4. Quadratic Discriminant Analysis (QDA)

Group Means	Alcohol	Malic	Ash	Alcalinity	Magnesium	Phenols	Flavanoids	Nonflavanoids	Proanthocyanins
1	13.745	2.011	2.456	17.037	106.3390	2.840	2.982	0.290	1.899
2	12.279	1.933	2.245	20.238	94.549	2.259	2.081	0.364	1.630
3	13.154	3.334	2.437	21.417	99.313	1.679	0.782	0.448	1.154

Color	Hue	Dilution	Proline	Sum of Posterior	Mean of Posterior	Prior probabilities of groups:
5.528	1.062	3.158	1115.712	178	0.335	0.332
3.087	1.056	2.785	519.507			0.399
7.396	0.683	1.684	629.896			0.270

Table 5. Naïve Bayes

Conditional Probabilities:

Var.	Alcohol		Malic		Ash		Alcalinity		Magnesium		Phenols		Flavanoids	
	Y[,1]	Y[,2]	Y[,1]	Y[,2]	Y[,1]	Y[,2]	Y[,1]	Y[,2]	Y[,1]	Y[,2]	Y[,1]	Y[,2]	Y[,1]	Y[,2]
1	13.745	0.462	2.010	0.689	2.456	0.227	17.037	2.546	106.33	10.499	2.840	0.339	2.982	0.398
2	12.279	0.538	1.932	1.016	2.245	0.316	20.238	3.350	94.549	16.754	2.259	0.545	2.081	0.706
3	13.154	0.530	3.334	1.088	2.437	0.185	21.417	2.258	99.313	10.891	1.679	0.357	0.782	0.294

B Discussion

From table 1 for multinomial logistic regression analysis, the sum of probabilities was estimated to be 178 with it mean calculated to be 0.333. The problem here is that can sum of probability of a proper defined probability space be >1?. No, it makes the space void. In a similar manner, table 2 to table 4 for Linear Discriminant Analysis (LDA) and Quadratic Discriminant Analysis (QDA) produced sum of posterior and mean of posterior of 178 and 0.335 respectively. Both the LDA and QDA produced similar prior probabilities of 0.332, 0.399, and 0.270 for the three classifiers: type 1, type 2, and type 3 of the wine respectively. LDA and QDA produced the same proportion of trace for (LD1) & (QDA1) and (LD2) & (QDA2) to be 0.6875 and 0.3125 respectively. QDA and LDA models have a slightly higher precision of 0.335 compare to lower precision of 0.303 for multinomial logistic regression model. Overall, Naïve Bayes possessed improved and higher A-priori probabilities of 0.334, 0.404, and 0.287 for the three classifiers: type 1, type 2, and type 3 of the wine respectively.

XI CONCLUSION

Overall, this analysis indicates the performance of classification of classifiers of multinomial logistic regression analysis, Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis and Naïve Bayes models. In conclusion, the four mentioned comparison models appealed to the three classified classifier of the wine dataset with Naïve Bayes dictating the pace for the ideal prior probabilities needed. The classifier of classification can be extended to K-Nearest Neighbors (KNN) model such that a value for "K", the number of nearest neighbors can be used as the classifier.

ACKNOWLEDGMENT

All authors approved the submission of the manuscript and to be published in honor of late Anthony Rotimi Hassan (PhD). The authors acquiesced to publish this manuscript in late memory of Anthony Rotimi Hassan (PhD). It was a selfless effort to honor the demise colleague; Anthony Rotimi Hassan (PhD). Anthony Rotimi Hassan during his time was a participating colleague of the UK-Africa Postgraduate Advanced Study Institute in Mathematical Sciences (UK-APASI).

References

[1] M. T. De Jong, M.T.C. Eijkemans, B.V. Calster, D. Timmerrman, K.G.M. Moons, E.W. Steyerberg, and M.V. Smeden, "Sample size considerations and predictive performance of multinomial logistic prediction models," Statistics in Medicine, vol. 38(9), 2019, pp. 1601-1619. doi: 10.1002/sim.8063.

[2] M. Fatima, and M. Pasha, "Survey of machine algorithms for disease diagnostic," Journal of Intelligent Learning Systems and Applications, vol. 9(1), 2017. doi: 10.4236/jilsa.2017.91001.

[3] K.M. Ghori, M. Imran, R.A. Nawaz-Abbasi, A. Ullah, L. Szathmary, "Performance analysis of machine learning classifiers for non-technical loss detection," Journal of Ambient and Humanized Computing," 2020. doi:10.1007/s12652-019-01649-9. Poor,

[4] T. Jebara, "Machine learning discriminative and generative," In: The International Series in Engineering and Computer Science, 2004. doi:10.1007/978-1-4419-9011-2.

[5] M. E. Kalan, R. Jebai, E. Zarafshan, Z. Bursac "Distinction between two statistical terms: Multivariable and multivariate logistic regression," Nicotine and Tobacco Research," vol. 23(8), 2021, pp. 1446-1447. https://doi.org/10.1093/ntr/ntaa055.

[6] N. T. Lee, P. Resnick and G. Barton "Algorithmic bias detection, and mitigation: Best practices and policies to reduce consumer harms," Brookings Research Report.

[7] S. Patrick, and R.V. Thomas, "Logistic regression in medical research," Anesthesia and Analgesia, vol. 132(2), 2021, pp. 365-366. doi: 10.1213/ANE.0000000000005247

[8] I. H. Sarker, "Machine learning: Algorithms, real-world applications and research directions," SN COMPUT. SCI., vol2(160), 2021. doi: 10.1007/s42979-021-00592-x.

[9] I. H. Sarker, A.S.M. Kayes and P. Watters, "Effectiveness analysis of machine learning classification models for predicting personalized context-aware smartphone usage. Journal of Big Data, vol. 6(57), 2019. doi:10.1186/s40537-019-0219-y.

[10] J. Schmidt, M.R.G. Marques, S. Botti, M.A.L. Marques, "Recent advances and applications of machine learning in solid-state materials science," npj Computational Materials, vol.83, 2019. doi:10.1038/s41524-019-0221-0.

[11] M-S. Shin and J-Y. Lee, "Building a nomogram for metabolic syndrome using logistic regression with a complex sample — A study with 39,991,680 Cases, Healthcare, vol.10 (372), 2022. https://doi:.org/10.3390/healthcare10020372.

[12] M. Yin, D. Zeng, J. Gao, Z. Wu and S. Xie, "Robust multinomial logistic regression based on RPCA," IEEE Journal of Selected Topics in Signal Processing, vol. 12(6), pp. 1144 – 1154. doi:10.1109/JSTSP.2018.2872460.

Contribution of individual authors to the creation of a scientific article (ghostwriting policy)

Rasaki Olawale Olanrewaju carved-out the mathematical derivations as well responsible for the Numerical Analysis Section.

Queensley C. Chukwudum and Sodiq Adejare Olanrewaju gave the introductory part and as well the detailed explanation of each concept.

Sunday E. Fadugba was responsible for the conclusion section and thorough proof reading of write-ups.

Sources of funding for research presented in a scientific article or scientific article itself.

There is no source of funding for this research work. Rasaki Olawale Olanrewaju intend to use his benefits has a reviewer to cover the publication charges that might arises after the manuscript has been accepted.

Creative Commons Attribution License 4.0 (Attribution 4.0 International, CC BY 4.0)