A Survey of Linear Discriminant Analysis with Applications to Star and Chest X-ray Data

Aatmun Baxi, Andy Chen, Johnny Mo, Abhi Vemulapati

***Abstract*— Linear Discriminant Analysis (LDA) is a supervised machine learning technique used for classification and dimensionality reduction. The objective is to create linear discriminant functions that maximize the ratio of between-class variance to within-class variance. In this paper, we outline the LDA approach and demonstrate its application on synthetic data, star data for classification and data visualization. Additionally, we consider the robustness of LDA to highly variant data with an application to chest X-ray images.**

I. INTRODUCTION

Supervised classification of datasets has proven to be one of the most versatile types of problems encountered in the study of machine learning. While the most prominent solutions to supervised classification, such as logistic regression, can tackle a diverse array of problems, other methods can more accurately classify datasets that satisfy certain assumptions. The method of linear discriminant analysis (LDA) is one such example [1]. LDA is used in a diverse array of problems, such as face recognition [2] and biomedical imaging [3]. Most notably, economist Edward Altman used LDA to create the Altman Z-score, a fairly accurate predictor of corporate defaults [4].

LDA provides a fix to the most significant shortcoming of logistic regression, which is its failure to find stable solutions to binary classification problems when classes are well-separated [5]. As a classifier, LDA assumes that input features are normally distributed in order to find a closed-form solution to the optimal separating hyperplane between classes. As a dimensionality reduction technique, LDA uses matrix diagonalization to detect the features of an input dataset which minimize the covariance among features within classes and maximize the covariance among features between classes.

The first iteration of LDA was introduced by statistician Ronald Fisher [6], who first proposed the technique when classifying two species of iris flowers. The later development of the multiclass LDA technique for classification and dimensionality reduction was done by C. R. Rao [7].

A. Notation and Definitions

Let N be the number of observations in a dataset. Let C be the number of classes (or labels) within the dataset and D be the number of features of each observation. We denote each of the N observations and their corresponding classes by $\mathbf{x}_i \in \mathbb{R}^D$ and $t_i \in \{1, \dots, C\}$ respectively for $i = 1, \dots, N$.

Let

$$X_k = \{\mathbf{x}_i \mid t_i = k\}$$

denote the set of *observations* within the kth class, and let

$$\mathcal{C}_k = \{i \mid t_i = k\}$$

denote the set of *indexes* of the observations belonging to the kth class. Let

$$N_k = |\mathcal{C}_k|$$

denote the *number* of observations belonging to the kth class. Additionally, let

$$\mathbf{m}_k = \frac{1}{N_k} \sum_{i \in \mathcal{C}_k} \mathbf{x}_i$$

denote the kth *(sample) class mean*, and let

$$\mathbf{m} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{x}_i$$

denote the *overall (sample) mean* of the dataset. Let

$$\mathbf{S}_k = \sum_{i \in \mathcal{C}_k} (\mathbf{x}_i - \mathbf{m}_k)(\mathbf{x}_i - \mathbf{m}_k)^\top$$

denote the *within-class scatter matrix* of the kth class.

B. Organization

We begin this paper by developing the mathematical foundations behind the various types of LDA. We then conduct and discuss several experiments that make use of these variations of the LDA technique, such as binary classification, dimensionality reduction, and multiclass classification.

II. MATHEMATICAL FOUNDATIONS

In this section, we present the assumptions and motivation for LDA as well as its mathematical formulation. We then show how it can be solved in the binary classification setting and subsequently the dimensionality reduction setting with the help of Fisher's criterion. We conclude with LDA as a multiclass classifier by taking the Bayesian approach.

A. Assumptions

We assume that the data within each class satisfies the following:

1) **Normality**: Each observation in X_k is sampled from a multivariate normal distribution $\mathcal{N}(\mathbf{x} \mid \mathbf{m}_k, \mathbf{\Sigma}_k)$, where $\mathbf{\Sigma}_k$ is the *within-class covariance matrix* of the kth class.
2) **Homogeneity**: The within-class covariance matrices $\mathbf{\Sigma}_k$ are the same for all $k \in \{1, \dots, C\}$. Let us denote this common covariance matrix as $\mathbf{\Sigma}$.

Note that $\mathbf{S}_k / N_k$ approximates $\mathbf{\Sigma}$ for sufficiently large N_k, as $\mathbf{S}_k$ is $N_k - 1$ times the unbiased estimate for $\mathbf{\Sigma}_k$ and we assume that $\mathbf{\Sigma}_k = \mathbf{\Sigma}$.

