A Dual Approach to Linear Discriminant Analysis Application to the Financial Sector

Imperial College London
Supervised by Ioanna Papatsouma

Raquel Rogero Morón (01564308), Alix Vermeulen (01564896), Kyle Patel (01509564), Kailun (Helen) Peng (01575789)

June 16, 2020

Outline

Introduction

Linear Discriminant Analysis (LDA)

Classification method

Figure 1: R.A. Fisher and C.R. Crao

- developed by R. A. Fisher in 1936, traditionally only for binary cases [1].
- extended for multi-class cases in 1948 by C.R.Rao (MDA) [2].

Dimensionality reduction technique

- Aim: projecting the group features observed on higher dimensional spaces onto smaller subspaces without losing information.
- Reduces the number of dimensions, the difficulty of the computations, avoids a curse of dimensionality problem...[3].

Linear Discriminant Analysis (LDA)

Dimensionality reduction technique

Two different approaches:

- unsupervised approach.
- supervised approach.

LDA is one of the most famous techniques of the supervised approach [4].

Linear combination of the continuous independent variables [4, 5]:

- class-dependent transformation.
- class-independent transformation.

LDA requires of four strong assumptions that are easily violated [6].

Motivation

- Supervised method commonly used to classify data.
- Interestingly popular over time as we seek for patterns in data to extract meaningful conclusion to be able to solve problems.

Aim of the report:

Assess the accuracy of two following methods on an example in the financial sector [4].

The two methods are:
- simply as a classification algorithm.
- reducing the dimension of the problem as a pre-processing step for classification.

Advantages and Disadvantages

Advantages

- Very simple and clear method to implement.
  Although it can have large matrix computations, can still be calculated fast.
- Produces strong results, also due to the strong assumptions made as a precursor.
- Not complex to train.
- Can be expanded to a multi class classifier unlike other methods.
- LDA still functions correctly even when the two classes are perfectly separable.

Advantages and Disadvantages

Disadvantages

- Strong assumptions.
- Fisher's linear discriminant is sensitive to outliers [7]. Stronger techniques have been developed that are not as sensitive to them.
- The SSS problem: total number of training samples smaller than the number of dimensions of the feature vector [8].
- Large matrix computations: if a large number of dimensions are being used, can be computationally very expensive.

Literature Review

LDA has a wide range of applications in the modern world:

Pattern recognition

- Face recognition [9], re-indentification [10].
- Emotion recognition, hand motion categorisation.
- Classification process of speech, music classification [11].
- Food quality evaluation [12].

Biostatistics

- Classification of genes [13].
- Identification of chemical toxicants [14].

Literature Review

Medicine

LDA plays a significant role to help doctors with disease diagnostics [15, 16].

- Distinguish between benign and malignant lumps.
- Categorise the condition of a patient with a certain disease.
- Provide information on what kind of treatments the patients should receive.

Social and behavioural sciences

- Predict recidivism [17].

Business and finances

- Predict the likelihood of bankruptcy of a business [18].

Report Structure

Methodology

- Assumptions
- Univariate and Multivariate Gaussian Distribution
- Classification
- Reduction of Dimensions and Classification

Application

- Methodology in R
- Classification
- Reduction before classification
- Summary of Results
- Discussion

Conclusion

Assumptions

- 4 strong assumptions that are not always respected in numerous cases.
- LDA as a classification method: greatly affected when these assumptions are not satisfied [3].
- LDA as a dimensionality reduction technique: robust to the violation of the assumptions.

The 4 Assumptions

1. **Independence**

   LDA has one categorical dependent variable corresponding to the group variable (e.g. bankruptcy or not) and then deals with continuous independent variables.

2. **Multicollinearity**

   There exists linear relations between the different independent variables [6].

3. **Gaussian distribution**

   The data of each different group is Gaussian distributed.

4. **Homoscedasticity**

   The variance-covariance matrices of each group are equal.

