Arab American University

Faculty of Graduate Studies

Using Logistic Regression as a Classifier in Modeling with Normal Mixtures

By

Israa Younes Zaazaa

Supervisor

Prof. Hasan Hamdan Hanashi

This thesis was submitted in partial fulfillment of the requirements for the Master`s degree in Applied Mathematics

July/2019

Thesis Approval

Using Logistic Regression as a Classifier in Modeling with Normal Mixtures

By

Israa Younes Zaazaa

This thesis was defended successfully on 25/ 8 / 2019 and approved by:

Committee members	Signature
1. Prof. Hasan Hamdan Hanashi (Supervisor)	.....
2. Dr. Mahmoud Farah Almanassra (Internal Examiner)	.....
3. Dr. Inad Nawajah (External Examiner)	.....

DECLARATION

I am a student Israa Younes Zaazaa, owner of the university number 21520147 and signed below, I acknowledge that I have complied with all the regulations, instructions, decisions and laws of the Arab American University in force and the scientific and ethical standards accepted, including the scientific secretariat, and that I bear all the consequences and results and responsibility if it shows the contrary to the right of the Deans Council to terminate the degree and certificate I have obtained or will receive based on this decision and without the right to challenge or challenge the decision of the Council of Deans in this regard.

Signature : …… Date: 25/ 8/ 2019

DEDICATION

I dedicate this work to my father, who is my first teacher, he put me at the beginning of this path. To my mother, she encouraged me to continue and gave me all the support.

To my sisters and brothers, to everyone who encouraged me even with a word to complete this achievement.

ACKNOWLEDGMENTS

I would like to express my great thanks and gratitude to Prof. Hassan Hanashi who supervised my thesis with a dedicated effort and drove me with his blessed efforts and knowledge to this achievement.

I would also like to thank all members of the Committee members; Dr. Mahmoud Almanassra and Dr. Inad Nawajah.

ABSTRACT

The classification of observations plays an important role in statistics and all other fields. In this thesis, we studied Logistic Regression (LR) as a method of classification and compare its performance with the performance of Linear Discriminant Analysis (LDA), Gaussian Mixture Model (GMM), and Neural Networks (NN). Performance is compared by the Misclassification Table and Error Rate for each method.

Furthermore, the effect of sample size and presence of correlation were studied. In general, the results showed that when the linear discriminant analysis assumptions are met, the performance of the linear discriminant analysis method is best. If the conditions are not met, the logistic regression method outperforms the other classification methods.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Introduction

1.1 General Introduction

Classification and discrimination are important tools in the analysis of data. However, prediction takes the values of variables as inputs and then gives the closest value to the occurrence (Nikam, 2015). The Classification has recently become popular in regards to applications and computing, which is the focus of the thesis. These tools are used for differentiation and discriminating between observations. There is a slight overlap between the two concepts. The discrimination is numerical, but the classification is designed to separate the different data into classes.

There are two situations of classification. One of them is detecting the classes of objects, or assigning new observations into classes. It is appropriate to label these classes as $\pi_1, \pi_2, \ldots, \pi_j$, where j is the component's number. The data have commonly come in multivariate form, so the observed values of the independent variables $X' = [x_1, x_2, \ldots, x_p]$, where p the number of independent variables, differs from class to the other.

There are real examples of classification situations such as separating between successful or unsuccessful college students, with measured variables like entrance examination scores, high school grade point average, and the number of high school activities. In practice, no matter what there will always be uncertainty in identifying the class membership of the data. That is due to several reasons. First, predicting future values is not deterministic and can never be that way.

Besides getting accurate information from the correct source may lead to the things to be

destroyed. For example, to know the age of a battery, what we should use to exhaust energy stored in them, and then we can know the time necessary to exhaust its energy, which is known as battery life.

Another example, we need prediction to avoid medical errors that may occur during a surgical procedure, these mistakes should have occurred in the past, and the result was a human loss so that we can recognize them and avoid them at present. That is medical errors can't be avoided without their occurrence. This means the inevitability of a loss to obtain accurate information from the correct source.

The assignment of observations or objects into predefined homogeneous groups is a problem of major practical and research interest. It's important to know that classification may have errors and the populations are overlap, and the cost of misclassification is different. For example, we may use quantitative information in predicting who will or will not graduate from the Medical college, an unsuccessful graduate student in a medical college classified as successful is costlier than classifying a successful one as unsuccessful. This would be an example of simple binary classification problems, where the categorical dependent variable can only assume two distinct values. In other cases, there are multiple categories or classes for the categorical dependent variable. Here we will only study a binary dependent variable.

A categorical variable (or nominal variable) has a measurement unordered scale consisting of a set of categories. For nominal variables, the order of listing the categories is irrelevant. The statistical analysis should not depend on that ordering. Methods designed for nominal variables give the same results no matter how the categories are listed (Agresti, 2007). Categorical scales are pervasive in the social sciences for measuring attitudes and opinions. Categorical scales also occur frequently in the health sciences, for measuring responses such as whether a patient survives an

operation (yes, no), the severity of an injury (none, mild, moderate, severe), and stage of disease (initial, advanced).

Although categorical variables are common in the social and health sciences, they are by no means restricted to those areas. Often, categorical data result from n independent and identical trials with two possible outcomes for each, referred to as “success” and “failure.” These are generic labels, and the “success” outcome need not be a preferred result. Identical trials mean that the probability of success is the same for each trial. Independent trials mean the response outcomes are independent random variables. In particular, the outcome of one trial does not affect the outcome of another. These are often called Bernoulli trials.

1.2 Motivation

Classifier effectiveness depends surely on the characteristics of the data to be classified, it's the size and which distribution does it belong to. There is no certain classifier that doses the best in all given scenarios. For example, a simple model may lose some important information and a complex model, even if it fits well with the data, may not give good predictive accuracy. Various data were selected to examine the suitability of the logistic regression model and compare its performance with that of other models. There are many classification methods such as Support Vector Machine (SVM), Principal Component Analysis (PCA), Classification and Regression Trees (CART), Logistic Regression(LR), Linear Discriminant Analysis (LDA), Gaussian Mixture Model (GMM), Neural Networks (NN) and others.

Our focus will be mainly on comparing the performance of LR with that of the LDA, GMM and Neural Networks (NN). This is because LR and LDA are the most commonly used techniques in Machine Learning and Data Science, they are popularity in use, and they have a solid theoretical

basis. On the other side, in Computer Science, NN is commonly used.

1.3 Research Questions

In this study, several research questions will be addressed. Four classification techniques will be presented and compared. The performance of each will be evaluated through the accuracy tables or what is called confusion tables.

1.4 The Objectives of the Study

The goal of this study is classifying new observations by several methods. The fitted model using Logistic Regression (LR), Linear Discriminant Analysis (LDA), Gaussian Mixture Model (GMM) and Neural Networks (NN) will be compared using the corresponding error rates.

These aims are summarized as follows:

1. Using LR as a classifier in four Datasets.
2. Comparing the performance of LR, LDA, GMM and Neural Networks (NN) via normal mixtures using classification tables.

1.5 Methodology

Classification of multivariate normal observations will be done. The logistic regression will be used to discriminate between the components of the mixture. This technique will be compared with other existing discrimination and classification techniques such as the use of a Linear Discriminant Function.

The comparison will be done using real-life data through the classification tables. All statistical

analyses will be done using the R software.

1.6 Organization of the Thesis

In this chapter, a brief background is presented to serve the purpose of the thesis and its methodology. The classification methods used in this thesis were described briefly along with a modest review of the literature. In the second chapter, the theoretical aspects of LDA, GMM and Neural Networks (NN) are presented. In the third chapter, the theoretical framework of LR is presented and the details are discussed. In Chapter 4, the performance of the LR model is compared with that of the other three models, the comparison results are presented in the last chapter using real-life examples.

In the fifth and final chapter, the results will be discussed and the research recommendations will be included for the future**.**

1.7 Literature Review

Many studies that compare the accuracy of each method of classification. These studies did not cover the four methods presented in this thesis. And here are the most relevant studies that address these comparisons:

Kiang (2002) assessed the comparative of classification methods. His study aims to evaluate the performance of classification methods. The study showed that no one method that outperforms all methods all the time. This paper also found that classification methods show sensitivity to changes in data characteristics. He recommended the construction of classification systems using several methods of different classifications to form a hybrid work in response to the existence of different

biases of data. He also recommended studying the characteristics of the data because that may affect the performance of the classification methods. Also, Hossain, et al. (2002) concluded that the efficiency of logistic regression is $\frac{1}{2}$ to $\frac{2}{3}$ times better than the efficiency of the linear discrimination performance in case the data is multivariate normally distributed.

However, Pohar, Blas, and Turk (2004) compared LR and LDA in a simulation study. The objective of their paper was to develop guidelines to determine when to choose the most appropriate method.

They focused on the predictive power of methods based on the estimation of parameters.

The researchers found that the method of linear discriminant analysis (LDA) shows greater superiority when the data follows a normal distribution and fails if the number of categories is small (2 or 3). These differences become puny when the sample size becomes larger. Otherwise, LR is more efficient and appropriate regardless of the distributions. Later on, Maroco, et al. (2011) predicted Dementia using data mining methods, by making a comparison of the accuracy and sensibility of LDA, LR, SVM, CART, Random Forests and Neural Networks (NN).

Researchers in the previous study founded that the linear discrimination analysis method showed high predictive accuracy, sensitivity, and distinguished power among other classification methods. Further, Holden, Finch, and Kelley (2011) compared two-group classification methods and they noticed that increasing the size of the sample led to an increase in the accuracy of classification using LDA and decreased accuracy rates at the CART. They showed that in most of the conditions of the study, LDA has the highest classification truth, while the LR and linear discrimination showed lower ratings. On the other hand, Kiveu (2015) compared LR, LDA with analyzing PCA, principal component analysis. Kiveu (2015) concluded that the logistic regression achieved a slightly higher success rate compared to the linear discriminant analysis of 87% for the logistic

regression, and 85.60% for linear discriminant analysis. The researcher found that the linear discriminant function attained a lower rate of misclassification.

Besides, Yeun Liong and Fan Foo (2013) compared between LDA and LR for data classification. The researchers determined that the method of logistic regression showed predictive ability and a higher percentage of classification than linear discrimination analysis. But if the sample size is large, the computing time for the logistic regression is about 6.44 times higher than the computing time of the other method, so, they showed that overall, LR is better.

Therefore, the researchers recommended using the method of linear discrimination analysis if we have a large sample size and our time is limited, and they explained that the number of independent variables does not have a significant impact on the computing time of the two methods. While Demir (2014) compared CART and LR with other predictive methods in order of identifying patients who are at risk of readmission. This study showed that the predictive ability of the LR was higher than the other methods adopted in this paper, and classification trees had a similar result.

Similarly, Omurlu, et al; (2014) compared the execution of CART and LR with CART and Neural Networks (NN). The study was to compare the performance of classification methods including LR and CART and other methods in diagnosing the presence of albumin in patients with the second type of diabetes. In this study, the logistic regression (LR) was better than CART where the logistic regression (LR) ranked third with a precision of 74.24%, while CART was ranked fourth and final 69.7% accuracy.

At last, Brolin and Finch (2016) followed up a classification in the existence of misclassified practice data. The study aimed to find the effect of classification of structured training data on three groups of misclassifications by using classification methods like LDA and L.R and CART.

The researcher found that CART was ranked second in the most accurate methods in this study, with a large gap between its accuracy and accuracy of the method that won the first place (random forests).

In brief, our study will focus on comparing Logistic Regression (LR), Linear Discriminant Analysis (LDA), Gaussian Mixture Models (GMM) and Neural Networks (NN). The performance of these techniques as classifiers will be investigated through real-life examples.

Chapter 2

Overview of Classification Methods

2.1 Introduction

The task of classifying future observations into predefined homogeneous groups is of interest to researchers in many fields.

In classification procedures, we try to allocate a future observation in one of several predefined groups or categories. The primary purpose of the classification is to describe the features of observations from different known populations, and how to separate observations under two or more classes (Johnson & Wichern, 2007).

The classification classifies the observations to categories by a function of a combination of variables. The aim is to provide a formula of function that leads to optimal discrimination of observations in categories with the lowest error rate (Wehrens, 2011).

Some studies dealt with mixed types of variables. There are some strategies to solve this dilemma, the most important is to convert the variables to be all of the same types, all of them are continuous or all of the categories. But this process may lead to loss of information and this is a drawback (Hamid, 2010).

Rating of classification accuracy is based on 3 criteria: sample design and this depends on specific techniques, design response and finally analysis process (Lu, Weng, 2007)

In this chapter, the theoretical details of the four methods namely Linear Discriminant Analysis (LDA), Logistic Regression (LR), Gaussian Mixture Models (GMM) and Neural Networks (NN) are presented and discussed.