B. Binary Classification

First, we consider the case of binary classification. We aim to use dimensionality reduction to simplify the problem and allow for an easier time classifying the data. To accomplish this, we project the data down to one dimension by taking the inner product of the data with a projection vector $\mathbf{w}$, yielding a new set of data $\{y_i\}_{i=1}^N$, where $y_i = \mathbf{w}^\top \mathbf{x}_i \in \mathbb{R}$.

We would like to project the data in such a way so as to maximize class separability. To achieve this, we maximize the distance between the means of the projected classes. That is, if $\mathbf{m}_1$ and $\mathbf{m}_2$ are the means of each class before projection, then

$$m_1 = \mathbf{w}^\top \mathbf{m}_1$$
$$m_2 = \mathbf{w}^\top \mathbf{m}_2$$

are the means of the projected classes, and we seek to maximize

$$m_1 - m_2 = \mathbf{w}^\top(\mathbf{m}_1 - \mathbf{m}_2).$$

Since scaling the projection vector $\mathbf{w}$ does not change the shape of the transformed data, we are only seeking an optimal direction for $\mathbf{w}$. Thus, we restrict $\mathbf{w}$ to unit length, $\|\mathbf{w}\| = 1$. This restriction on the magnitude of $\mathbf{w}$ also allows for a solution to the maximization problem, as otherwise, $m_1 - m_2$ can be made arbitrarily large.

Additionally, we want to minimize the overlap between the two projected classes for less ambiguity when classifying the observations. Thus, we would like the projected data within each class to be clustered tightly together. Fisher's solution is to maximize the ratio between the separation of class means and the separation of data points within each class, or in other words, the ratio of between-class variance to within-class variance.

The between-class variance is given by $(m_1 - m_2)^2$, while the within-class variance is the sum of squares within each class, i.e., $s_1^2 + s_2^2$, where

$$s_k^2 = \sum_{i \in \mathcal{C}_k} (y_i - m_k)^2.$$

Thus, Fisher's criterion is

$$J(\mathbf{w}) = \frac{(m_1 - m_2)^2}{{s_1}^2 + {s_2}^2}.$$

Rewritten in terms of $\mathbf{w}$, this is equivalent to

$$J(\mathbf{w}) = \frac{\mathbf{w}^\top \mathbf{S}_b \mathbf{w}}{\mathbf{w}^\top \mathbf{S}_w \mathbf{w}}$$

where

$$\mathbf{S}_b = (\mathbf{m}_1 - \mathbf{m}_2)(\mathbf{m}_1 - \mathbf{m}_2)^\top$$

is the *between-class scatter matrix* and

$$\mathbf{S}_w = \mathbf{S}_1 + \mathbf{S}_2$$

is the total *within-class scatter matrix*. Note that since $\mathbf{S}_1 \approx N_1 \mathbf{\Sigma}$ and $\mathbf{S}_2 \approx N_2 \mathbf{\Sigma}$, we have that $\mathbf{S}_w \approx N \mathbf{\Sigma}$.

To maximize $J(\mathbf{w})$, we set the gradient of it with respect to $\mathbf{w}$ equal to $\mathbf{0}$ and obtain

$$(\mathbf{w}^\top \mathbf{S}_b \mathbf{w}) \mathbf{S}_w \mathbf{w} = (\mathbf{w}^\top \mathbf{S}_w \mathbf{w}) \mathbf{S}_b \mathbf{w}.$$

Noting that $\mathbf{w}^\top \mathbf{S}_b \mathbf{w}$ and $\mathbf{w}^\top \mathbf{S}_w \mathbf{w}$ are scalars and

$$\mathbf{S}_b \mathbf{w} = (\mathbf{m}_1 - \mathbf{m}_2)(\mathbf{m}_1 - \mathbf{m}_2)^\top \mathbf{w} \propto \mathbf{m}_1 - \mathbf{m}_2$$

since $(\mathbf{m}_1 - \mathbf{m}_2)^\top \mathbf{w}$ is a scalar, it follows that

$$\mathbf{w} \propto \mathbf{S}_w^{-1}(\mathbf{m}_1 - \mathbf{m}_2).$$

Due to our assumptions of normality and homogeneity, the within-class variances of our projected data must also be homogeneous for both classes. This stems from the fact that the probability distribution of a transformed observation from the kth class is