Classification

How to classify data

Decision boundary: $G(x) = arg\max_k P(Z = k \mid X = x)$

$$
\begin{aligned}
P(Z = k \mid X = x) &= \frac{P(X = x \mid Z = k)P(Z = k)}{P(X = x)} \\
&= \frac{f_k(x)\pi_k}{\sum_{i=1}^{n} P(X = x \mid Z = i)P(Z = i)} \\
&\propto f_k(x)\pi_k
\end{aligned}
$$

Using our assumptions

$$f_k(x) = \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma_k}|^{1/2}} \exp\left\{ -\frac{1}{2}(x-\mu_k)^T \boldsymbol{\Sigma_k}^{-1}(x-\mu_k)\right\}$$

μ_k - mean of the inputs from class k
$\boldsymbol{\Sigma_k} = \boldsymbol{\Sigma}$ - variance-covariance matrix equal for all classes
$|\boldsymbol{\Sigma_k}|$ - determinant of the variance-covariance matrix

$$
\begin{aligned}
P(Z = k \mid X = x) &\propto f_k(x)\pi_k \\
&\propto \frac{\pi_k}{(2\pi)^{p/2}|\mathbf{\Sigma}|^{1/2}} \exp\left\{ -\frac{1}{2}(x-\mu_k)^T \mathbf{\Sigma}^{-1}(x-\mu_k) \right\} \\
&\propto \pi_k \exp\left\{ -\frac{1}{2}(x-\mu_k)^T \mathbf{\Sigma}^{-1}(x-\mu_k) \right\} \\
&= C\pi_k \exp\left\{ -\frac{1}{2}(x-\mu_k)^T \mathbf{\Sigma}^{-1}(x-\mu_k) \right\}
\end{aligned}
$$

where C is a constant of proportionality.

Maximising the posterior probability

$$
\begin{aligned}
\log(P(Z = k \mid X = x)) =& \log(\pi_k) - \frac{1}{2}(x - \mu_k)^T \boldsymbol{\Sigma}^{-1}(x - \mu_k) \\
=& \log(\pi_k) - \frac{1}{2}[x^T \boldsymbol{\Sigma}^{-1} x + \mu_k^T \boldsymbol{\Sigma}^{-1} \mu_k] \\
& + x^T \boldsymbol{\Sigma}^{-1} \mu_k \\
=& D + \log(\pi_k) - \frac{1}{2}\mu_k^T \boldsymbol{\Sigma}^{-1} \mu_k + x^T \boldsymbol{\Sigma}^{-1} \mu_k
\end{aligned}
$$

where $D = -\frac{1}{2} x^T \boldsymbol{\Sigma}^{-1} x$
and hence we obtain the linear discriminant functions for each class k:

$$
\delta_k(x) = \log(\pi_k) - \frac{1}{2}\mu_k^T \boldsymbol{\Sigma}^{-1} \mu_k + x^T \boldsymbol{\Sigma}^{-1} \mu_k
$$

Estimating the parameters

$$\delta_k(x) = \log(\pi_k) - \frac{1}{2}\mu_k^T \boldsymbol{\Sigma}^{-1}\mu_k + x^T \boldsymbol{\Sigma}^{-1}\mu_k$$

$$\hat{\pi}_k = \frac{N_k}{N} \quad \text{where } N_k \text{ is the number of observations in class } k$$

$$\hat{\mu}_k = \frac{1}{N_k}\sum_{z_i=k} x_i$$

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{N-K}\sum_{k=1}^{K}\sum_{z_i=k}(x_i - \hat{\mu}_k)(x_i - \hat{\mu}_k)^T$$

Decision boundary: $G(x) = arg\max_k \delta_k(x)$
We assign x to the class k that yields the greatest discriminant function (largest posterior probability).