2.2 Linear Discriminant Analysis (LDA)

The first technique or method used in classification is LDA, which was developed in 1936 by Sir Ronald Fisher (Fisher, 1936 indirect). The power of this technique stems from the fact that when two populations are normal with equal covariance matrices, then the method is (asymptotically) optimal; thus, the probability of misclassification is minimum (Rencher, 2003).

A discriminant rule is a separation of the sample spaces, into disjoint sets R_1, R_2,…, R_j such that $S= \bigcup_{k=1}^{j} R_k$. If $x \in R_k$, it is predicted to be a member of population $\boldsymbol{\vartheta}_k$, such that the error of misclassification is as small as possible. In our case, j=2, that is there are only two populations $\boldsymbol{\vartheta}_1$ and $\boldsymbol{\vartheta}_2$.

In particular, let $\boldsymbol{\vartheta}_1$ is the population of observations with the values of X's which classified as the first class, and $\boldsymbol{\vartheta}_2$ are the residual observations which classified as the second class (Johnson & Wichern, 2007). Discriminant analysis is used in situations where the clusters are previously known.

Discriminant analysis aims are using the variables that discriminate between two or more normally occurring groups. Then find a function that can be used to classify future observations and predict its membership in groups based on measured variables.

In LDA, a subject is classified into the group for which its classification function score is highest (Maroco, 2011). Among the advantages that make LDA is good in the classification, the shape and location of the original data does not change if we make a transformation on this data, while using Principle Component Analysis (PCA), it changes its location and shape if we do the same transformation (Balakrishnama, Ganapathiraju, 1998).

LDA searches for the best projection to project the input data, on a lower dimensional space, in which the data of different classes are well-separated as much as possible. The projection maximizes the distance between the means among several competing classes while minimizes the variance within each class (Al-jazzar, 2012) as in Figure 2.1.

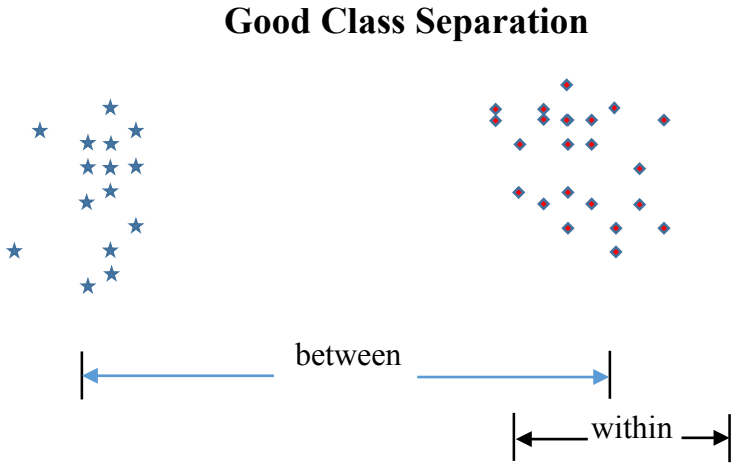


Figure 2.1: Good class separation between 2-classes.

Assume we have binary outcome variable populations, i.e. $\boldsymbol{\vartheta}(x)_j, j = 1, 2$ and we have to allocate an observation x to one of two groups. These groups or populations can be described by probability density function $f_1(x)$ and $f_2(x)$.

Let that the sample space Ω which contains all observations' variables, R_1 is the set for all variables which make an observation classified as $\boldsymbol{\vartheta}_1$, and R_2 is the set for all variables which make an observation classified as $\boldsymbol{\vartheta}_2$. The classification regions R_1 and R_2 are disjoint, i.e. $R_1 \cap R_2 = \phi$ and $R_1 \cup R_2 = \Omega$.

The conditional probability, $P(2|1)$ is the probability of misclassifying an object as it is in $\boldsymbol{\vartheta}_2$ but it is in $\boldsymbol{\vartheta}_1$, so $P(2|1)$ which is the volume formed by the integration of density function $f_1(x)$as the

height, over the classification region R_2, and expressed as :

$$P(2|1) = p(\ x \in R_2 \mid \pi_1) = \int_{R_2} f_1(x)\ dx \qquad (2.1)$$

and similarly for $P(1|2)$.

Let p_1 be the prior probability of $\boldsymbol{\vartheta}_1$, and p_2 the prior probability of $\boldsymbol{\vartheta}_2$. We define the prior probabilities as follows: p_1 is the proportion of observation in $\boldsymbol{\vartheta}_1$, and p_2 is the proportion of observation in $\boldsymbol{\vartheta}_2$, where $p_1 + p_2 = 1$ for every observation.

Then the overall probabilities of correctly or incorrectly classifying objects can be derived as the product of the prior and conditional classification probabilities as following:

$$P(\text{observation is correctly classified as } \boldsymbol{\vartheta}_1) = P(x \in R_1 \mid \boldsymbol{\vartheta}_1)\ p(\boldsymbol{\vartheta}_1) = P(1|1)\ p_1 \qquad (2.2)$$

$$P(\text{observation is misclassified as } \boldsymbol{\vartheta}_1) = P(x \in R_1 \mid \boldsymbol{\vartheta}_2)\ p(\boldsymbol{\vartheta}_2) = P(1|2)\ p_2 \qquad (2.3)$$

$$P(\text{observation is correctly classified as } \boldsymbol{\vartheta}_2) = P(x \in R_2 \mid \boldsymbol{\vartheta}_2)\ p(\boldsymbol{\vartheta}_2) = P(2|2)\ p_2 \qquad (2.4)$$

$$P(\text{observation is misclassified as } \boldsymbol{\vartheta}_2) = P(x \in R_2 \mid \boldsymbol{\vartheta}_1)\ p(\boldsymbol{\vartheta}_1) = P(2|1)\ p_1 \qquad (2.5)$$

The expected cost of misclassification (ECM) is given by

$$\text{ECM} = c(2|1)\ P(2|1)\ p_1 + c(1|2)\ P(1|2)\ p_2, \qquad (2.6)$$

where $c(2|1)$ is the cost of misclassification when an observation from $\boldsymbol{\vartheta}_1$ is incorrectly classified as $\boldsymbol{\vartheta}_2$, and vice versa for $c(1|2)$.

The regions R_1 and R_2 which minimize ECM formed by:

$$R_1 : \frac{f_1(x)}{f_2(x)} \geq \left(\frac{c(1/2)}{c(2/1)}\right)\left(\frac{p_2}{p_1}\right) \quad (2.7)$$

$$R_2 : \frac{f_1(x)}{f_2(x)} < \left(\frac{c(1/2)}{c(2/1)}\right)\left(\frac{p_2}{p_1}\right) \quad (2.8)$$

To assign an observation x to $\boldsymbol{\vartheta}_1$ or $\boldsymbol{\vartheta}_2$, Fisher (1936) employed his Linear Discriminant Function (LDF). To apply the rule, he assumed that $\sum_1 = \sum_2 = \sum$ and no assumption made about the probability density function (PDF). Fisher's rule does not require normality (Timm, 2002) but some favored the use of logistic regression if the hypothesis of the normal distribution of data was excluded (Kiang, 2003). He also assumed that $p_1=p_2$ and c(1|2) = c(2|1), so (2.7) and (2.8) give us:

$$R_1: \frac{f_1(x)}{f_2(x)} \geq 1$$

$$R_2 : \frac{f_1(x)}{f_2(x)} \leq 1$$

In this case, for any new observation x_0 , assigning x_0 to $\boldsymbol{\vartheta}_1$ if $\frac{f_1(x)}{f_2(x)} \geq 1$, otherwise it is assigned to $\boldsymbol{\vartheta}_2$.

2.2.1 LDA Assumptions

Discriminant analysis, like all statistical procedures, is restricted by certain assumptions, and these are:

1. Sample size: There is no objection to existing samples of different sizes, but the size of the set should exceed the number of variables. For example, the maximum number of variables = N -2, where N is the sample size.
2. Normal distribution: it is common that the data are normally distributed with outliers values (Johanson & Wichern, 2007). It's important to know that LDA is highly sensitive

to the inclusion of outliers. The included in the evaluating of LDA (Reimann et al, 2002). It is preferred that the data being normal distributed. Two popular tests check the normality, these are Kolmogorov-Smirnov (K-S) and shapiro-Wilk test.

3. Homogeneity of variances/covariances: LDA is strongly affected by the heterogeneity of the covariance matrices of all groups, it requires the equalization of the variance matrices for all of them (Al-jazzar, 2012)

   $(\text{i.e. } \Sigma_1 = \Sigma_2 = \ldots = \Sigma_k)$.

4. Non-Multicollinearity: LDA assumes the independence of variables. The violation of this hypothesis as the presence of dependent variables correlated with one of the variables, or the existence of a variable as a function of another variable; leads to the fact that the coefficients will not estimate the relative weight correctly (Poulsen & French, 2002).

2.2.2 Classification with Two Multivariate Normal Populations

Parametric procedures are used when classification rules make assumptions regarding the *pdfs* $f_1(x)$ and $f_2(x)$, such as normality because it's reasonability and simplicity across population models (Johnson, Wichern, 2007).

While Fisher's LDF is nonparametric, it is only asymptotically optimal under normality and requires the covariance matrices to be equal (Timm, 2002).

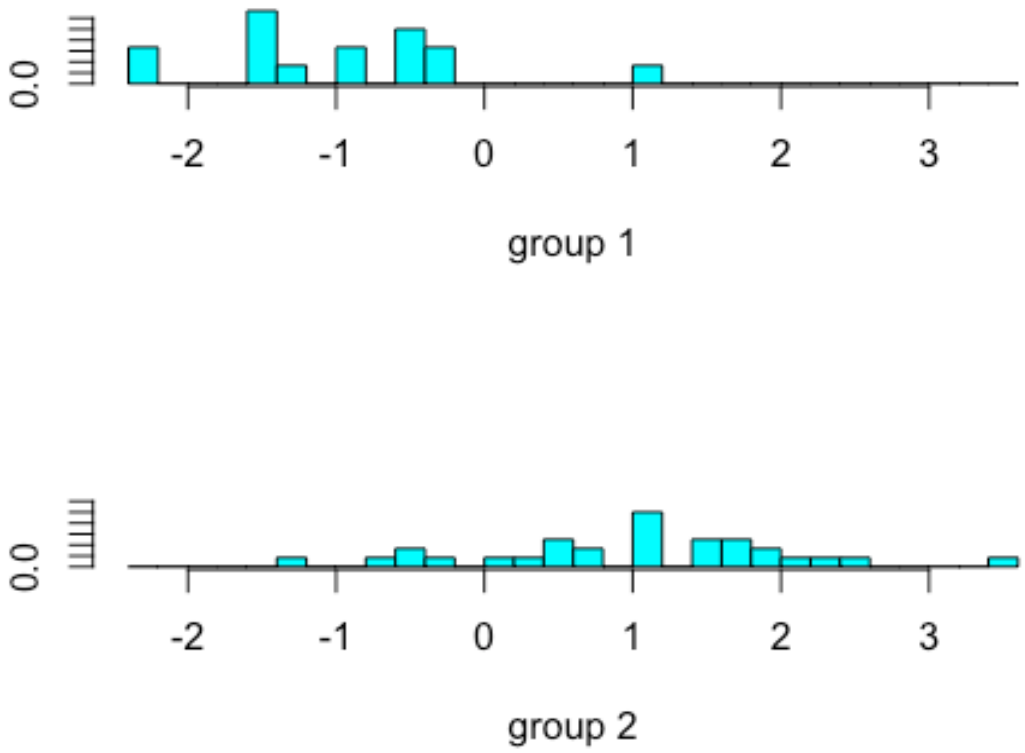


Figure 2.2: Plot of LDA for the Depression dataset.