$$p(y_i) = p(\mathbf{w}^\top \mathbf{x}_i) = \mathcal{N}(\mathbf{x}_i \,|\, \mathbf{m}_k, \mathbf{w}^\top \mathbf{\Sigma}_k \mathbf{w})$$

by [5], but as $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2 = \mathbf{\Sigma}$, it follows that for both classes,

$$\text{var}[y_i] = \mathbf{w}^\top \mathbf{\Sigma} \mathbf{w}.$$

In the special case where class priors are the same, a new observation $\mathbf{x} \in \mathbb{R}^D$ is more likely to belong to class 1 than class 2 if $|\mathbf{w}^\top \mathbf{x} - \mathbf{w}^\top \mathbf{m}_1| < |\mathbf{w}^\top \mathbf{x} - \mathbf{w}^\top \mathbf{m}_2|$ and vice versa. Without loss of generality, if $\mathbf{w}^\top \mathbf{m}_1 < \mathbf{w}^\top \mathbf{m}_2$, then $\mathbf{x}$ is more likely to belong to class 1 if $\mathbf{w}^\top \mathbf{x}$ is less than the midpoint of the projected class means. Thus, we obtain a discriminant value

$$c = \frac{1}{2} \mathbf{w}^\top (\mathbf{m}_1 + \mathbf{m}_2)$$

where we classify $\mathbf{x}$ into class 1 if $\mathbf{w}^\top \mathbf{x} < c$ and classify $\mathbf{x}$ into class 2 otherwise.

When class priors are unequal, we can take a more general approach, which we outline in section II-D.

C. Dimensionality Reduction

As in the case of binary classification with LDA, we apply the principle of maximizing the ratio of between-class variance to within-class variance using scatter matrices in order to project high-dimensional datasets onto a low-dimensional subspace. We maintain the assumptions of normality and homogeneity of within-class covariances.

Define the scatter matrix of the set of class means,

$$\mathbf{S}_b = \sum_{k=1}^{C} N_k (\mathbf{m}_k - \mathbf{m})(\mathbf{m}_k - \mathbf{m})^\top$$

to be the *between-class scatter matrix*, and define the sum of the C within-class scatter matrices

$$\mathbf{S}_w = \sum_{k=1}^{C} \mathbf{S}_k,$$

to be the total *within-class scatter matrix*, which by our previous assumptions is approximately equal to $N\mathbf{\Sigma}$.

Our objective is to find a basis $\{\mathbf{v}_1, \ldots, \mathbf{v}_M\}$ for an M-dimensional subspace of $\mathbb{R}^D$ such that the ratio of the between-class scatter to the total within-class scatter is maximized when the data is projected onto the subspace. Intuitively, we would like to find a linear transformation which, when applied to the dataset, maximizes the separation between the classes while preserving the closeness of the observations within each class.

First, let us consider the case when $M = 1$. The objective proposed by Fisher is to find $\mathbf{v}_1 \in \mathbb{R}^D$ for which the between-class to within-class ratio is maximized for the set of points $\{\mathbf{v}_1^\top \mathbf{x}_i\}_{i=1}^N$. We determine the within-class scatter of the kth class when projected onto an arbitrary vector $\mathbf{v} \in \mathbb{R}^D$ as follows:

$$\begin{aligned}
&\sum_{i \in \mathcal{C}_k} (\mathbf{v}^\top \mathbf{x}_i - \mathbf{v}^\top \mathbf{m}_k)^2 \\
&\qquad = \sum_{i \in \mathcal{C}_k} \mathbf{v}^\top (\mathbf{x}_i - \mathbf{m}_k)(\mathbf{x}_i - \mathbf{m}_k)^\top \mathbf{v} \\
&\qquad = \mathbf{v}^\top \left[\sum_{i \in \mathcal{C}_k} (\mathbf{x}_i - \mathbf{m}_k)(\mathbf{x}_i - \mathbf{m}_k)^\top \right] \mathbf{v} \\
&\qquad = \mathbf{v}^\top \mathbf{S}_k \mathbf{v}.
\end{aligned}$$

Similarly, we find that the projected between-class scatter $\{\mathbf{v}^\top \mathbf{m}_k\}_{k=1}^C$ is equal to $\mathbf{v}^\top \mathbf{S}_b \mathbf{v}$.