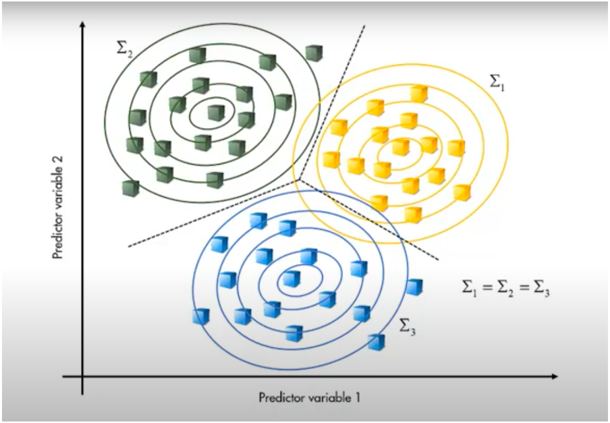

Figure 2: Linear boundaries between 3 classes

Reduction of Dimensions

Binary Case

P - dimensional input vector x

Project down to a one dimensional space.

$y = w^T x$

Separability

$$y = w^T x$$

Then if a condition is imposed on y, we can classify $y \geq c$ as class Z_1 and $y < c$ as Z_2.

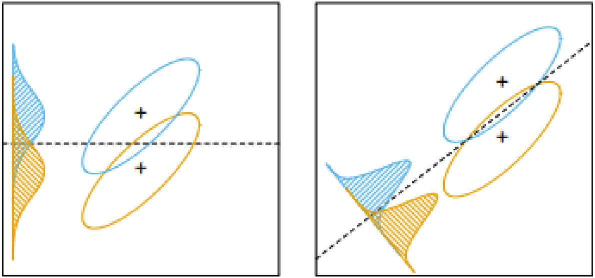

Figure 3: Separability of two classes depending on projection [20]

Fisher's Linear Discriminant

Consider there being N_1 points in class Z_1 and N_2 points in class Z_2. The 2 classes have mean vectors:

$m_i = \frac{1}{N_i} \sum_{z \in Z_i} x_z \quad$ for i = 1,2 in this case.

$\uparrow$ between class variance whilst $\downarrow$ within class variance.

We define the within class variance as
$s_i^2 = \sum_{z \in Z_i} (y_z - m_i)^2 \quad$ where $y_z = w^T x_z$

We define the total within class variance to be $s_1^2 + s_2^2$.

Fisher's Linear Discriminant

Fisher's criterion:

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

To maximise with respect to w.

Explicitly we have so far:

$$J(w) = \frac{w^T (m_2 - m_1) (m_2 - m_1)^T w}{w^T (C_1 + C_2) w}$$

where C_i is the covariance matrix of x_i

Scatter Matrices

Scatter Matrices are defined as $S_i = N_i C_i$

Now we introduce:

$$S_B = (m_2 - m_1)(m_2 - m_1)^T$$
$$S_W = S_1 + S_2$$

We can therefore define Fisher's criterion as

$$J(w) = \frac{w^T S_B w}{w^T S_W w}$$

Obtaining the optimal w

$$
\begin{aligned}
S_W w &= \frac{w^T S_W w}{w_T S_B w} S_B w \\
&= \frac{w^T S_W w}{w^T S_B w} (m_2 - m_1)(m_2 - m_1)^T w \\
&= \frac{w^T S_W w}{w^T S_B w} (m_2 - m_1)^T w \, (m_2 - m_1)
\end{aligned}
$$

Since $\frac{w^T S_W w}{w^T S_B w} (m_2 - m_1)^T w$ is a scalar value we omit this. This means that

$$w \propto S_W^{-1}(m_2 - m_1)$$

Extending to K classes

We first define:

$$X_k = \{x_i | y_i = k\},$$

$$m_k = \frac{1}{N_k} \sum_{x \in X_k} x$$

$$S_k = N_k C_k = \sum_{x \in X_k} (x - m_k)(x - m_k)^T$$

$$S_W = \sum_{k=1}^{K} S_k$$

Between Scatter

$$S_T = \sum_{i=1}^{N} (x_i - m)(x_i - m)^T \quad \text{where in this case } m = \frac{1}{N}\sum_{i=1}^{N} x_i$$

$$S_T - S_W = \sum_{k=1}^{K} N_k (m_k - m)(m_k - m)^T = S_B$$

Which allows us to then attempt to maximise:

$$J(w) = \frac{w^T S_B w}{w^T S_W w}$$

Obtaining the optimal w

Solutions determined by the eigenvectors of $S_W^{-1}S_B$, when S_W is invertible. The optimal w is given by the corresponding eigenvector of the largest eigenvalue.

