Volume 8, Issue 12, December 2023 ISSN (online): 2456-0006

International Journal of Science Technology Management and Research

Available online at: www.ijstmr.com

# Feature Extraction For Dimension Reduction A Better Approach

Prabodhini Bholankar
M Tech 4th Sem
Computer Science and Engineering Department
LNCT(Bhopal) Indore Campus
Indore M.P. India

Mr Nilesh Avinash Joshi
Assistant professor
Computer Science and Engineering Department
LNCT(Bhopal) Indore Campus
Indore M.P. India

***Abstract: Dimensionality reduction is a technique used to reduce the number of features in a dataset while retaining as much of the important information as possible. In other words, it is a process of transforming high-dimensional data into a lower-dimensional space that still preserves the essence of the original data. Feature selection involves selecting a subset of the original features that are most relevant to the problem at hand. The goal is to reduce the dimensionality of the dataset while retaining the most important features. There are several methods for feature selection, including filter methods, wrapper methods, and embedded methods. Filter methods rank the features based on their relevance to the target variable, wrapper methods use the model performance as the criteria for selecting features, and embedded methods combine feature selection with the model training process. Linear Discriminant Analysis, or LDA, is a machine learning algorithm that is used to find the Linear Discriminant function that best classifies or discriminates or separates two classes of data points. LDA is a supervised learning algorithm, which means that it requires a labelled training set of data points in order to learn the Linear Discriminant function.***

***Keywords: Linear, Discriminant, Analysis, Dimension, Reduction, Precision, Recall***

## I. Introduction

Machine Learning: Machine learning is nothing but a field of study which allows computers to “learn” like humans without any need of explicit programming. Predictive modeling is a probabilistic process that allows us to forecast outcomes, on the basis of some predictors. These predictors are basically features that come into play when deciding the final result, i.e. the outcome of the model.

Dimensionality reduction is the process of reducing the number of features (or dimensions) in a dataset while retaining as much information as possible. This can be done for a variety of reasons, such as to reduce the complexity of a model, to improve the performance of a learning algorithm, or to make it easier to visualize the data. There are several techniques for dimensionality reduction, including principal component analysis (PCA), singular value decomposition (SVD), and linear discriminant analysis (LDA). Each technique uses a different method to project the data onto a lower-dimensional space while preserving important information.

Why is Dimensionality Reduction important in Machine Learning and Predictive Modeling?

An intuitive example of dimensionality reduction can be discussed through a simple e-mail classification problem, where we need to classify whether the e-mail is spam or not. This can involve a large number of features, such as whether or not the e-mail has a generic title, the content of the e-mail, whether the e-mail uses a template, etc. However, some of these features may overlap. In another condition, a classification problem that relies on both humidity and rainfall can be collapsed into just one underlying feature, since both of the aforementioned are correlated to a high degree. Hence, we can reduce the number of features in such problems. A 3-D classification problem can be hard to visualize, whereas a 2-D one can be mapped to a simple 2-dimensional space, and a 1-D problem to a simple line. The below figure illustrates this concept, where a 3-D feature space is split into two 2-D feature spaces, and later, if found to be correlated, the number of features can be reduced even further

## II. Linear Discriminant Analysis

Linear Discriminant Analysis, or LDA, is a machine learning algorithm that is used to find the Linear Discriminant function that best classifies or discriminates or separates two classes of data points. LDA is a supervised learning algorithm, which means that it requires a labelled training set of data points in order to learn the Linear Discriminant function. Once the Linear Discriminant function has been learned, it can then be used to predict the class label of new data points. LDA is similar to PCA (principal component analysis) in the sense that LDA reduces the dimensions. However, the main purpose of LDA is to find the line (or plane) that best separates data points belonging to different classes. The key idea behind LDA is that the decision boundary should be chosen such that it maximizes the distance between the means of the two classes while simultaneously minimizing the variance within each classes data or within-class scatter. Whenever there is a requirement to separate two or more classes having multiple features efficiently, the Linear Discriminant