Thus, to maximize the ratio of the between-class variance to the within-class variance for the projected data, we seek $\mathbf{v}_1$ such that

$$\mathbf{v}_1 = \underset{\mathbf{v}}{\operatorname{argmax}} \frac{\mathbf{v}^\top \mathbf{S}_b \mathbf{v}}{\mathbf{v}^\top \mathbf{S}_w \mathbf{v}}.$$

Note that this ratio depends only on the direction of $\mathbf{v}$, as replacing $\mathbf{v}$ with $a\mathbf{v}$ for any scalar a yields the same ratio. So, we may impose an equality constraint on $\mathbf{v}^\top \mathbf{S}_w \mathbf{v}$ and solve the constrained optimization problem given by

$$\begin{aligned}
\max \quad & \mathbf{v}^\top \mathbf{S}_b \mathbf{v} \\
\text{s.t.} \quad & \mathbf{v}^\top \mathbf{S}_w \mathbf{v} = 1.
\end{aligned}$$

Assuming that $\mathbf{S}_w$ is invertible, the Lagrange condition is given by

$$\begin{aligned}
2\mathbf{S}_b \mathbf{v} - 2\lambda \mathbf{S}_w \mathbf{v} &= \mathbf{0} \\
\mathbf{S}_b \mathbf{v} &= \lambda \mathbf{S}_w \mathbf{v} \\
\mathbf{S}_w^{-1} \mathbf{S}_b \mathbf{v} &= \lambda \mathbf{v}.
\end{aligned}$$

Thus, we see that the vector which solves the optimization problem must be an eigenvector of $\mathbf{S}_w^{-1}\mathbf{S}_b$. Due to the norm constraint, we have that

$$\mathbf{v}^\top \mathbf{S}_b \mathbf{v} = \lambda \mathbf{v}^\top \mathbf{S}_w \mathbf{v} = \lambda,$$

so the maximizing vector $\mathbf{v}_1$ must be an eigenvector associated with the largest eigenvalue of $\mathbf{S}_w^{-1}\mathbf{S}_b$.

When $M > 1$, inductive methods suggest that we take our basis $\{\mathbf{v}_1, \ldots, \mathbf{v}_M\}$ to be eigenvectors associated with the M largest eigenvalues of $\mathbf{S}_w^{-1}\mathbf{S}_b$. Note that the rank of $\mathbf{S}_w^{-1}\mathbf{S}_b$ is bounded above by the rank of $\mathbf{S}_b$. Without loss of generality, if our dataset is centered at the origin (i.e., $\mathbf{m} = \mathbf{0}$), then the sum of all the observations is zero, and we have that

$$\sum_{k=1}^C N_k \mathbf{m}_k = \mathbf{0},$$

so $\mathbf{m}_C$ is linearly dependent on the $C-1$ other class means $\{\mathbf{m}_k\}_{k=1}^{C-1}$. Thus the rank of $\mathbf{S}_b$ is at most $C-1$, and as a result, when performing dimensionality reduction with LDA, we are restricted to projecting data onto an M-dimensional subspace where $M \leq C-1$.

The above derivation assumes that our matrix $\mathbf{S}_w$ is invertible. However, in practice, if the number of features exceeds the number of samples per class, the matrix $\mathbf{S}_w$ will not have full rank. Possible ways to avoid this problem include using the pseudo-inverse of $\mathbf{S}_w$, reducing the dimensionality of the data beforehand with PCA, or using shrinkage to estimate the covariance matrix described by Ledoit and Wolf [8]. One such algorithm for dimension reduction is given below.

Algorithm 1 LDA for Dimensionality Reduction

Require: $\{\mathbf{x}_i\}_{i=1}^N \subseteq \mathbb{R}^D$.

1: Form $\mathbf{S}_w = \sum_{k=1}^C \sum_{i \in C_k} (\mathbf{x}_i - \mathbf{m}_k)(\mathbf{x}_i - \mathbf{m}_k)^\top$.
2: Form $\mathbf{S}_b = \sum_{k=1}^C N_k (\mathbf{m}_k - \mathbf{m})(\mathbf{m}_k - \mathbf{m})^\top$.
3: Compute the M eigenvectors $\{\mathbf{v}_i\}_{i=1}^M$ associated with the largest eigenvalues of $\mathbf{S}_w^{-1}\mathbf{S}_b$, using pseudoinverse if necessary.
4: **for** $j = 1, \ldots, N$ **do**
5: Compute the coefficients of the projection of $\mathbf{x}_j$ onto the subspace spanned by $\{\mathbf{v}_i\}_{i=1}^M$, given by the set $\{\mathbf{v}_i^\top (\mathbf{x}_j - \mathbf{m}_k)\}_{i=1}^M$ where $k = t_j$.
6: **end for**