In our case, we will restrict ourselves and assumes that $f_1(x)$ and $f_2(x)$ are multivariate normal densities. I.e. $\boldsymbol{\vartheta}_1$ is $\mathcal{N}_p(\mu_1, \Sigma)$ and $\boldsymbol{\vartheta}_2$ is $\mathcal{N}_p(\mu_2, \Sigma)$, with

$\mathbf{X} = (x_1, x_2, \ldots, x_p)$, then the joint densities for populations $\boldsymbol{\vartheta}_1$ and $\boldsymbol{\vartheta}_2$ expressed by

$$f_i(\mathrm{X}) = \frac{1}{(2\pi)^{p/2}} \frac{1}{|\Sigma|^{1/2}} \quad exp\left[-\frac{1}{2} \ (\mathrm{X} - \mu_i)` \ \Sigma^{-1} \ (\mathrm{X} - \mu_i)\right], \text{ for } i = 1,2 \qquad (2.9)$$

For any realization x of **X,** suppose that the population parameters $(\mu_1, \sum)$ and $(\mu_2, \sum)$ are known. Then the minimum Expected Cost of Misclassification (ECM) regions found in (2.7) and (2.8) become:

$$\mathrm{R}_1: \frac{-1}{2} \ (x - \mu_1)' \ \sum^{-1} \ (x\text{-}\mu_1) + \frac{1}{2} \ (x - \mu_2)' \sum^{-1} \ (x\text{-}\mu_2) \geq \ ln\left(\frac{\mathrm{c}(1/2)}{\mathrm{c}(2/1)}\right) \ \left(\frac{p_2}{p_1}\right) \qquad (2.10)$$

$R_2: \frac{-1}{2}(x-\mu_1)' \sum^{-1}(x-\mu_1) + \frac{1}{2}(x-\mu_2)' \sum^{-1}(x-\mu_2) < ln\left(\frac{c(1/2)}{c(2/1)}\right)\left(\frac{p_2}{p_1}\right)$

(2.11)

The allocation rule that minimizes the ECM is as follows:

x_0 allocated to $\boldsymbol{\vartheta}_1$ if:

$$(\mu_1 - \mu_2)' \sum^{-1} x_0 - \frac{1}{2}(\mu_1 - \mu_2)' \sum^{-1} (\mu_1 + \mu_2) \geq ln\left[\left(\frac{c(1/2)}{c(2/1)}\right)\left(\frac{p_2}{p_1}\right)\right] \quad (2.12)$$

Otherwise x_0 is allocated as $\boldsymbol{\vartheta}_2$.

In most popular cases, the population quantities μ_1, μ_2 and $\sum$ are unknown, the rule (2.12) must be adjusted. Wald and Anderson (1984), substituted the population parameters by their sample estimates (Johnson and Wichern, 1999).

Suppose, then that we have n_1 observations from $\boldsymbol{\vartheta}_1$, and n_2 observations from $\boldsymbol{\vartheta}_2$, such that $n_1 + n_2 - 2 \geq p$, each of the form $X` = (x_1, x_2, \ldots, x_p)$. Then the sample mean vectors and covariance matrices are determined by

$$\bar{x}_1 = \frac{1}{n_1}\sum_{j=1}^{n_1} x_{1j}, \quad \bar{x}_2 = \frac{1}{n_2}\sum_{j=1}^{n_2} x_{2j}$$

$$S_1 = \frac{1}{n_1-1} \sum_{j=1}^{n_1} (x_{1j} - \bar{x}_1)(x_{1j} - \bar{x}_1)', \quad S_2 = \frac{1}{n_2-1} \sum_{j=1}^{n_2} (x_{2j} - \bar{x}_2)(x_{2j} - \bar{x}_2)'$$

Since it is assumed the present population have the same covariance matrix $\sum$, the sample covariance matrices S_1 and S_2 are pooled to conclude unbiased estimate of $\sum$.

$$S_{pooled} = \left[\frac{n_1-1}{n_1-1+n_2-1}\right] S_1 + \left[\frac{n_2-1}{n_1-1+n_2-1}\right] S_2 \quad (2.13)$$

Substituting $\bar{x}_1$ for μ_1, $\bar{x}_2$ for μ_2 and S_{pooled} for $\sum$ in (2.11) gives the sample classification rule:

x_0 allocated to π_1 if:

$$(\bar{x}_1 - \bar{x}_2)' S_{pooled}^{-1} x_0 - \frac{1}{2} (\bar{x}_1 - \bar{x}_2)' S_{pooled}^{-1} (\bar{x}_1 + \bar{x}_2) \geq \ln\left[\left(\frac{c(1/2)}{c(2/1)}\right) \left(\frac{p_2}{p_1}\right)\right] \qquad (2.14)$$

Otherwise x_0 allocated as π_2.

In summary, Fisher's rule has some optimal properties if $\sum_1 = \sum_2$, $p_1 = p_2 = \frac{1}{2}$, and $c(1/2) = c(2/1)$ as long as samples are acquired from normal populations. When this is not the case, Fisher's rule is to be averted (Timm, 2002)

2.3 Gaussian Mixture Models (GMMs)

For Gaussian Mixture Models, in particular, we'll use 2D normal distributions, meaning that our input is now a vector instead of a scalar. This also changes our parameters: the mean is now a vector as well. The mean represents the center of our data so it must have the same dimensionality as the input.

The variance changes less intuitively into a covariance matrix Σ. The covariance matrix, in addition to telling us the variance of each dimension, also tells us the relationship between the inputs, i.e., if we change *x*, how does *y* tend to change?

We can plot the 2D Gaussian, as we see in Figure 2.3, The X and Y axes are the two inputs and the Z-axis represents the density. The lower plot is a contour plot. The lighter the color, the larger the density, the larger is the enclosed probability.

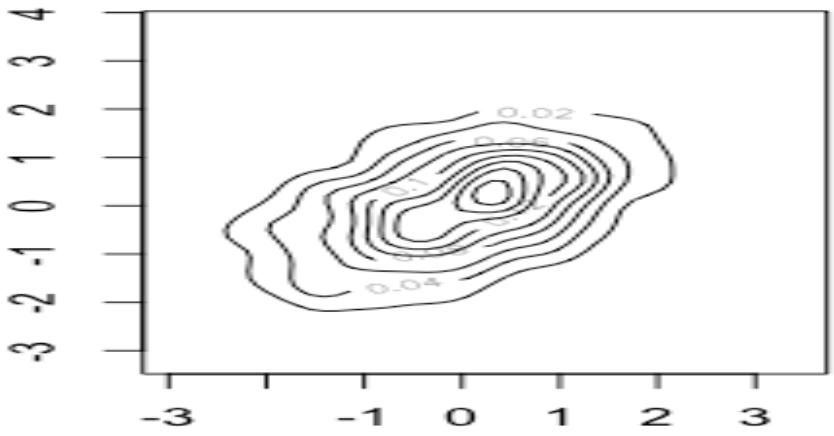


Figure 2.3: Contour plot that shows this our contour Gaussian in 2D.

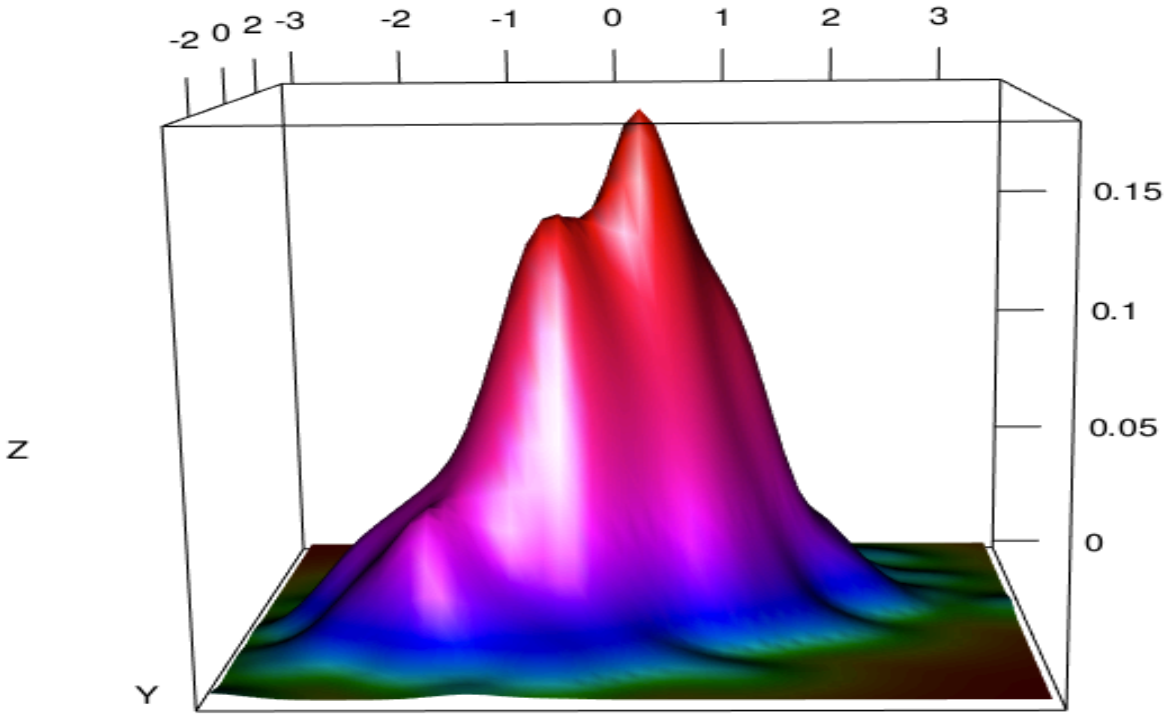


Figure 2.4: the Surface plot that shows the 3D density.

If the datasets are not normal, then CLT can be used. In our case and since we know these data are normally distributed, why not try to fit the normal case to them? The idea behind Gaussian Mixture

Models, GMM, is to find the parameters of the Gaussians that explain our data better. However, determining the optimal number of ingredients is not easy (Zhang, Huang, 2015).

As a condition of using this road is to determine the optimum count of the components to get good performance in the classification. Few of them reduce the performance of this method, especially if training data are already small. And if the components are large, it means that the complexity of calculations will be overly exaggerated and this is something we do not seek at all (Huang et al., 2005).

The most common method is the maximum likelihood technique because of its optimal properties. This method is the most widely used among other proposed methods (Oliveira, Martins, 2005).

We are assuming that these data are normally distributed and we want to find parameters that maximize the likelihood of observing these data. In other words, we consider each point as being generated by a mixture of normal distributions and can compute that probability as follows:

$$P(x) = \sum_{j=1}^{k} \emptyset_j \ \mathcal{N}_j(x;\ \mu_j\ ,\ \sum_j), \tag{2.18}$$

where k is the number of components of normal mixtures, and the result is $3k$ groups of parameters to be estimated (Cao, 2010), such that each $\emptyset_j$ is positive and $\sum_{j=1}^{k} \emptyset_j = 1$ (Anifowose, 2012), and every ϕ_j is the weight of the j^{th} normal, which represents the size of contribution of that component. We have three sets of different parameters that we need to estimate: the weights for each normal component ϕ_j, the means of the j^{th} normal component μ_j, and the covariances of each j^{th} normal component $\sum_j$.

$$\mathcal{N}_j(x;\ \mu_j\ ,\ \sum_j) = \frac{1}{(2\pi)^{|x|/2}\sqrt{|\Sigma_j|}}\ \exp\left(-\frac{1}{2}\ (x-\mu_j)'\ \sum_j^{-1}\ (x-\mu_j)\right)\ ,\ x \in \mathbb{R}\ ,\ \mu \in \mathbb{R},$$

and $\Sigma > 0$. (2.19)

What characterizes GMM is that it treats the data as a linear combination of Gaussian distributions, either uses one single probability distribution for all data or completely ignores its structure (Cao, 2010).

To classify an observation in GMM, we use a Bayesian maximum likelihood classifier (Povinelli & Johnson, 2004). As θ represents the parameters, the component is determined by calculating the conditional probability of each model for every single observation and adopting max $P(x^n; \theta)$ as follows:

$$\underset{\theta}{\text{argmax}}\ log\ P(x^n;\ \theta) = \underset{\theta}{\text{argmax}} \log\left(\prod_i P(x_i|\ \theta)\right)$$

$$= \underset{\theta}{\text{argmax}}\ \sum_i log(P(x_i|\ \theta)\ ) \qquad (2.20)$$

If we try to directly solve for these, it turns out that we can find closed-forms. But there is one huge catch: we have to know the ϕ_j's. In other words, if we knew exactly which combination of the normal mixture a particular point was taken from, then we could easily figure out the means and covariances.

But this one critical flaw prevents us from solving GMMs using this direct technique. Instead, we have to come up with a better approach to estimate the weights, means, and covariances. We will estimate these parameters by using the well-known Expectation-Maximization (EM) Algorithm and applying it to our sample.

2.4 Neural Networks (NN)

The idea of naming this method came from the principle that neural networks work in the brain and are therefore designed to simulate the way the brain works and Neural networks are powerful

tools in classifying patterns in terms of artificial intelligence (Russell and Norvig,2003), also it is nonlinear because its idea is inspired by the synthesis of biological neurons in its nonlinear form (Tiwary, 2014). The history of neural networks began in the 1940s (Kriesel, 2011)

A neural network is a massively parallel distributed processor made up of simple processing units that has a natural propensity for storing experiential knowledge and making it available for use (Haykin, 2008). In this way, the weights in the network are adjusted to achieve the goal.