Analysis model is considered the most common technique to solve such classification problems. For e.g., if we have two classes with multiple features and need to separate them efficiently. When we classify them using a single feature, then it may show overlapping.
This can be used to project the features of higher dimensional space into lower-dimensional space in order to reduce resources and dimensional costs. In this topic, "Linear Discriminant Analysis (LDA) in machine learning", we will discuss the LDA algorithm for classification predictive modeling problems, limitation of logistic regression, representation of linear Discriminant analysis model, how to make a prediction using LDA, how to prepare data for LDA, extensions to LDA and much more. So, let's start with a quick introduction to Linear Discriminant Analysis (LDA) in machine learning.
Linear Discriminant Analysis uses both the axes (X and Y) to create a new axis and projects data onto a new axis in a way to maximize the separation of the two categories and hence, reducing the 2D graph into a 1D graph.
Let's assume we have to classify two different classes having two sets of data points in a 2-dimensional plane as shown below image:

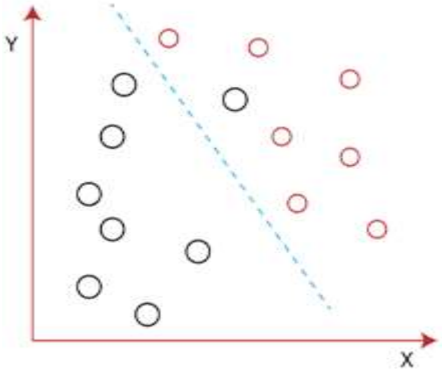

Fig 1 Two-dimensional plan separating data points

it is impossible to draw a straight line in a 2-d plane that can separate these data points efficiently but using linear Discriminant analysis; we can dimensionally reduce the 2-D plane into the 1-D plane. Using this technique, we can also maximize the separability between multiple classes

## How to Prepare Data for LDA

some suggestions that one should always consider while preparing the data to build the LDA model:

- **Classification Problems:** LDA is mainly applied for classification problems to classify the categorical output variable. It is suitable for both binary and multi-class classification problems.
- **Gaussian Distribution:** The standard LDA model applies the Gaussian Distribution of the input variables. One should review the univariate distribution of each attribute and transform them into more Gaussian-looking distributions. For e.g., use log and root for exponential distributions and Box-Cox for skewed distributions.
- **Remove Outliers:** It is good to firstly remove the outliers from your data because these outliers can skew the basic statistics used to separate classes in LDA, such as the mean and the standard deviation.
- **Same Variance:** As LDA always assumes that all the input variables have the same variance, hence it is always a better way to firstly standardize the data before implementing an LDA model. By this, the Mean will be 0, and it will have a standard deviation of 1.

## III. Literature Survey

In 2019 Peter Boedeker and Nathan T. Kearns proposed "Linear Discriminant Analysis for Prediction of Group Membership: A User-Friendly Primer" In psychology, researchers are often interested in the predictive classification of individuals. Various models exist for such a purpose, but which model is considered a best practice is conditional on attributes of the data. Under certain conditions, linear discriminant analysis (LDA) has been shown to perform better than other predictive methods, such as logistic regression, multinomial logistic regression, random forests, support-vector machines, and the K-nearest neighbor algorithm. The purpose of this Tutorial is to provide researchers who already have a basic level of statistical training with a general overview of LDA and an example of its implementation and interpretation. This Tutorial is meant to serve as a practical and applied overview of LDA for prediction of group membership[1].

In 2020 Ivan Rodrigues Alitta Parayil proposed "Use of Linear Discriminant Analysis (LDA), K Nearest Neighbours (KNN), Decision Tree (CART), Random Forest (RF), Gaussian Naive Bayes (NB), Support Vector Machines (SVM) to Predict Admission for Post Graduation Courses". Today's world has become really competitive when it comes to higher education. There are many options that bewilder students in the selection of their university. There are many consultancy facilities that help students in choosing the optimal university according to their needs. Many algorithms can be used to build a recommendation system . They are attempting to compare various machine learning algorithms that cater to the needs of building a recommendation system that will give the probability of admission of a student applying for Masters abroad[2].