D. Multiclass Classification

We briefly consider an application of Bayes' theorem to the problem of multiclass classification using LDA. Bayes' theorem tells us that the probability that a new observation $\mathbf{x} \in \mathbb{R}^D$ belongs to the kth class is

$$p(t = k|\mathbf{x}) = \frac{p(\mathbf{x}|t = k)p(t = k)}{p(\mathbf{x})},$$

where we simply take the prior probability to be

$$p(t = k) = \frac{N_k}{N}.$$

As $p(\mathbf{x})$ and N are constant with respect to the class t, the most probable class of $\mathbf{x}$ is the one which maximizes the expression $p(\mathbf{x}|t = k)N_k$. Then, because the all classes are normally distributed with homogeneous covariances, we have

$$p(\mathbf{x}|t = k) = \mathcal{N}(\mathbf{x}|\mathbf{m}_k, \mathbf{\Sigma})$$

where in practice, $\mathbf{\Sigma}$ is estimated from $\mathbf{S}_w$.

We can simplify this problem to finding the class k which maximizes the logarithm of $\mathcal{N}(\mathbf{x}|\mathbf{m}_k, \mathbf{\Sigma})N_k$, ignoring terms which do not depend on k:

$$\begin{aligned}
\hat{t} &= \underset{k}{\operatorname{argmax}} -\frac{1}{2}(\mathbf{x} - \mathbf{m}_k)^\top \mathbf{\Sigma}^{-1}(\mathbf{x} - \mathbf{m}_k) + \log N_k \\
&= \underset{k}{\operatorname{argmax}} \ \mathbf{x}^\top \mathbf{\Sigma}^{-1} \mathbf{m}_k - \frac{1}{2}\mathbf{m}_k^\top \mathbf{\Sigma}^{-1} \mathbf{m}_k + \log N_k
\end{aligned}$$

Note that the quadratic term $\mathbf{x}^\top \mathbf{\Sigma}^{-1} \mathbf{x}$ is dropped due to the homogeneity assumption, so we obtain a linear decision boundary. We can see that when $C = 2$, the difference of objective functions yields the discriminant function found at the end of section II-B with some correction for class priors.

In the case when within-class covariance is isotropic (i.e., $\mathbf{\Sigma} = \sigma^2\mathbf{I}$) and class sizes are equal, we simply find the class mean $\mathbf{m}_k$ which is closest to $\mathbf{x}$.

III. EXPERIMENTS

A. LDA on Synthetic Data

We would like to observe the expected behavior of LDA on a dataset well-suited to its application. To do this, we create a synthetic dataset, sampling $N = 1000$ points in $\mathbb{R}^{100}$, each of which belong to one of $C = 5$ classes. To satisfy our assumptions in II-A, we sample points for each class from a Gaussian distribution, with homogeneous (and isotropic) covariance among classes. We then apply LDA dimensionality reduction to project the data onto a 2-dimensional subspace of $\mathbb{R}^{100}$ and visualize the results, and then use LDA for multiclass classification. We expect LDA to find a projection subspace that optimally separates the classes, which should present itself as a distinct separation between each class when the projected data is plotted.

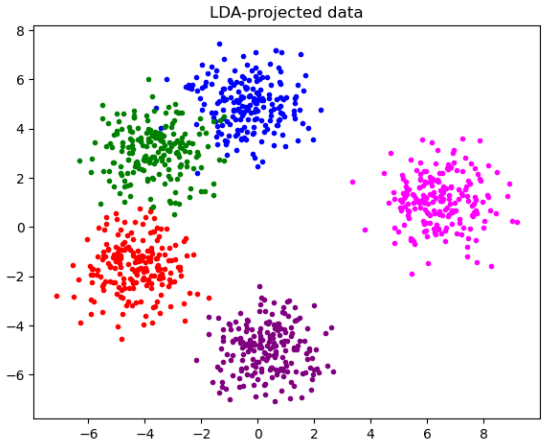

Fig. 1. LDA projection of synthetic data in $\mathbb{R}^{100}$.

The results of this projection are shown in Figure 1. We find that LDA does indeed find a projection subspace that separates classes well, despite the subspace being very low-dimensional ($M = 2$) when compared to the original feature space ($D = 100$). This shows that when the assumptions of normality and homogeneity are satisfied, LDA is an effective tool to extract the most discriminating features from high-dimensional data. In addition, because the classes were well separated and satisfied the assumptions of the model, LDA produced 100% accuracy on the synthetic data when used as a classifier.