Neural networks have succeeded in the classification especially for data on the topics of bankruptcy forecasting, handwriting recognition and speech, diagnosis of faults and medical diagnoses, valuation of bonds and others.

Advantages of Neural Networks:

1- Nonlinearity: Their neuronal cells are connected in a non-linear network, which means they are distributed throughout the retina and function abnormally.
2- The possibility of mapping inputs and outputs: The weights of the neural networks are modified to reduce the difference between the actual response and the desired response, where the network is trained several times to reach the stability in the weights of networks.
3- Adaptability: Neural networks deal easily with minor changes. If the variables are not stable, we can change the weights. The availability of this feature ensures our system stability, which means a more stable performance. Adaptation does not mean durability.
4- Contextual information: This means that the neuron may be affected by the global activity of all neurons in the network.
5- Fault tolerance: This ability to bear the error, because of the nature of the complex and wide distribution of the damage without showing a catastrophic level of performance.

Neural networks (NN) have other advantages like the Uniformity of Analysis and Design, Neurobiological Analogy, and Evidential Response.

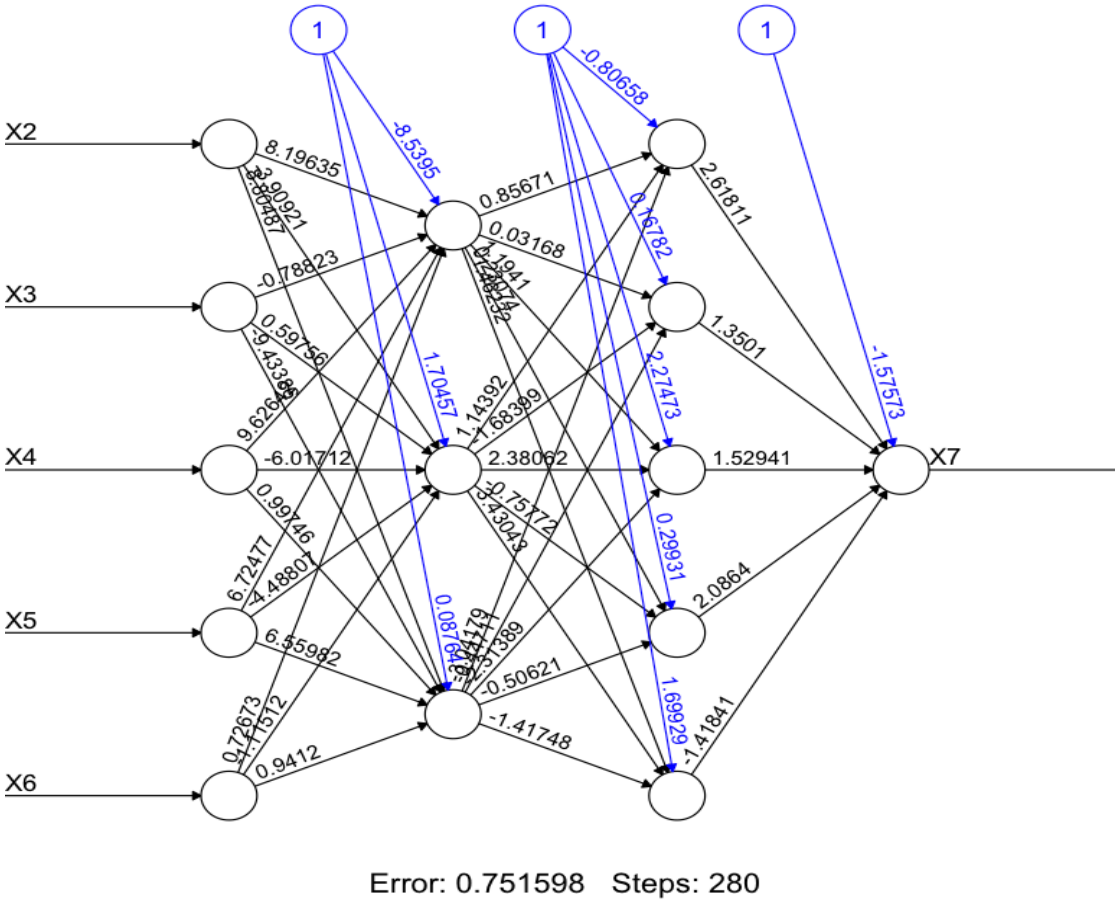


Figure 2.5: Plot of Neural Networks for the Depression Dataset.

As we see, neural networks are a nonlinear model. The range of the output of a neuron is as the unit interval [0,1].

As we note from Figure 2.5, the neural networks (NN) have 3 elements:

1- Input layer: A set of synapses and its function is the interconnecting, and the *m* input signals are x_j's where j=1, …m. In Figure 2.5 input signals are x_1, x_2, …, x_6. Each tangled node

produces a weight symbolized by w_{kj}, where k indicates for the neuron, and j indicates to the signal (Haykon, 2008). Weights are the numbers listed above the stock in Figure 2.5.

2- Hidden layer: A collector that Synthesizes the weighted input signals, in the sense that it forms a linear combiner. The collector of the neural networks (NN) for the Depression Data is x_7, and it needed 280 steps as shown in Figure 2.5.

3- Output layer: The Activation Function $\varphi(.)$ (not shown in Figure 2.5) determines the allowable range in the output signal to a finite value which is as the closed unit interval [0,1] (Haykon, 2008).

The Neural Networks (NN) model may contain a certain amount of bias (b_k) affects the net input value of the Activation Function either by increasing or decreasing. This is expressed mathematically as follows: $u_k = \sum_{j=1}^{m} w_{kj} x_j$ (2.21)

and the output of the Activation Function is $y_k = \varphi(u_k + b_k)$. (2.22)

By the Threshold Function: The Activation Function is:

$$\varphi(u_k + b_k) = \begin{cases} 1 & if\ uk + bk \geq 0 \\ 0 & if\ uk + bk \leq 0 \end{cases} \quad (2.23)$$

We will talk about the Logistic Regression (LR) method in some details in the next chapter.

2.5 Datasets

We have 4 datasets, we want to compare the performance of the four methods (LDA, Mixture models, Neural Networks (NN), and LR) and observe the accuracy of performance according to the Error rates.

In what follows are the details of these four datasets. These datasets have different sizes, different

class variables, and different numbers of independent variables.

Example 1:

The first dataset is the Depression Dataset, which consists of 45 penitents (15 ill and 30 well) with one categorical dependent variable coded to 1 = ill and 2 = well, with five qualitative independent variables coded according to the penitents' answers of these five questions:

1.Have you recently felt that you are playing a useful part in things?

1 =No to 4 =Yes

2. Have you recently felt contented with your lot? 1= No to 4 =Yes

3. Have you recently felt capable of making decisions about things?

1=No to 4 Yes.

4. Have you recently felt that you're not able to make a start on anything?

1= Yes to 4 =No

5. Have you recently felt yourself dreading everything you have to do?

1 =Yes to 4 =No

Example 2:

The second dataset is the Bankruptcy Data. The annual financial data listed in the attached data file, ($\pi = 0$ is for bankrupt, and $\pi = 1$ for no bankrupt firms) have been analyzed by Johnson, W

(1987). Their analysis was focused on studying the impact of outlier observations on discriminant analysis. The Bankruptcy Dataset has four variables which are:

1. CF/TD which indicates to (cash flow/total debt).
2. NI/TA which indicates to (net income/total assets).
3. CA/CL which indicates to (current assets/ current liabilities).
4. CA/NS which indicates to (current assets/net sales).

Example 3:

The third dataset is the Frog Data that is a new data set that was collected by the Department of Biology at James Madison University in the United States. Data occurred from 2010-present. The dependent variable CLADE.DES is based on DNA sequence data or locality collected. The data classified 534 frogs into 21 classes, some of which remained unknown, and adopted 80 variables, reduced to two classes of frogs, with 10 variables and 100 observations. The variables in Frog Data (measured with cm) are quantitative. Character abbreviations are:

1. ED - Eye diameter
2. HL - Head Length
3. HW - Head Width
4. LAL – Lower arm length
5. MN - Mandible-nostril distance
6. TBL - Tibia Length
7. UEW - Upper Eyelid Width
8. OL - Odontoid Length (measured from the bottom of the mandible to the top of the odontoid)
9. MD - Mandible depth (the thickness of the mandible at the base of the odontoid. This is used to calculate the actual length of the odontoid by subtracting MD from OL)

Example 4:

The last dataset is Women Empowerment dataset from UNDP and CIA sites which published in 2004. These data ranked 177 countries of the world if they had the advantage of women's empowerment or not through 134 variables reduced to 7 quantitative and qualitative variables which are mixed of:

1. Log GDP: which is converted into purchasing power parity terms to eliminate differences in national price levels.
2. Education enrollment: which is the combined gross enrollment in primary, secondary, and tertiary schools.
3. Power position: which is female legislators, senior officials, and managers.
4. Employment: which is female professional and technical workers.
5. Income: This is the ratio of female to male earned income.
6. Empowerment: which is the categorical – dependent variable, i.e. it's coded by 0 to countries which classified as one where women are not empowered, and the opposite for 1.

In this chapter, the theoretical aspects of the three classification methods which are LDA, GMM, and NN were discussed.

In the next chapter, theoretical aspects of logistic regression will be presented.

Chapter3

Binomial Logistic Regression.

3.1 introduction

The regression methods aim is to describe the relationships between one dependent variable and one or more than independent variables. Some of these methods are linear regression and others are nonlinear such as logistic regression.

Logistic regression (LR), is a statistical method for analyzing datasets in which there are independent variables used to predict the outcome of a dichotomous variable (Titterington, 1985). Logistic regression is commonplace in many fields such as education, sciences, health, biology, etc. (Barkinson & Read, 1985).

Logistic regression has two wide applications in research, one of them is classification (Predicting membership of new observations into labeled classes), and separation (Distinguish observations according to its nature) (Tensey, et al., 1996).

Logistic regression classifies observations by constructing a regression function that enables us to predict and requires that the dependent variable be two levels, success, and failure. This is similar to the binomial theory requirement (Wang, 2008). Since the objective of the regression method is to find and describe the relationship between independent and dependent variables. The logistic regression is an appropriate method to examine the hypotheses about the nature of the relationship between the dependent variable in the case of categorical, and independent variables, whether they were continuous or categorical.

If we have k independent variables, the mean of Y given $X=x$, $E(Y|X=x)$, can be expressed as $E(Y|X=x) = \beta_0+\beta_1 x_1+\ldots + \beta_k x_k$, where $\beta_0, \beta_1, \ldots, \beta_k$ are partial slopes coefficients of the

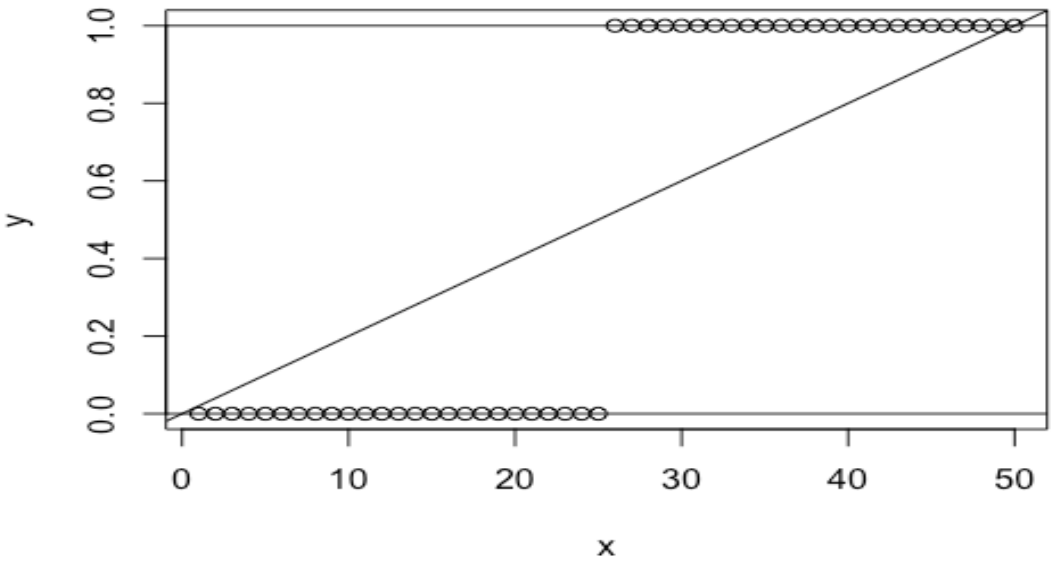
corresponding x_1 ,x_2 , … , x_k independent variables.