In 2021 Yanni Li and Bing Liu proposed "3E-LDA: Three Enhancements to Linear Discriminant Analysis". Linear discriminant analysis (LDA) is one of the important techniques for dimensionality reduction, machine learning, and pattern recognition. However, in many applications, applying the classical LDA often faces the following problems: (1) sensitivity to outliers, (2) absence of local geometric information, and (3) small sample size or matrix singularity that can result in weak robustness and efficiency. Although several researchers have attempted to address one or more of the problems, little work has been done to address all of them together to produce a more effective and efficient LDA algorithm[3].

In 2021 Abbas F. H. Alharan and Zahraa M. Algelal proposed “Improving Classification Performance for Diabetes with Linear Discriminant Analysis and Genetic Algorithm”. In the modern-day, Diabetic disease is one of the most chronic and appalling diseases humanity faces. There are 463 million people had Diabetes worldwide, and it caused approximately 4.2 million deaths, according to the International Diabetes Federation (IDF) Diabetes Atlas Ninth edition 2019. Therefore diabetic patients need state-of-the-art healthcare against such diseases and propose early prediction to help decrease the risks related to such diseases. In this context, this research, a diabetes diagnosis system, has proposed to analyze two different diabetes datasets, namely PIMA Indian Diabetes and data[4].

In 2021 Vijay, 2Dr. Pushpneel Verma proposed “Linear Discriminant Analysis for Hate Speech Text Classification”. Social media has enabled the people to share their ideas widely online. Social media has many advantages. It provides a platform to people to express their talent. It provides a way to communicate with large number of people. Many people use social media to grow their network and strengthen their business. As the numbers of users are increasing on social media, the problem of hate speech is also increasing on social media. Hate speech on social media can provoke violence. There are many supervised machine learning based algorithms which can be used to detect hate speech on social media[5].

In 2022 Marion Olubunmi Adebiyi Micheal Olaolu Arowolo proposed “A Linear Discriminant Analysis and Classification Model for Breast Cancer Diagnosis”. The use of machine-learning methods will allow for the classification and prediction of cancer as either benign or malignant. This investigation applies the machine learning algorithms of random forest (RF) and the support vector machine (SVM) with the feature extraction method of linear discriminant analysis (LDA) to the Wisconsin Breast Cancer Dataset. The SVM with LDA and RF with LDA yielded accuracy results of 96.4% and 95.6% respectively. Evidence from this study shows that better prediction is crucial and can benefit from machine learning methods[6].

In 2022 Chunyan Wang and Wenjie Wang proposed “Regularized Linear Discriminant Analysis Via A New difference-of-convex algorithm with extrapolation”. They transformed the classical linear discriminant analysis (LDA) into a smooth difference-of-convex optimization problem. Then, a new difference-of-convex algorithm with extrapolation is introduced and the convergence of the algorithm is established. Finally, for face recognition problem, the proposed algorithm achieves better classification performance compared with several current algorithms in the literature. They proposed a new RLDA is proposed. A new DC algorithm with extrapolation is introduced for smooth DC problem and the convergence of this algorithm is given. Numerical results show that the proposed algorithm achieves better classification performance compared with current algorithms for face recognition. They may consider several more practical applications of RLDA in optimal control and so on[6].

In 2022 Juhaina , Terrance Frederick Fernandez proposed “Improved Accuracy in Stock Price Prediction System Using A Novel Decision Tree Algorithm Compared to Linear Discriminant Analysis (LDA) Algorithm”. The Main target of proposed work is comparative study of Novel Decision Tree Algorithm and Linear Discriminant Analysis (LDA) Algorithm for optimizing Stock price prediction to improve the Accuracy of Stock Exchange. Novel Decision Tree Algorithm (N=10) and Linear Discriminant Analysis Algorithm (N=10) are simulated by varying the Novel Decision Tree parameter and Linear Discriminant Analysis parameter to optimize the pH. Sample size is calculated using Gpower 80% for two groups and there are 20 samples used in this work[7].

