A report on multiclass linear discriminant analysis

Mathav Raj - 194102311

***Abstract*— To study LDA as a dimensionality reduction technique and visualize it's application as a classifier. The goal is to project a data set onto a lower-dimensional space with good class-separability in order to avoid over fitting ("curse of dimensionality").**

I. INTRODUCTION

Linear Discriminant Analysis is a dimensionality reduction technique which is commonly used for the supervised classification problems. It is used for modeling differences in groups i.e. separating two or more classes. It is used to project the features in higher dimension space into a lower dimension space.

II. ALGORITHM

A. Algorithm

Consider a set of d dimensional points X, $x_1, x_2, ..., x_n$ with c classes. $d \geq c$. Fischer's linear discriminant involves projection from a d dimensional space to a $c-1$ dimensional subspace. The net within class scatter is the sum of all within class scatters,

$$S_W = \sum_{i=1}^{c} S_i \tag{1}$$

where,

$$S_i = \sum_{x \in D_i}^{n} (x - m_i)(x - m_i)^T \tag{2}$$

where m_i is the mean of each individual class i varying from 1 to c. The proper generalization of between class scatter can be framed from the total scatter as,

$$\begin{aligned} S_T &= \sum_{x} (x - m)(x - m)^T \\ &= \sum_{i=1}^{c} \sum_{x \in D_i} (x - m_i + m_i - m)(x - m_i + m_i - m)^T \\ &= \sum_{i=1}^{c} \sum_{x \in D_i} (x - m_i)(x - m_i)^T \\ &\quad + \sum_{i=1}^{c} \sum_{x \in D_i} (m_i - m)(m_i - m)^T \\ &= S_W + \sum_{i=1}^{c} N_i (m_i - m)(m_i - m)^T \end{aligned}$$

$$S_B = \sum_{i=1}^{c} N_i (m_i - m)(m_i - m)^T \tag{3}$$

The projection from d dimensional space to a $c-1$ dimensional space is accomplished by $c-1$ discriminant functions,

$$y = W^T x \tag{4}$$

we want to find a transformation matrix that maximizes the ratio of between class scatter to within class scatter in the projected subspace.

$$S_W = W^T S_W W \tag{5}$$

$$S_B = W^T S_B W \tag{6}$$

$$J(W) = \frac{|W^T S_B W|}{|W^T S_W W|} \tag{7}$$

where W is the projection matrix of size $dxc-1$ where each column is the eigen vector found by solving,

$$S_B w_i = \lambda_i S_W w_i \tag{8}$$

Algorithm 1 Multiple Linear discriminant analysis

```
Compute the d-dimensional mean vectors for the different classes from the data set
Compute the scatter matrices (within-class and in-between-class scatter matrix) according to equations 1,2 and 3
According to equation 8 find the d eigen values and the corresponding eigen vectors of $S_W^{-1} S_B$
Sort the eigen vectors by decreasing eigenvalues and choose $k$ eigen vectors with the largest eigenvalues to form a $dk$ dimensional matrix $W$ (where every column represents an eigen vector)
Transform the data samples $X$ onto the new subspace by the matrix multiplication: $Y = XW$ (where $X$ is a $nd$-dimensional matrix representing the $n$ samples, and $Y$ is the transformed $nk$-dimensional samples in the new subspace)
LDA learns a linear decision boundary assuming a gaussian distribution for each class on the projected data and learns the parameters from the projected data and argmax of the linear discriminant function is decided as that particular data point's class
```

B. Experiment on IRIS and WINE dataset:

Iris data set has 4 features and 3 classes. Wine data set has 13 features and 3 classes. Both PCA and LDA can reduce the dimensions of data but LDA is a supervised whereas PCA is unsupervised – PCA ignores class labels. The linear

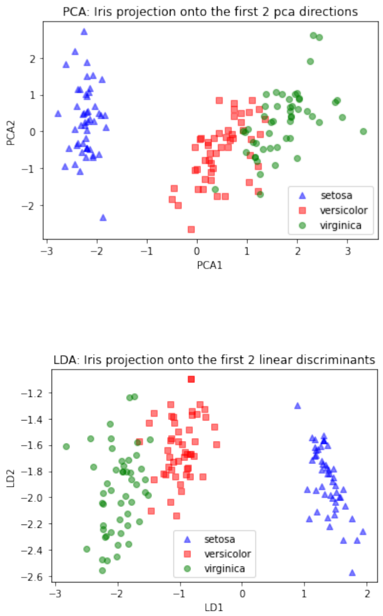

Fig. 1: LDA vs PCA for Iris

Discriminant analysis takes the mean value for each class and considers variants to make predictions assuming a Gaussian distribution. Here it can be seen that LDA does better in separating out the classes. For Iris the considerable overlap between Versicolor and Virginica is comparatively lesser in LDA than in PCA. For Wine dataset it is much more clear how the three classes are separable.

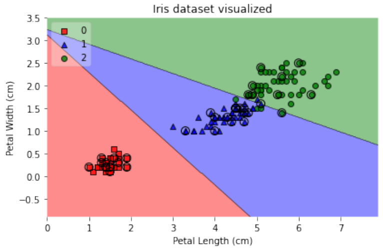

Fig. 3: Linear decision boundary between selected features

The data set has four different features which makes it harder to visualize the decision boundary. To make it easier to visualize, we consider only two features petal length and petal width. We use sklearn library to learn the decision boundary from training data which is 75% of original data points.

LDA is robust. The class separation is not affected by standardization. Though the scatter matrices are different it was observed that the eigen values were almost the same. Standardization is necessary in PCA to cancel the illegal dominating effect of wider ranges in features which are measured in different scales.

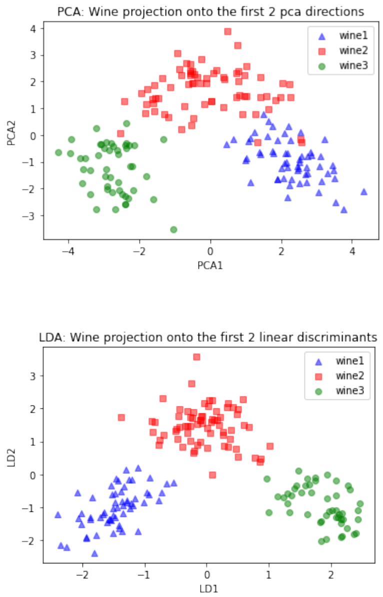

Fig. 2: LDA vs PCA for Wine

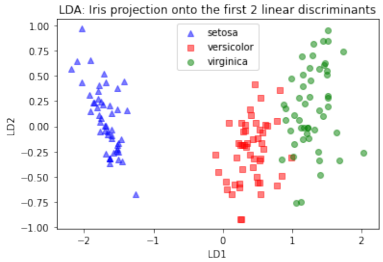

Fig. 4: Effect of standardization

C. *Drawbacks:*

- Linear decision boundaries may not adequately separate the classes. Support for more general boundaries is desired.
- In a high-dimensional setting, LDA uses too many parameters. A regularized version of LDA is desired.

III. CONCLUSION AND FURTHER STUDIES

One of the challenges of these subspace projection techniques for classification is when we encounter small sample size and also a study on incorporating different weightage to different classes is needed for a cost sensitive classification scheme

IV. REFERENCES

- "Pattern Classification", second edition,David G. Stork, Peter E. Hart, and Richard O. Duda
- https://jss367.github.io/Visualize-shallow-learning.html
- https://machinelearningmastery.com/linear-discriminant-analysis-for-machine-learning