The explicate form of the equation is:

$Y = E(Y|X=x) + \varepsilon$, i.e $Y= \beta_0+\beta_1 x_1+... + \beta_k x_k + \varepsilon$, if we have more than one explanatory variable where β_0 is the intercept, $\beta_1,...,$ β_k are regression coefficients, and ε is the error term, where is assumed to be independent and normally distributed with mean zero and constant variance.

In linear regression, the main estimate method is the ordinary least squares, but in logistic regression, the main method is the maximum likelihood to estimate the coefficients. The differences between logistic regression and linear regression are summarized below (Hosmer, Lemeshow,2000):

1- The outcome variable in logistic regression is binary or dichotomous, so the value of the predicted value $\pi(x)$ = y may even be outside the permissible range of 0 to 1, as shown in Figure 3.1



i. Figure 3.1: Linear regression for a binary dependent variable.

2- In any regression problem, the mean value of the outcome variable, given the value of the

independent variable is the key quantity. This quantity is called the conditional mean and which is expressed as *"E* (Y|X= *x)",* where *Y* denotes the outcome variable and x denotes a value of the independent variable. In linear regression, we assume that this mean may be expressed as a linear equation in *x* (or some transformation of *x* or *Y*), such that *E* (Y|X= *x*)*, ranges* between (- ∞, +∞) if β not equal to zero with taking any value of *x.* In dichotomous data, conditional mean must be greater than or equal to zero and less than or equal to 1, i.e.

[$0 \leq E(Y|X= x) \leq 1$].

3- Another difference between the linear and logistic regression models is the conditional distribution of the outcome variable. In the linear regression model, $Y= E(Y|X= x)+ \varepsilon$, such that the error expresses an observation's deviation from the conditional mean. The most common assumption is that ε follows a normal distribution with mean zero and some variance that is constant across levels of the independent variable. It follows that the conditional distribution of the outcome variable given *x* will be normal with mean $E(Y|X= x)$, and a constant variance. This is not the case with a dichotomous outcome variable, the value of the outcome variable given *x* as $Y = \pi(x) + \varepsilon$. Here the quantity may assume one of two possible values. If y = 1 then $\varepsilon = 1 - \pi(x)$ with probability $=\pi(x)$ which is equal to $E(Y|X= x)$, and if y = 0 then $\varepsilon = \pi(x)$ with probability = $1- \pi(x)$. Thus, ε has a distribution with mean zero and variance equal to $\pi(x)[1- \pi(x)]$. That is, the conditional distribution of the outcome variable follows a binomial distribution, not the normal distribution, with probability given by the conditional mean, $\pi(X)$.

3.2 The logistic regression model

We want to find the probability that a case is categorized as one of two categories of outcomes. Let us assume that we have one independent variable X; when $X=x$, Y could be 0 or 1. So let us define $\pi(x) = P(Y=1|X=x) = 1-P(Y = 0|X = x)$, if we know one probability, we know the other. The expected value of Y_i given $X=x$ is, by definition of expectation,

$E(Y_i|X=x) = 1\times P(Y_i=1) + 0\times P(Y_i=0)$. This reduces to $E(Y_i|X=x) = P(Y_i=1|X=x) = P_i$.

$E(Y_i|X=x) = E(\beta_0+\beta_1 x_i+ \varepsilon_i) = E(\beta_0)+E(\beta_1 x_i)+E(\varepsilon_i) = \beta_0+ \beta_1 x_i$, so we get that

$P_i = \beta_0 + \beta_1 x_i = \pi(x)$. We conclude that $\pi(x) = \beta_0+\beta_1 x_i$.

Unfortunately, this is not a good model. As long as $\beta_1 \neq 0$, extreme values of x will give values of $E(Y_i|X=x) = \beta_0+\beta_1 x_i$ that is inconsistent with the fact that $0 \leq p \leq 1$.

To solve this problem, we substitute the odds that $Y=1$ for the probability that $Y=1$, with using the natural logarithm. The odds that $Y=1$, written odds($Y=1$), is the ratio of the probability that y=1 to the probability that $Y\neq 1$. The odds that $Y=1$ is equal $\frac{p(Y=1)}{p(Y=0)} = \frac{p(Y=1)}{1-p(Y=1)}$. We use the term log odds for this transformation. We model the log odds as a linear function with one explanatory variable: $\ln\left(\frac{p}{1-p}\right) = \beta_0+\beta_1 x$. So the natural logarithm of the odds, $\ln\left(\frac{\pi(x)}{1-\pi(x)}\right)$, is called the logit of Y. Now, extend this case to k independent variables $\boldsymbol{X}= (x_1, x_2,\ldots, x_k)$, therefore, the logistic regression model expressed as :

$$\ln\left(\frac{\pi(x)}{1-\pi(x)}\right) = \beta_0 +\beta_1 x_1+ \ldots + \beta_k x_k = \beta' \mathbf{X} \qquad (3.24)$$

which is the logistic regression model.

We will convert the logit and return to the odds via exponentiation for both sides of (3.19), so we get

$$\frac{\pi(x)}{1-\pi(x)} = e^{\beta' x} \qquad (3.25)$$

By Cross-product, the equation will be: $\frac{1-\pi(x)}{\pi(x)} = \frac{1}{e^{\beta' x}}$,

$\frac{1}{\pi(x)}$ -1 = $\frac{1}{e^{\beta' x}}$, so we get $\frac{1}{\pi(x)}$ = 1+ $\frac{1}{e^{\beta' x}}$.

Solving for $\pi(x)$, we get $\pi(x) = \frac{e^{\beta' x}}{1+e^{\beta' x}}$. (3.26)

Whatever the value of β's and the X's, it is sure that the numerator is less than the

denominator in (3.26), and thus ensure that the value of π(x) is limited between 0 and 1.

Aldrich and Nelson showed that according to the previous formula (3.26), values of $\beta' x$ are ranging from negative infinity to positive infinite and this gives the Gaussian distribution S-shape (Cabrera,1994).

3.3 Fitting logistic regression model

If Y is coded as 1 or 0, then $\pi(x)$ equals to the conditional probability when Y =1 given *x* which is donated by *P*(Y=1|X=*x*), and the conditional probability when Y=0 given *x* which donated as *P*(Y=0 | *x*) equals 1-π(x).

To fit the logistic regression model in equation (3.19) using the data with the unknown parameters $\beta_0+\beta_1 x+\ldots+\beta_k x_k$, that requires that we estimate the values of the unknown

parameters. Maximum likelihood is a common method of estimation; maximum likelihood estimates of the parameters are the values that maximize the probability of getting the observed set of data for a given *x*. Below is the outline of the steps of obtaining the Maximum Likelihood Function (MLE). Finding probability by Maximum Likelihood is preferred because it is more practical, especially for the data that violate the normality assumption (Press, Wilson, 1978).

First, we have to find the likelihood function. Since the dependent variable is a dichotomous variable, then we start with Bernoulli trials. A convenient way to express the contribution to the likelihood function for the pair (x_i ,y_i) obtained as follows:

since for a given $X_i = x_i$, y_i =1 with probability $\pi(x)$ and y_i =0 with probability 1- $\pi(x)$, so the distribution of y_i for a given $X= x_i$ can be written as:

$$\pi(x_i)^{y_i}\,[1 - \pi(x_i)]^{1-y_i} \qquad (3.27)$$

Since the observations are assumed to be independent, the likelihood function is obtained as the product of the terms given in expression (3.28) as follows:

for all (x_1, y_1), …, (x_n,y_n), the likelihood function performed as

$$L(\beta)= P(y_1, \ldots, y_n| x_1, \ldots, x_n) = \prod_{i=1}^{n} \pi(x_i)^{y_i}\,[1 - \pi(x_i)]^{1-y_i} \qquad (3.28)$$

The principle of maximum likelihood states that we use as our estimate of β as the value which maximizes the expression in equation (3.28).

Secondly, the *log likelihood* is obtained as:

$$\ell(\beta)= \ln L(\beta) = \sum_{i=1}^{n}(y_i\,[\ln \pi(x_i)] + [1-y_i\,] \ln[1 - \pi(x_i)]) \qquad (3.29)$$

So, we maximize $\ell(\beta)$ instead of ln L(β) because it is easier to work with given that it attains its

maximum at the same point as that of $\ell(\beta)$. That is because the natural log is one to one function as its always increasing.

Thirdly, $\ell(\beta)$ is a differentiable as a function of β. Therefore, to find the value of that maximizes we differentiate function concerning the *k*+1coefficients and set the resulting expressions equal to zero. There will be *k*+1 likelihood equations that are obtained by differentiating the log likelihood. The likelihood equations that result may be expressed as follows:

$$\sum[\, y_i - \frac{e^{\beta' x}}{1+e^{\beta' x}}] = 0 \qquad i = 1, \ldots, n \tag{3.30}$$

$$\sum(x_{ij}\,[y_i - \frac{e^{\beta' x}}{1+e^{\beta' x}}]) = 0 \quad i = 1, \ldots, n,\ j = 1,2,\ldots, k \tag{3.31}$$

The value of β given by the solution to equations (3.30) and (3.31) is called the maximum likelihood estimate and is denoted as $\hat{\beta}$.

Chapter 4

Comparing Logistic Regression with Other Commonly Used Models

4.1 Introduction

In our study, the classification error rate is used as an evaluation method. The performance of the four classification methods used will be evaluated by comparing their classification rates. Factors that may contribute to the performance will be also investigated. These factors are the sample size and the correlation between the explanatory variables. (I did this in section 4.3)

The process is performed 15 times, the classification error is found for each case and then the average of the 15 times is taken as the classification error rate which is done for all datasets with all methods. Then we compared the error of the classification of each data for the four methods.

In particular, four datasets are used and the performance of each method is reported. All computations are done using the statistical package R and explore the factors that might impact the performance of these classification methods.

4.2 Using four Real Data Sets for Comparing

All four classification methods are statistical methods designed to predict the value of a category variable based on the given values of independent variables. These techniques are widely used in medical and economic sciences.

We will conduct the comparison process through five real datasets which described in chapter 2 with a very brief description:

1- The Depression Dataset: This data is used to predict whether a person is in a depression state or not based on his answers to 5 questions, data has 45 observations.

2- The Bankruptcy Dataset: This data is used to predict the company's occurrence in bankruptcy depending on several variables. This was done for 46 companies. The Data constructed by four variables. The source: 1968, 1970, 1971, 1972 Moody's Industrial Manual.

3-Frog Dataset: This data distinguishes between two types of frogs according to biological measurements of the head, eye, arm, the distance between nostrils and mandible, and the decomposition of the lower teeth and measurements in the unit centimeter.

4- Women Empowerment: This data is from UNDP and CIA sites which published in 2004. Its aim is to rank 177 of countries around the world if they empower their women or not based on 7 variables.

4.2.1 Testing LDA assumptions

Since real-world data are usually rough and may not follow the required pattern, the model examination of the conditions and assumptions should be checked because this affects its performance. The assumptions are:

1. Normality:

Which means that the variables are normally distributed or approaching the normal distribution. There are many tests of normality, the one we used here is the Shapiro-Wilk test. The effect of the normality or lack of normality will be studied and investigated. So we have two hypotheses:

Null Hypothesis: The variables follow the normal distribution.

Alternative Hypothesis: The variables do not follow the normal distribution.

We hope that the value of the p-value will be > 0.05 to exit the rejection region at the 0.05 level of significance.

2. Equality of the covariance matrix

Suppose we have n_1 objects from π_1 and n_2 objects from π_2, with a vector of multivariate

independent random variables $\grave{X}= [x_1, x_2, \ldots, x_p]$, such that $n_1+ n_2- 2 \geq p$, then the unbiased the pooled estimator of covariance matrix obtained in equation (2.13) which we inferred obviously in Chapter 2.

And this is done by Box's M test, with two hypotheses:

Null Hypothesis: The covariance matrices for variables are equal, i.e. $\Sigma_1= \Sigma_2$.

Alternative Hypothesis: The covariance matrices for variables are not equal, i.e. $\Sigma_1 \neq \Sigma_2$.

We also hope to have the p-value > 0.05 at the 0.05 level of significance.

Another assumption is that the independent variables are not highly correlated to each other, so the correlation between each pair of variables must be found.

Exploratory Data Analysis:

Exploring and describing the important features and data screening is the start in any data analysis process. This is done to find out if there are any errors in the data, such as missing information correct alignment of the variables, outliers, etc. Besides, it gives us a rough and visual idea about the correlation between them, which can be noticed by scatter plots.

Scatter plots help us detect the nature of the trend between each pair of variables as linear or nonlinear. Scatter plots show us the trends between the explanatory variables.