B. LDA for Star Classification

We use LDA to classify stars from a set of observable attributes. More specifically, we classify stars as brown dwarfs, red dwarfs, white dwarfs, main sequence, supergiants, or hypergiants ($C = 6$) depending on a number of their features present in our data [9], namely the stars' temperature, radius, luminosity, absolute magnitude, apparent color, and spectral type ($D = 6$). Our data contains $N = 240$ samples total, with each class containing $N_k = 40$ observations.

We turn our attention to spectral type, which is an attribute of a star given by properties already present in our data, such as temperature and luminosity. Critical to nearly all classification algorithms, LDA requires the data's attributes to be independent of each other, and in our testing, we find that excluding the spectral type results in a slightly more accurate model.

We train our model by randomly splitting our data into 70% for training and 30% for testing. Over 50 such splits, our model achieved a mean testing accuracy of 98.9% with a standard deviation of 1.1%.

C. LDA for Data Visualization

Though not discussed at great length in this paper, LDA can also be used as a tool to assist in the visualization of labeled data. Due to limitations of human vision, visualization of data with dimensionality greater than 3 is unintuitive at best. However, LDA's ability to optimally discriminate between classes during projection can provide an effective means to visualize the relationships between classes. We briefly demonstrate this ability by projecting our star data to two dimensions and visualizing the results in Figure 2.

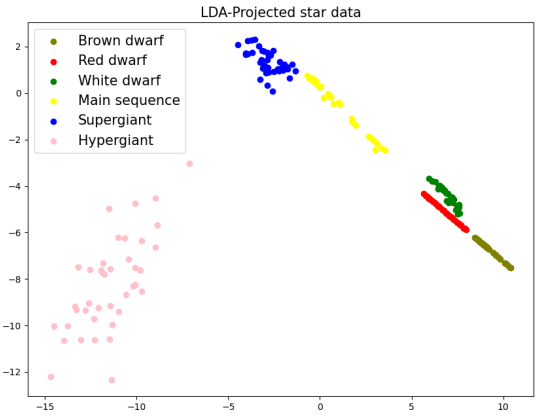

Fig. 2. LDA reduction of star data to two dimensions.

Coincidentally, we find that LDA seems to approximate the structure of a Hertzsprung-Russell (HR) diagram, the standard diagram used to visualize star types in relation to their temperature and luminosity: see Figure 3 for one such example. The axes of the HR diagram are often simply temperature and luminosity, so we cannot expect the axes to be equivalent in our projection subspace, however the overall shape of the HR diagram is preserved modulo some affine transformation of our projected data points. There is a distinct diagonal main sequence branch, with white dwarfs clearly separated from the main sequence branch, and the giants diverging off to the opposite side of the white dwarfs.

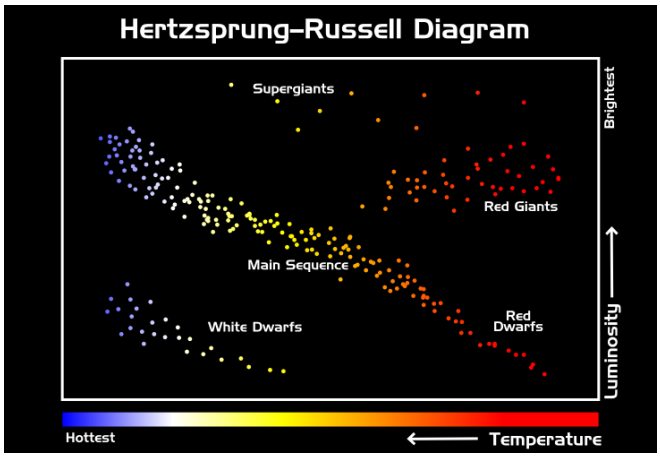

Fig. 3. Example of a Hertzsprung-Russell diagram. Source: `http://aspire.cosmic-ray.org/Labs/StarLife/studying.html`

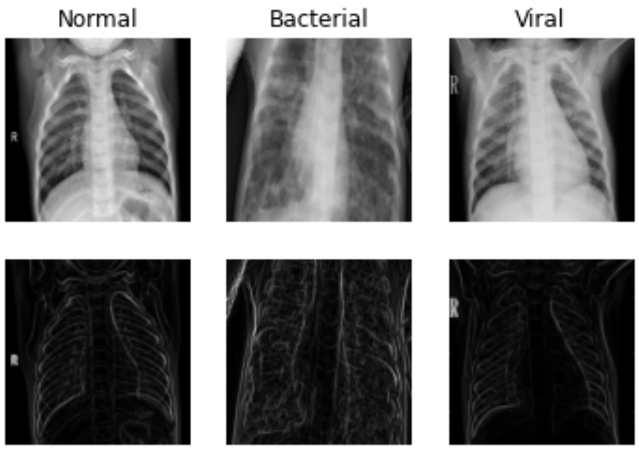

Fig. 4. X-ray images with and without Sobel filter.