## IV. WORKING LINEAR DISCRIMINANT ANALYSIS (LDA)

Originally developed in 1936 by R.A. Fisher, Discriminant Analysis is a classic method of classification that has stood the test of time. Discriminant analysis often produces models whose accuracy approaches (and occasionally exceeds) more complex modern methods. Discriminant analysis can be used only for classification (i.e., with a categorical target variable), not for regression. The target variable may have two or more categories. To explain discriminant analysis, let's consider a classification involving two target categories and two predictor variables.

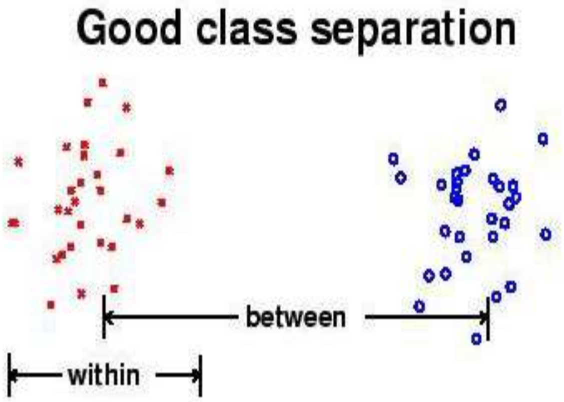

Fig 2 Separation within and between classes

A transformation function is found that maximizes the ratio of between-class variance to within-class variance as illustrated by this figure produced by Ludwig Schwardt and Johan du Preez transformation seeks to rotate the axes so that when the categories are projected on the new axes, the differences between the groups are maximized. The figure 4.2 (also by Schwardt and du Preez) shows two rotates axes. Projection to the lower right axis achieves the maximum separation between the categories; projection to the lower left axis yields the worst separation.

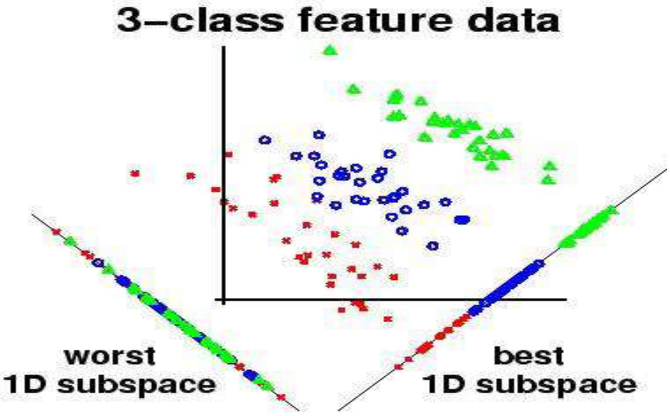

Fig 3 Worst 1D case and best 1D case

## V. Result Analysis

*A Show the scatter plot between two features*

This screen show scatter plot between two features and target class is denoted by 1 and 2 class 1 is denoted by blue colour and class2 denoted by red colour