For the Depression dataset, all explanatory variables are qualitative, so we can't see if there is any linear relationship between them.

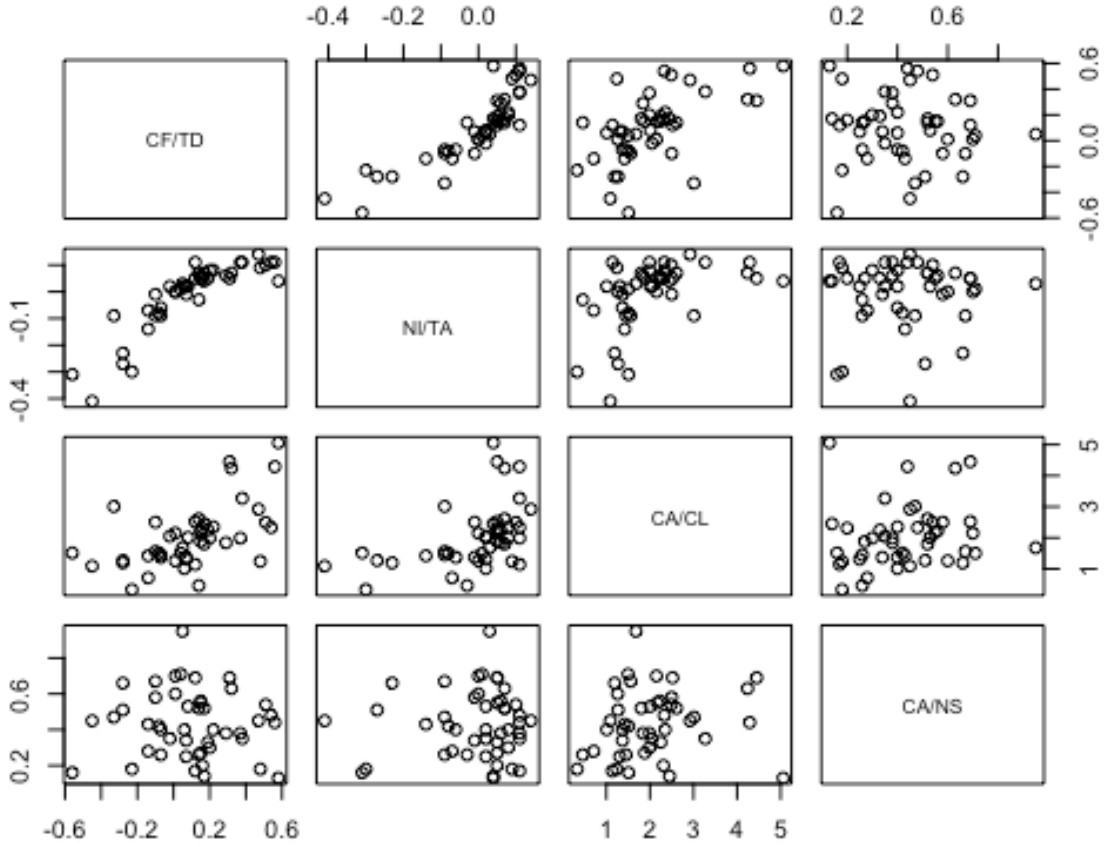


Figure 4.1: Scatter plot of the Bankruptcy Dataset.

Figure 4.1 is a matrix of scatter plots between variables in the Bankruptcy Dataset. It shows that there is a positive linear relationship between CF/TD and NI/TA.

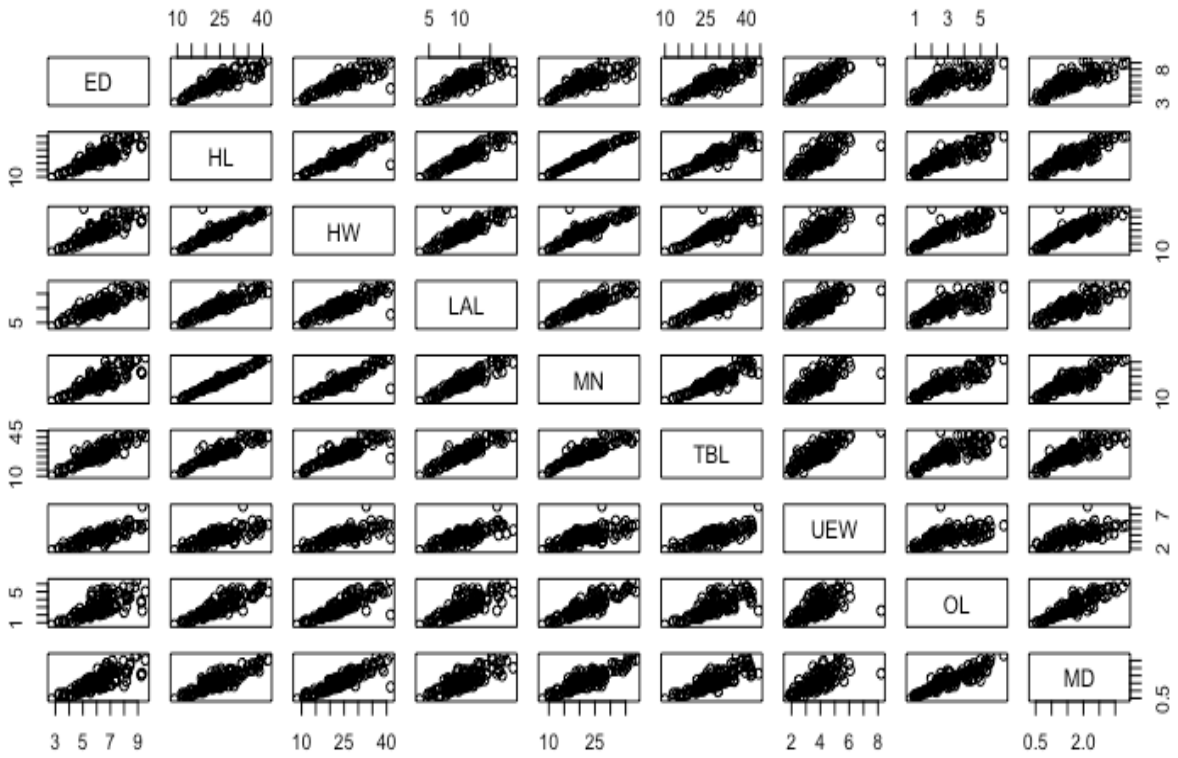


Figure 4.2: Scatter plot of the Frog Dataset.

Figure 4.2 is a matrix of scatter plots between variables in the Frog Dataset. It shows that there are very strong positive linear relationships between the ED variable with all of them. Also, there are strong positive linear relationships between the HL variable with all remained variables, but the relationship between HL and MN is a very strong positive relationship.

Also, the HW variable has strong linear relationships with other remaining variables, but the relationship between the HL and HW variables is the best. Besides the previous, there are strong positive linear relationships between the LAL variable with all remaining variables. The same for the MN variable. Also, there are very strong positive linear relationships between the TBL variable with all remaining variables, but the relationship between TBL and OL is the weakest. And the relationships between the UEW variable with all remaining are moderate linear relationships, and this applies to the variables OL and MD.

Univariate Tests of Normality:

The normality of a data can be tested by the Shapiro-Wilk test with a package shapiro.test in R. In this case, the Null Hypothesis: the variable is normally distributed, and the Alternative Hypothesis: the variable is not normally distributed. The p-value of the Shapiro-Wilk test of the four datasets was followed in Table 4.1:

TABLE 4.1 Result of the Normality Test for the Four Datasets.

Dataset	Variables	P-value of Shapiro-Wilk test
Depression	PUA FC MD FNA FD	0.0002511 2.382e-05 2.424e-10 1.464e-05 1.281e-05
Bankruptcy	CF/TD NI/TA CA/CL CA/NS	0.5431 5.302e-06 0.001916 0.3457
Frog	ED HL HW LAL MN TBL UEW OL MD	0.2739 0.00302 0.04535 0.03218 0.0009113 0.02366 0.001216 0.0001036 0.008917

As we notice, the Depression Dataset, all variables are abnormal. Concerning the Bankruptcy Dataset, just CF/TD and CA/NS variables are normal. Also, in Frog Dataset, the variable ED and the p-value of the remained variables are almost too close to 0.05.

Test the Equality of Covariance Matrix:

Box's M test measures the hypothesis of equal covariance between variables, and that is used to test the equality of the covariance matrix. The results showed a p-value ≅ 0, which means rejecting the null hypothesis that assumes equality. This result is the same in all datasets.

4.2.2 Applying Linear Discriminant Analysis

Tables below show linear discrimination coefficients for each dataset, which in turn measure the relative weight and importance of all variables in each data, such that the variable that possesses the absolute larger coefficient means that it has the heaviest of the variables and vice versa. This means that the groups are most affected by this variable.

Example 1:

TABLE 4.2 Standardized Canonical Discriminant Function Coefficients for the Depression Data.

Variable	PUA	FC	MD	FNA	FD
LD1 Coefficient	0.64634	0.34609	-0.1384	0.34501	0.68123

The coefficients in Table 4.2 indicate that the FD has the strongest discriminatory effect, followed by PUA, and MD is the least discriminating.

Example 2:

TABLE 4.3 Standardized Canonical Discriminant Function Coefficients for the Bankruptcy Data.

Variable	CF/TD	NI/TA	CA/CL	CA/NS
LD1 Coefficient	0.66124	0.39352	0.88721	-1.1785

The coefficients indicate that the first independent variable which is CA/NS has the strongest discriminatory effect, followed by the third independent variable. But the second variable is the

least discriminating.

Example 3:

TABLE 4.4 Standardized Canonical Discriminant Function Coefficients for the Frog Data.

Variable	ED	HL	HW	LAL	MN	TBL	UEW	OL	MD
LD1 Coefficient	0.021	0.23	0.004	0.012	-0.1	0.18	-0.52	-2.02	0.96

As we notice the OL variable has the strongest discriminatory effect, followed by the MD variable. However, HW is the least discriminating.

4.2.3 Applying Logistic Regression

The Pseudo $R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y}_i)^2}$, where N is the sample size, y is the dependent variable, $\bar{y}_i$ is the mean of y values, and $\hat{y}_i$ is the predicted value of y_i .

R^2 as a refinement from null to fitted model, as you notice, the denominator of the ratio is dealing with the sum of squared errors from the null model which predicts the dependent variable without the combination of independent variables, and it considers that every value of y is predicted to be $\bar{y}$. But numerator dealing with the sum of squared errors from the fitted model.

We seek to reduce the difference between real values and predictive values, which increases the value of the numerator in the R^2 ratio. Therefore, obtaining a higher value means that the estimate of parameters of the model is improved. Also, in simple linear regression, i.e. when you have only one independent variable, R^2 considered as a square of the correlation between predicted and the actual values, so it ranges from 0 to 1. Besides, this ratio can tell us how much $\hat{y}$ varies from y.

In linear regression, Pseudo R^2 is the coefficient of determination and its function is to summarize the variation in the dependent variable associated with independent variables. The greater its value indicates that the model interprets the greater variance. Its maximum value is 1.

But in the logistic regression where the dependent variable is categorical, it is not possible to calculate Pseudo R^2 at once and having the characteristics of Pseudo R in the linear regression. In the case of logistic regression, the estimates are calculated using the log likelihood method, and there are three ways:

1. Cox and Snell: Based on a comparison of the log likelihood of the model to the log likelihood of the basic model, and because of categorical dependent variables, the maximum value is less than 1.
2. Nagelkerke: It is a modified version of Cox and Snell to cover the full range from 0 to 1.
3. McFadden: A modified version also depends on the log-likelihood kernels for the intercept-only model and the full estimated model.

When comparing competing ways to the same data, the way which has the highest R^2 statistic is the best way.

TABLE 4.5 The Pseudo R^2 of the Datasets.

Datasets	McFadden	Cox and Snell	Nagelkerke
Depression	0.552190	0.504879	0.701204
Bankruptcy	0.567287	0.542570	0.725264
Frog	0.33505	0.371237	0.509015

Table 4.5 shows that for the Depression Dataset, 55.219% of the variance of explanatory variables was explained by the dependent variable according to McFadden value, and 50.4879% according to Cox and Snell value, also 70.1204% according to Nagelkerke value.

Also, it shows that for the Bankruptcy Dataset, 56.728% of the variance of explanatory variables was explained by the dependent variable according to McFadden value, and 54.257% according to Cox and Snell value, also 72.5264% according to Nagelkerke value.

TABLE 4.6 Results of Fitting the Logistic Regression Model of the Depression Data.