D. X-Ray Image Classification

Li, Zhu, and Ogihara [10] find that LDA is an effective multiclass classifier in facial and object recognition, despite the image data used in these tasks frequently violating the homogeneity assumption necessary for LDA. In view of this, we want to explore LDA's robustness to highly variant data.

We consider the problem of diagnosing pneumonia in patients using images of their chest X-rays. Our training dataset consists of a set of $N = 2250$ grayscale chest X-ray images with varying dimensions belonging to $C = 3$ classes (with $N_k = 750$ for each class) that have been sampled from a database of around 5000 images [3]. In order to standardize the dimensionality of the dataset, we resize all images to dimensions 220×220 pixels and flatten each image array to generate our dataset $\{\mathbf{x}_i\}_{i=1}^{2250} \subset \mathbb{R}^{220*220}$. Then, we apply horizontal and vertical Sobel filters to each image in order to exploit some common differences between normal and pneumonia X-rays, such as some cloudiness in the ribcage as seen in the pneumonia X-rays in Figure 4.

Each image is classified with one of three diagnoses: normal, bacterial pneumonia, or viral pneumonia. Our classification technique takes two steps: We first consider all three diagnosis types as separate classes in order to reduce the dimension of each $\mathbf{x}_i$ to 2 using the LDA technique described in section II-C. Then, in order to optimally train our binary classifier, we group both types of pneumonia under a single class and sample 750 images from this new class to maintain evenness between the size of our classes. Finally, we apply a binary classifier to the dimensionality-reduced data. Figure 4 shows a sample of images from the dataset. After finding the two eigenvectors satisfying Fisher's criterion for the training dataset, we plot the projection of the data onto the subspace spanned by this basis, which is shown in Figure 5. The projection of the test dataset using the train-fitted subspace is shown in Figure 6. The training dataset seems to be well separated due to overfitting of the LDA model to the training set. Nevertheless, our various binary classifiers, such as logistic regression, SVM, and binary LDA

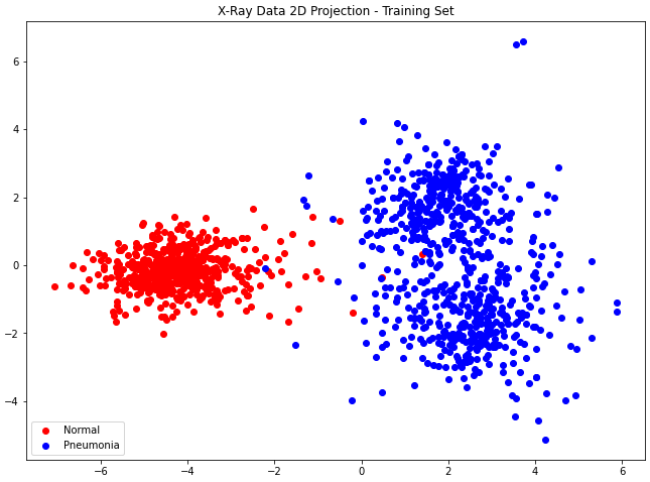

Fig. 5. Projection of training set to two dimensions.

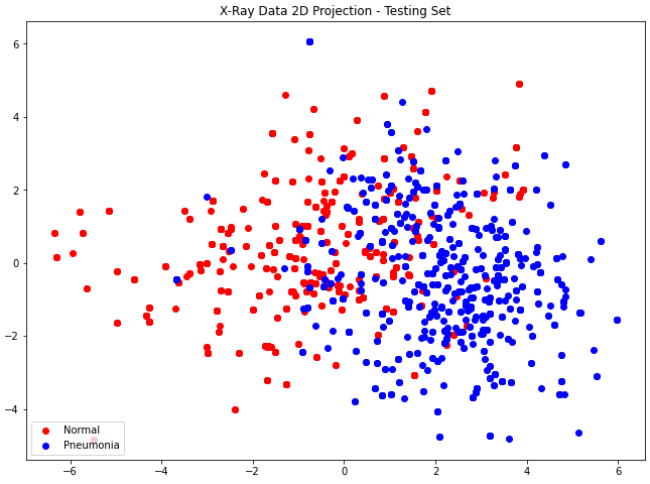

Fig. 6. Projection of test set to two dimensions using model from training set.

all achieve a prediction accuracy rate of about 70% without a Sobel filter and 75% with a Sobel filter.