Up to $(K-1)$ features for K classes [19].

Direct consequence of $m = \frac{1}{N}\sum_{i=1}^{N} x_i = \frac{1}{N}\sum_{k=1}^{K} N_k m_k$

Modelling in R

Predicting Bankruptcy

Performance in predicting Bankruptcy.

Data exhibits 64 statistics based around Bankruptcy which we use to predict the status of Polish companies in reference to Bankruptcy for up to 5 years [21].

Classification vs Reduction of Dimensions then Classification.

In R...

- Make use of MASS library to perform LDA on the data with the lda function.
- Small sample size problem encountered $\rightarrow$ singular within scatter matrix.
- 4:1 Ratio
- Classified so N^{th} class will go bankrupt in n years.

Results and Discussion

Results: confusion matrices

Classification

		Actual					
		0	1	2	3	4	5
Predicted	0	3854	16	20	18	18	7
	1	13	0	0	0	1	0
	2	7	1	0	0	0	0
	3	14	0	0	0	0	0
	4	4	0	0	0	0	0
	5	19	0	1	0	0	0

Results: confusion matrices

Reduction-classification

		Actual					
		0	1	2	3	4	5
Predicted	0	3881	16	22	18	18	7
	1	17	0	0	0	1	0
	2	2	1	0	0	0	0
	3	5	0	0	0	0	0
	4	2	0	0	0	0	0
	5	4	0	1	0	0	0

Overall Statistics

Classification

```
Overall Statistics

               Accuracy : 0.9647
                 95% CI : (0.9585, 0.9702)
```

Figure 4: overall statistics of classification method

Reduction-classification

```
Overall Statistics

               Accuracy : 0.9715
                 95% CI : (0.9658, 0.9764)
```

Figure 5: overall statistics of reduction-classification method

Overall Statistics

Classification

Statistics by Class:

	Class: 0	Class: 1	Class: 2	Class: 3	Class: 4	Class: 5
Sensitivity	0.99233	0.000000	0.0000000	0.000000	0.0000000	0.000000
Specificity	0.03571	0.995475	0.9992447	0.998743	0.9994970	0.998746

Figure 6: Statistics by class of classification method

Reduction-classification

Statistics by Class:

	Class: 0	Class: 1	Class: 2	Class: 3	Class: 4	Class: 5
Sensitivity	0.98543	0.000000	0.000000	0.000000	0.000000	0.000000
Specificity	0.05952	0.996732	0.997734	0.996480	0.998994	0.994483

Figure 7: Statistics by class of reduction-classification method

Figures on Separability

Classification

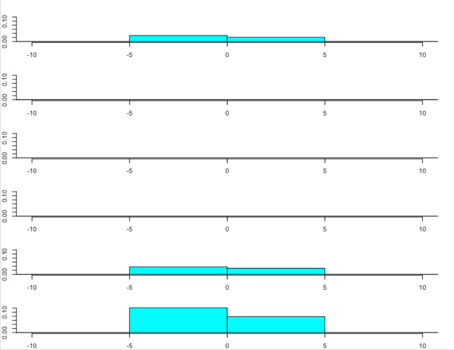

Figure 8: Histogram of projection on LD1

Reduction-classification

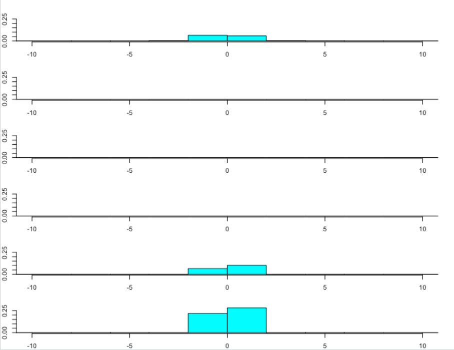

Figure 9: Histogram of projection on LD1

Assumptions

Normality of data

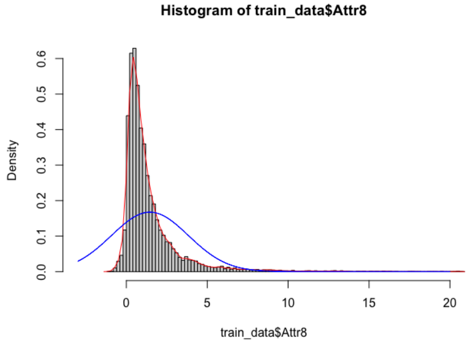

Figure 10: Distribution of X8(profits on operating activities/total assets)

- In general, the distribution of variables in our data set is not Gaussian.