Variables	Estimated coefficients	Std. Error	Z value	Sig.	Exp(β)
Intercept	-15.2226	6.3321	-2.404	0.0162	
PUA	0.6757	0.8045	0.840	0.4009	1.9654
FC	1.1471	1.0622	1.080	0.2802	3.149
MD	2.9016	2.2733	1.276	0.2018	18.203
FNA	1.0316	0.7080	1.457	0.1451	2.80556
FD	2.0280	1.2023	1.687	0.0917	7.59888

Most of odds ratio are greater than 1, and its logit equation = - 15.2226 + 0.6757 PUA + 1.1471 FC + 2.9016 MD + 1.0316 FNA + 2.0280 FD.

Also, as the p-values of all variables are more than 0.05, then all of them are not significant variables in the logistic regression model.

TABLE 4.7 Results of Fitting the Logistic Regression Model of the Bankruptcy Data.

Variables	Estimated coefficients	Std. Error	Z value	Sig.	Exp(logit)
Intercept	-5.320	2.366	-2.248	0.02459	
CF/TD	7.138	6.002	1.189	0.23433	1259.0
NI/TA	-3.703	13.670	-0.271	0.78647	0.0246
CA/CL	3.415	1.204	2.837	0.00455	30.416
CA/NS	-2.968	3.065	-0.968	0.33286	0.051

Most of odds ratio are greater than 1, and its logit equation = - 5.320 + 7.138 CF/TD -3.703 NI/TA + 3.415 CA/CL - 2.968 CA/NS.

Also, as the p-values of CF/TD, NI/TA, CA/NS variables are more than 0.05, then all of them are not significant variables in the logistic regression model, but the p-values of CA/CL is less than 0.05, so CA/CL is a significant variable in the logistic regression model.

TABLE 4.8 Results of Fitting the Logistic Regression Model of the Frog Data.

Variables	Estimated coefficients	Std. Error	Z value	Sig.	Exp(logit)
Intercept	- 3.33426	1.47900	- 2.254	0.024171	
ED	0.09358	0.55725	0.168	0.866631	1.098
HL	0.39634	0.42067	0.942	0.346100	1.486
HW	- 0.04487	0.09621	- 0.466	0.640986	0.956
LAL	- 0.02424	0.30230	- 0.080	0.936099	0.976
MN	- 0.29124	0.38461	-0.757	0.448911	0.7473
TBL	0.38688	0.18036	2.145	0.031952	1.4723
UEW	- 0.49622	0.67691	- 0.733	0.463520	0.6088
OL	- 2.47920	0.65163	- 3.805	0.000142	0.082.
MD	0.73336	1.33118	0.551	0.581697	2.082

Its logit equation = - 3.33426 + 0.09358 ED + 0.39634 HL - 0.04487 HW - 0.02424 LAL - 0.29124 MN + 0.38688 TBL - 0.49622 UEW - 2.47920 OL + 0.73336 MD.

Also, as the p-values of ED, HL, HW, LAL, MN, UEW, and MD variables are more than 0.05, then all of them are not significant variables in the logistic regression model, but the p-values of TBL and OL are less than 0.05, so they are significant variables in the logistic regression model.

4.2.4 Classification Results of the Diabetes Dataset

Now we will study the performance of LR, and compare it with the performance of LDA, GMM, and Neural Networks (NN).

The Comparison determined according to the classification rates, which will be illustrated in tables for each dataset. Classification table contains n_{00} which is the number of observations which equal 0 and correctly predicted as 0, n_{01} is the number of observations which equal 0 and incorrectly predicted as 1, n_{10} is the number of observations which equal 1 and incorrectly predicted as 0, n_{11} is the number of observations which equal 1 and correctly predicted as 1.

The error rate is the ratio of the summation of the number of incorrectly predicted observations (n_{01}, n_{10}) to the total number of observations in the dataset (n_{00}+ n_{01} +n_{10} +n_{11}).

It's important to know that in comparing the error rates, we assume that the costs of misclassification are equal and have the same penalty, i.e. c_{01}=c_{10}, and c_{00}= c_{11}=0.

TABLE 4.9 Classification Table for the Two Groups.

Actual Group	Number of Observations	Predicted Group	
		0	1
0	n_0	n_{00}	n_{01}
1	n_1	n_{10}	n_{11}

And the Error Rate = $\frac{n_{01}+n_{10}}{n_{00}+n_{01}+n_{10}+n_{11}}$.

TABLE 4.10 Confusion Matrix of the Depression Data Classification Error Rates for the Four Methods.

Data Name	Actual		Predicted		Error Rate (with equal cost)
			0	1	
LR	0	15	12	3	0.1333333
	1	30	3	27	
LDA	0	15	12	3	0.1555556
	1	30	4	26	
GMM	0	15	14	1	0.2000000
	1	30	8	22	
Neural Networks (NN)	0	4	3	1	0.1818181
	1	7	1	6	

The Depression dataset has 45 observations. Here as we notice, for LR, there are 12 observations of the first group were classified correctly, and 27 observations from the second group were correctly classified also.

So LR succeeded in classifying 86.667% of observations.

Now, the LDA classified 12 observations from the first group and 26 from the second group correctly. That means that the LDA succeeded in classifying 84.45% of the observations.

Besides, the GMM classified 14 observations from the first group and 22 from the second group correctly. That means that the GMM succeeded in classifying 80% of the observations. Also, the NN method made for a sample of the Depression dataset with sample size = 11. It classified 3 observations from the first group and 6 from the second group correctly.

That means that Neural Networks (NN) misclassify 81.8181% of the observations correctly. For more accuracy, this step is repeated 15 times, then we take the average of those correct classification rate.

TABLE 4.11 The Average of Error Rates for Each Method for the Depression Data.

Method	LR	LDA	GMM	Neural Networks (NN)
Error Rate	0.14942	0.16110	0.22466	0.169697

As you can see, LR has the smallest error rate, it followed by the LDA, then the GMM has the biggest error rate followed by the NN method.

TABLE 4.12 Confusion Matrix of the Bankruptcy Data Classification Error Rates for the Four Methods.

Data Name	Actual		Predicted		Error Rate
			0	1	
LR	0	21	18	3	0.08696
	1	25	1	24	
LDA	0	21	18	3	0.08696
	1	25	1	24	
GMM	0	21	21	0	0.021734
	1	25	1	24	
Neural Networks (NN)	0	6	6	0	0.166667
	1	6	2	4	

The Bankruptcy dataset has 46 observations. Here as we noticed, for LR, there are 18 observations of the first group were classified correctly, and 24 observations from the second group were correctly classified. So LR succeeded in classifying 91.304% of observations correctly. The LDA has the same error rate.

Besides, the GMM classified 21 observations from the first group and 24 from the second group correctly. That means that the GMM succeeded to classify 97.8266% of the observations correctly.

Also, the NN method made for a sample of the Bankruptcy Dataset with sample size = 12.
It classified 3 observations from the first group and 6 from the second group correctly.
That means that the NN method succeeded in classifying 83.333% of the observations correctly.

For more accuracy, this step is repeated 15 times, then we take the average of those correct classification rate.

TABLE 4.13 The Average of Error Rates for Each Method for the Bankruptcy Data.

Method	LR	LDA	GMM	Neural Networks (NN)
Error Rate	0.1111	0.08696	0.222	0.58333

As you can see, the LDA method has the smallest error rate so it behaved very well and it is followed by the LR method, and the NN method has the biggest error rate.

TABLE 4.14 Confusion Matrix of the Frog Data Classification Error Rates for the Four Methods.

Data Name	Actual		Predicted		Error Rate
			0	1	
LR	0	8	7	1	0.1724
	1	11	4	7	
LDA	0	8	5	3	0.2069
	1	11	3	8	
GMM	0	8	2	6	0.2413
	1	11	1	10	
Neural Networks (NN)	0	3	2	1	0.166667
	1	3	0	3	

The Frog dataset has 132 observations. Here as we noticed, for the LR, there are 7 observations of the first group were classified correctly, and 7 observations from the second group were correctly classified. So LR succeeded in classifying 82.76% of observations correctly. Now, the LDA classified 5 observations from the first group and 8 from the second group correctly. That means that the LDA succeeded in classifying 79.31% of the observations.

Besides, the GMM method classified 2 observations from the first group and 10 from the second group correctly. That means that the GMM method succeeded to classify 75.9% of the observations correctly. Also, the NN method made for a sample of the Depression dataset with sample size = 6. It classified 2 observations from the first group and 3 from the second group correctly. That means that the NN method succeeded in classifying 83.333% of the observations

correctly. For more accuracy, this step is repeated 15 times, then we take the average of those correct classification rate.

TABLE 4.15 The Average of Error Rates for Each Method for the Frog Data.

Method	LR	LDA	GMM	Neural Networks (NN)
Error Rate	0.1458	0.1736	0.2152	0.1111

As you can see, the NN method has the smallest error rate. It followed by LR, the LDA method comes in the third place and GMM has the biggest error rate.

Also, the GMM method classified 21 observations from the first group and 3 from the second group correctly. That means that GMM succeeded in classifying 48% of the observations correctly.

Also, the NN method made for a sample of the Depression dataset with size 13. It classified zero observations from the first group and 10 from the second group correctly. That means that the NN method succeeded in classifying only 25% of the observations correctly.

For more accuracy, this step is repeated 15 times. Then the average of those correct classification rate is reported.

In general, we note that the LR method recorded the lowest error rate in the classification of all data in the event that assumptions of LDA are overturned, and therefore LR shows its superiority over the other three classification methods.

LR followed by the LDA method in all cases. In most cases, Neural Networks (NN)was ranked last and had the highest rate of error.

4.3 Study the Effect of Some Data Properties

In this chapter, we will examine the properties of the data y affect the efficiency of the performance of each method of classification.

For the Women's Empowerment Dataset, the variables were collected from the United Nations Development Programme's (UNDP) and the Central Intelligence Agency's (CIA) sites. The data included 6 variables and 177 observations. These variables were added to other variables in the paper "A Multivariate Statistical Analysis of Female Empowerment ", by the students: Adri Anne Demski and Janelle Jones.

Women's Empowerment Data was chosen because of the importance of the topic, the selected variables are of the global interest to members of the UN. It also shows the classification and rank of Palestine among the countries of the world. The most important that it serves our purpose and satisfies the conditions of the method of Linear Discriminant Analysis (LDA).

The covariance matrices of variables are equal, and the unit matrix is adapted for easy handling. The means of the variables are not equal (Demski, Jones, 2004).

While the requirements of the method of linear discrimination analysis are met, the extent of the factors of sample size and distance between means for both classes, classification matrices and the correlation between forecasters will be examined for each method of classification, with error rate as the valuation method.

4.3.1 Results for the Effect of Sample Size

TABLE 4.16: Results of error rates for the effect of sample size.

Sample size	LR	LDA	GMM	Neural Networks (NN)
28	0.0805	0.2281	0.9798	0.214
51	0.1194	0.2222	0.9762	0.70588
96	0.1481	0.1728	0.9753	0.8395
177	0.1016	0.2033	0.9830	0.2130

The error rate sometimes increases and sometimes decreases with increasing sample size, so there is no clear correlation between error rate and increasing of the sample size, we may get more clear results if sample size is too large

We also note that the LR method showed superiority over the rest of the roads, followed by LDA then Neural Networks (NN), but the error rate of GMM method is very high.

4.3.2 Results for the Effect of Existing Correlated Variables

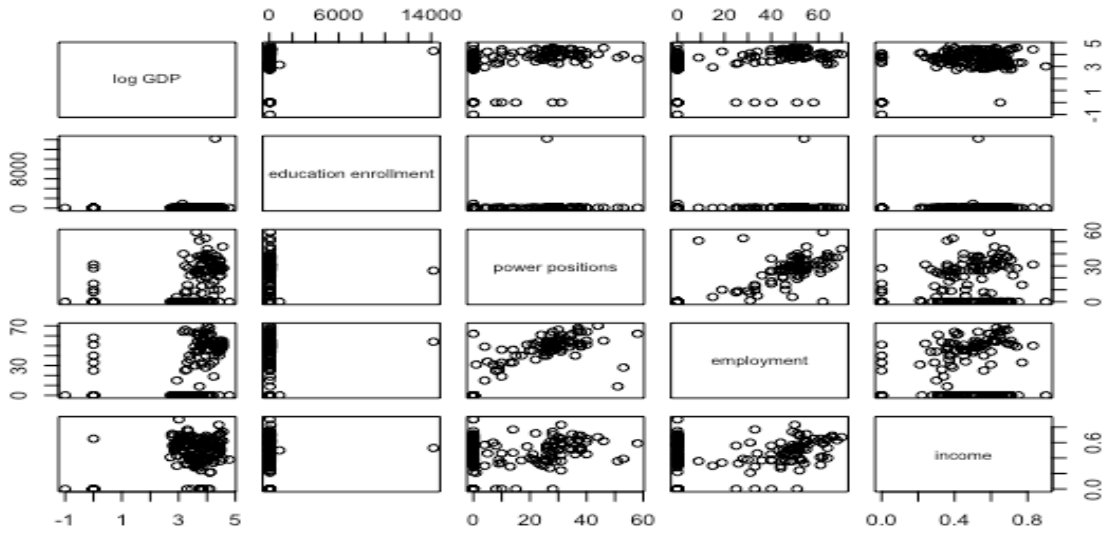


Figure 4.3: Scatter plot of Women Empowerment dataset.