In addition to overfitting of the model, we can also explain the poor generalization of the train-fitted projection subspace to the test data by noting the high variance of our data. Figure 4 shows the variation in anatomy and positioning, causing the images to be highly irregular. To improve LDA's performance on this dataset, one should consider restricting the dataset to images that are controlled for body type, positioning, etc. or consider another regularization routine to reduce the variance of the data.

IV. Advantages and Disadvantages

One of LDA's most useful features is its ability to reduce the dimensionality of data while also serving as a robust classifier method. This enables us to visualize data in lower dimensions, allowing for easily interpretable results. The fact that the decision boundary is linear means that the algorithm is easy to implement and yields a simple but robust classification model.

One of LDA's disadvantages is that it is perhaps too simple. A linear decision boundary may not yield an adequate classification accuracy. The assumption of normally distributed data also limits its ability to work with more complex datasets. In the high dimensional problem, if the sample size is small, it becomes impossible to solve for eigenvalues and eigenvectors of $\mathbf{S}_w^{-1}\mathbf{S}_b$ without modifications to the derivation.

Our experiment with the chest X-ray dataset also points to LDA showing a lack of robustness to highly variable data, with its ability to discriminate classes not generalizing well to test data when the train data is highly variable with many attributes.

V. Conclusions

In this paper, we explained the mathematical foundations of LDA and applied it to various types of classification and dimension reduction problems in order to demonstrate its efficacy. We have used Fisher's criterion to solve for the optimal directions of projection and considered the application of Bayes' theorem to the general multiclass classification problem.

In our experiments, we have applied LDA to a synthetic data set in order to showcase its performance in the multiclass classification problem, yielding promising results despite the reduction of a 100-dimensional space to a 2-dimensional space. We achieved a high degree of accuracy when applying LDA for star classification while demonstrating LDA's data visualization capabilities by comparing its results to the Hertzsprung-Russell diagram. Finally, we obtained surprisingly accurate results when using LDA to help diagnose instances of pneumonia in a highly variant dataset of X-ray images. All content of this project, including the code for our experiments, are available on our GitHub repository: `https://github.com/warewaware/LDA156FinalProj`.

References

[1] J. Friedman, T. Hastie, and R. Tibshirani, *The Elements of Statistical Learning*. Springer series in statistics New York, 2013, vol. 1(10).

[2] H. Yu and J. Yang, "A Direct LDA Algorithm for High-Dimensional Data - with Application to Face Recognition," *Pattern Recognition*, vol. 34, pp. 2067–2070, 2001.

[3] D. S. Kermany *et al.*, "Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning," *Cell*, vol. 172, no. 5, pp. 1122–1131, Feb. 2018. DOI: 10.1016/j.cell.2018.02.010.

[4] E. I. Altman, "Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy," *The Journal of Finance*, vol. 23, no. 4, pp. 589–609, Sep. 1968.

[5] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.

[6] R. Fisher, "The Use of Multiple Measurements in Taxonomic Problems," *Annals of Eugenics*, vol. 7, no. 2, pp. 179–188, 1936.

[7] C. Rao, "The Utilization of Multiple Measurements in Problems of Biological Classification," *Journal of the Royal Statistical Society*, vol. 10, no. 2, pp. 159–203, 1948.

[8] O. Ledoit and M. Wolf, "A well-conditioned estimator for large-dimensional covariance matrices," *Journal of Multivariate Analysis*, vol. 88, no. 2, pp. 365–411, 2004, ISSN: 0047-259X. DOI: https://doi.org/10.1016/S0047-259X(03)00096-4.

[9] D. Baidya. (2019). "Star dataset to predict star types," [Online]. Available: https://www.kaggle.com/deepu1109/star-dataset.

[10] T. Li, S. Zhu, and M. Ogihara, "Using Discriminant Analysis for Multi-class Classification: an Experimental Investigation," *Knowledge and Information Systems*, vol. 10, pp. 453–472, Mar. 24, 2006.