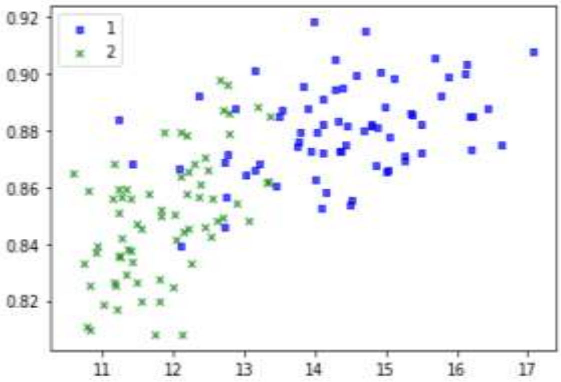

Fig 4 Graph shows linear relationship

*B Reducing dimension using LDA to plot the objects*

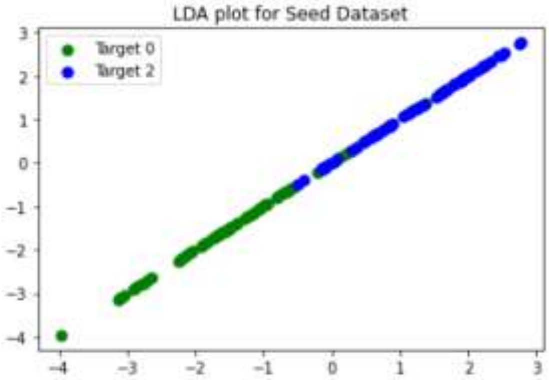

Figure 5 Reducing dimension using LDA

## Conclusion

Dimensionality reduction is the process of reducing the number of features (or dimensions) in a dataset while retaining as much information as possible. This can be done for a variety of reasons, such as to reduce the complexity of a model, to improve the performance of a learning algorithm, or to make it easier to visualize the data. There are several techniques for dimensionality reduction, including principal component analysis (PCA), singular value decomposition (SVD), and linear discriminant analysis (LDA). Each technique uses a different method to project the data onto a lower-dimensional space while preserving important information. When the classes with response variable are well-separated, the parameter estimates for the logistic regression model are surprisingly unstable.

## REFERENCE

1. Peter Boedeker "Linear Discriminant Analysis forPrediction of Group Membership:A User-Friendly PrimerAdvances in Methods and Practices in Psychological Science 2019, Vol. 2(3) 250–263 DOI: 10.1177/2515245919849378.
2. Ivan Rodrigues Use of Linear Discriminant Analysis (LDA), K Nearest Neighbours (KNN), Decision Tree (CART), Random Forest (RF), Gaussian Naive Bayes (NB), Support Vector Machines (SVM) to Predict Admission for Post Graduation Courses Electronic copy. Don Bosco Institute of Technology Mumbai 400070, India ivanrods19@gmail.com 2-19ailable at: https://ssrn.com/abstract=3683065.
3. Yanni Li, Xidian University 3E-LDA: Three Enhancements to Linear Discriminant Analysis Yanni Li, Bing Liu, Yongbo Yu, Hui Li, Jiacan Sun, and Jiangtao Cui. 2021. 3E-LDA: Three Enhancements to Linear Discriminant Analysis. ACM Trans. Knowl. Discov. Data 15, 4, Article 57 (March 2021), 20 pages.
4. Abbas F. H. Alharan Improving Classification Performance for Diabeteswith Linear Discriminant Analysis and GeneticmAlgorithm 2021 Palestinian International Conference on Information and Communication Technology (PICICT).
5. Vijay, Dr. Pushpneel Verma Linear Discriminant Analysis for Hate Speech Text Classification International Journal of Engineering Development and Research (www.ijedr.org) Year 2021, Volume 9, Issue 2.
6. Marion Olubunmi Adebiyi A Linear Discriminant Analysis and Classification Model forBreast Cancer Diagnosis A Linear Discriminant Analysis and Classification Model for Breast Cancer Diagnosis. Appl. Sci. 2022, 12,11455. https://doi.org/10.3390/ app122211455.
7. Chunyan Wang Regularized Linear Discriminant Analysis Via A Newdifference-of-convex algorithm with extrapolation October 4th, 2022 DOI: https://doi.org/10.21203/rs.3.rs-2112437/v1
8. Juhaina1, Terrance Frederick Fernandez2 *Improved Accuracy in Stock Price Prediction System Using a Novel Decision Tree Algorithm Compared to Linear Discriminant Analysis (LDA) Algorithm* Department of Computer Science and Engineering, Saveetha School of Engineering, Eur. Chem. Bull. 2023, 12 (S1), 4719– 4726