As we see in Figure 4.3 there is a positive linear relationship between Employment with Power positions and Income with Employment variables. So we will calculate the error rate with and without correlated variables, then compare.

TABLE 4.17: Results of the error rate for the effect of existing of correlated data.

Classification method	With correlated variables	Without correlated variables
LR	0.1316	0.1129
LDA	0.2133	0.2090
GMM	0.9830	0.9152
Neural Networks (NN)	0.2930	0.27273

As we see that the error rate increases little when data has a correlated variable, which means that the existence of correlated variables increase the error rate, but not too much.

It is different when studying the effect of variables associated with each other on the error rate of the regression method and the rest of the other methods, the effect on the performance of the four classification methods was found to be negligible.

The rate of error was decreased by approximately 2% of LR if the associated variables were removed. In LDA, the error rate decreased by 0.77%, it also approximately decreased by 6%, but GMM's error rate decreased by 7%.

Chapter 5

Conclusion and Recommendation

5.1 Conclusion

In this study, the performance of Logistic Regression (LR) was compared with the performance of 3 other methods: Linear Discrimination Analysis (LDA), Gaussian Mixture Models (GMM), and Neural Networks (NN). We have relied on the comparison of the error rates when the costs of misclassifications are equal.

We used the same three real data to evaluate the performance of each of the four classification methods: Depression Data, Bankruptcy and Frogs Data sets.

We have used the Women's Empowerment data to examine the impact of certain data

characteristics on the performance of the four methods of classification.

The Depression data didn't meet the requirements of normality. LR showed superiority in its performance from the rest of the four classification methods, which recorded the lowest error rate, followed by LDA, GMM ranked in the final and preceded by the Neural Networks (NN).

As well as in Bankruptcy statements, GMM showed clear superiority over the other three classification methods followed by LDA and LR. The method of Neural Networks (NN) was obtained at the highest rate of error.

As for the Frog data, which followed the natural distribution, we are surprised that the method that recorded the lowest error rate is Neural Networks (NN), followed by LR with a slight difference, then by the LDA and GMM at last.

Women Empowerment data is used to match the conditions of the LDA, which is normality,

equality of the covariance matrix, and the difference in the means of variables to study the effect of the sample size, and the existence of the correlation between the variables on the performance of the four classification methods, then investigate whether the LR will remain superior to LDA or not.

However, using the Women Empowerment dataset, as we noted earlier matched the conditions of LDA which is normality and equal covariance matrix of, and the different means.

As LDA assumptions are met, we found that the was ranked second lowest error rate among the other three classification methods after LDA. In the case of the study of the effect of sample size on LR performance and the other methods of classification, there was no clear correlation between the error rate and increasing of the sample size.

Where LR showed the lowest error rate followed by LDA, then by Neural Networks (NN) and finally GMM, i.e. sample size didn't effect sufficiently on the performance of methods.

It is different when studying the effect of variables associated with each other on the error rate of the LR method and the rest of the other methods. Error rates have increased for LR, LDA, and Neural Networks (NN) slightly decreased for GMM.

5.2 Recommendations

Based on the inferences, we recommend that:

1. Before using one of the classification methods, it's important to check the data and explore the data graphically and numerically because we rarely find normal data.

2. As we note that the error rates for LR and LDA are very close to whether the data met the conditions of LDA or did not. Besides, LR exceeded LDA in the classification rates in general. Therefore, using LR in classifying data in case the data does not meet the conditions of LDA is recommended.

3. GMM can be affected by the correlation between independent variables. So it is preferable to get rid of one of them which is less important than the other correlated variables.

4. The thesis included the data of the number of observations less than 200, the effect of sample size on the performance of classification methods may be large if the size exceeds this number.

5. The thesis was limited to binary variables, so these results and recommendations may change of the dependent variable has more than two levels.

6. The cost of misclassification is the simple one, the results may change with a more complex cost function.

7. This study is limited in scope, so we need to increase the variety of examples to make more concrete recommendations.

Bibliography

[1] Agresti, A. (2018). *An introduction to categorical data analysis*. Wiley.

[2] Anifowose, F. A. (2012). A comparative study of gaussian mixture model and radial basis function for voice recognition. *arXiv preprint arXiv:1211.2556*.

[3] Balakrishnama, S., & Ganapathiraju, A. (1998). Linear discriminant analysis-a brief tutorial. *Institute for Signal and information Processing*, *18*, 1-8.

[4] Bolin, J. E., & Finch, H. (2014). Supervised classification in the presence of misclassified training data: a Monte Carlo simulation study in the three group case. *Frontiers in psychology*, *5*, 118.

[5] Cabrera, A. F. (1994). Logistic regression analysis in higher education: An applied perspective. *Higher education: Handbook of theory and research*, *10*, 225-256.

[6] Cao, M. (2010). Practice on classification using gaussian mixture model. *Course Project Report for COMP-135, Fall*.

[7] Demir, E. (2014). A decision support tool for predicting patients at risk of readmission: A comparison of classification trees, logistic regression, generalized additive models, and multivariate adaptive regression splines. *Decision Sciences*, *45*(5), 849-880.

[8] El-Jazzar, M., & El-Habil, A. (2014). A Comparative Study between Linear Discriminant Analysis and Multinomial Logistic Regression.

[9] Guar, P. (2012). Neural Network in data mining. *International Jornal of Electrinics and Cumputer Science Enginering (IJECSE), I(03*), 1994-1453.

[10] Hamid, H. (2010). A new approach for classifying large number of mixed variables. In *International Conference on Computer and Applied Mathematics* (pp. 156-161).

[11] Haykin, S. S., Haykin, S. S., Haykin, S. S., Elektroingenieur, K., & Haykin, S. S. (2009). *Neural networks and learning machines* (Vol. 3). Upper Saddle River: Pearson education.

[12] Holden, J. E., Finch, W. H., & Kelley, K. (2011). A comparison of two-group classification methods. *Educational and psychological measurement*, *71*(5), 870-901.

[13] Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression* (Vol. 398). John Wiley & Sons.

[14] Hossain, M., Wright, S., & Petersen, L. A. (2002). Comparing performance of multinomial logistic regression and discriminant analysis for monitoring access to care for acute myocardial infarction. *Journal of Clinical Epidemiology*, *55*(4), 400-406.

[15] Huang, Yonghong, et al. "A Gaussian mixture model based classification scheme for myoelectric control of powered upper limb prostheses." *IEEE Transactions on Biomedical Engineering* 52.11 (2005): 1801-1811.

[16] Huang, Y., Englehart, K. B., Hudgins, B., & Chan, A. D. (2005). A Gaussian mixture model based classification scheme for myoelectric control of powered upper limb prostheses. *IEEE Transactions on Biomedical Engineering*, *52*(11), 1801-1811.

[17] Johnson, R. A., & Wichern, D. W. (1999). *Applied multivariate statistical analysis* (Vlo. 5, No. 8). Upper Saddle River, NJ: Prentice hall.

[18] kiang, M. Y. (2003). A comparative assessment of classification methods. Decision support system, 35(4), 441-454.

[19] Kiveu, C. W. (2015). *Appraisal of Credit Applicant Using Logistic and Linear Discriminant Models With Principal Component Analysis* (Doctoral dissertation).

[20] Kriesel, D. (2011). A Brief Introduction to Neural Networks. dkriesel. com.

[21] Lu, D., & Weng, Q. (2007). A survey of image classification methods and techniques for improving classification performance. *International journal of Remote sensing*, *28*(5), 823-870.

[22] Liong, C. Y., & Foo, S. F. (2013, April). Comparison of linear discriminant analysis and logistic regression for data classification. In *AIP Conference Proceedings* (Vol. 1522, No. 1, pp. 1159-1165). AIP.

[23] Maroco, J., Silva, D., Rodrigues, A., Guerreiro, M., Santana, I., & de Mendonça, A. (2011). Data mining methods in the prediction of Dementia: A real-data comparison of the accuracy, sensitivity and specificity of linear discriminant analysis, logistic regression, neural networks, support vector machines, classification trees and random forests. *BMC research notes*, *4*(1), 299.

[24] Nikam, S. S. (2015). A comparative study of classification techniques in data mining algorithms. *Oriental journal of computer science & technology*, *8*(1), 13-19.

[25] Oliveira-Brochado, A., & Martins, F. V. (2005). *Assessing the number of components in mixture models: a review* (No. 194). Universidade do Porto, Faculdade de Economia do Porto.

[26] Omurlu, I. K., Ture, M., Unubol, M., Katranci, M., & Guney, E. (2014). Comparing performances of logistic regression, classification & regression trees and artificial neural networks for predicting albuminuria in type 2 diabetes mellitus. *Int J Sci Basic Appl Res*, *16*(1), 173-87.

[27] Pohar, M., Blas, M., & Turk, S. (2004). Comparison of logistic regression and linear discriminant analysis: a simulation study. *Metodoloski zvezki*, *1*(1), 143.

[28] Poulsen, J., & French, A. (2008). Discriminant function analysis. *Retrieved from*.

[29] Povinelli, R. J., Johnson, M. T., Lindgren, A. C., & Ye, J. (2004). Time series classification using Gaussian mixture models of reconstructed phase spaces. *IEEE Transactions on Knowledge and Data Engineering*, *16*(6), 779-783.

[30] Press, S. J., & Wilson, S. (1978). Choosing between logistic regression and discriminant analysis. *Journal of the American Statistical Association*, *73*(364), 699-705.

[31] Reimann, Clemens, et al. "Statistical data analysis explained." *Applied environmental statistics with R. England: Wiley & Sons Ltd* (2008).

[32] Rencher, A. C. (2003). *Methods of multivariate analysis* (Vol. 492). John Wiley & Sons.

[33] Russell, S. J., & Norvig, P. (2016). *Artificial intelligence: a modern approach*. Malaysia; Pearson Education Limited,.

[34] Tansey, R., White, M., Long, R. G., & Smith, M. (1996). A comparison of loglinear modeling and logistic regression in management research. *Journal of management*, *22*(2), 339-358.

[35] Timm, N. H. (2002). *Applied Multivariate Analysis: Springer Texts in Statistics*. Springer-Verlag New York Incorporated.

[36] Titterington, D. M., Smith, A. F., & Makov, U. E. (1985). *Statistical analysis of finite mixture distributions*. Wiley,.

[37] Tiwary, D. K. (2014). A Comparative study of classification algorithms for credit card approval using weka. *GALAXY International Interdisciplinary Research Journal, GIIRJ*, *2*(3), 165-174.

[38] Wang, Y. (2008). Comparing linear discriminant analysis with classification trees using forest landowner survey data as a case study with considerations for optimal biorefinery siting.

[39] Wehrens, R. (2011). *Chemometrics with R: multivariate data analysis in the natural sciences and life sciences*. Springer Science & Business Media.

[40] Zhang, G. P. (2000). Neural networks for classification: a survey. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, *30*(4), 451-462.

[41] Zhang, H., & Huang, Y. (2015). Finite mixture models and their applications: A review. *Austin Biometrics and Biostatistics*, *2*(1), 1-6.

الملخص

يلعب تصنيف الملاحظات دورًا مهمًا في علم الإحصاء وجميع المجالات الأخرى. في هذه الأطروحة، ندرس الانحدار اللوجستي كوسيلة للتصنيف ومقارنة أدائه بأداء تحليل التمايز الخطي، ونموذج خليط غاوس، والشبكات العصبية، تتم مقارنة الأداء من خلال جدول التصنيف الخاطئ ومعدل الخطأ لكل طريقة.

علاوة على ذلك، تم دراسة تأثير حجم العينة ووجود الارتباط. بشكل عام، أظهرت النتائج أنه عندما يتم استيفاء افتراضات تحليل التمييز الخطي، فإن أداء طريقة تحليل التمايز الخطي هو الأفضل. إن لم تستوف الشروط، فإن أداء طريقة الانحدار اللوجستي يتفوق على أداء أساليب التصنيف الأخرى.