Limitations

- Skewness of our data: more than 97% of observations are from class 0.
- Small sample size problem
- Due to the time constraint, LDA was only applied to one data set.

Conclusion

Conclusion

- In this project, we studied the maths behind LDA and applied it twice to a Polish company dataset for bankruptcy prediction.
- LDA has a high accuracy as both methods, with reduction-classification(97.15%) slightly higher than classification only(96.52%) in our case.
- The performance of LDA relies on the nature of the actual data set.
- Can consider using other methods which are not sensitive to imbalanced data.

References

[1] R. A. Fischer. The use of multiple measurements in taxonomic problems. *Annals of eugenics*, Vol. 7, No.2, pp.179–188, 1936.

[2] C. R. Rao. The Utilization of Multiple Measurements in Problems of Biological Classification. *Journal of the Royal Statistical Society. Series B (Methodological)*, Vol. 10, No. 2, pp. 159-203, 1948.

[3] S. Raschka. *Linear Discriminant Analysis - Bit by Bit*, viewed 31 May 2020, `http://sebastianraschka.com/Articles/2014_python_lda.htm`

[4] A. Tharwat, T. Gaber, A. Ibrahim, A. E. Hassanien. Linear discriminant analysis: A detailed tutorial *AI Communications*, Vol. 30, No. 2, pp 169–190, 2017.

[5] S. Balakrishnama, A. Ganapathiraju. Linear discriminant analysis - a brief tutorial, *Institute for Signal and Information Processing*, 1998.

[6] G.J. McLachlan. Discriminant analysis, *Wiley Interdisciplinary Reviews: Computational Statistics*, Vol. 4, No. 5, pp.421-431, 2012.

[7] F. Nie, H. Wang, Z. Wang, H. Huang. *Robust Linear Discriminant Analysis Using Ratio Minimization of L1,2-Norms* July 2, 2019.

[8] M. Kamel, A. Campilho. *Image Analysis and Recognition* 10th International Conference, ICIAR 2013, Póvoa do Varzim, Portugal, June 26-28, 2013.

[9] J. Lu, K.N. Plataniotis, A.N. Venetsanopoulos. Regularization studies of linear discriminant analysis in small sample size scenarios with application to face recognition. *Pattern Recognition Letters*. 26 (2): pp.181-191, 2005.

[10] L. Wu, C. Shen, A. v.d. Hengel. Deep linear discriminant analysis on fisher networks: A hybrid architecture for person re-identification. *Pattern Recognition*, 65 238-250, 2017.

[11] E. Alexandre-Cortizo, M. Rosa-Zurera, F. Lopez-Ferreras. Application of Fisher Linear Discriminant Analysis to Speech/Music Classification. *EUROCON 2005 - The International Conference on "Computer as a Tool"* : IEEE, pp. 1666-1669, 2005.

[12] D. Sun *Computer vision technology for food quality evaluation.* 1. ed. ed. Burlington [u.a.]: Academic Press; 2008.

[13] Y. Guo, T. Hastie, R. Tibshirani, Regularized linear discriminant analysis and its application in microarrays. *Biostatistics*, Volume 8, Issue 1, Pages 86–100, January 2007.

[14] W. Zhong, K.S. Suslick, Matrix Discriminant Analysis With Application to Colorimetric Sensor Array Data. *Technometrics*, 57 (4): 524-534, 2015.

[15] A. J. Izenman. *Mordern Multivariate Statistical Techniques*, 2008.

[16] H. Rajaguru, S. K. Prabhakar. *Bayesian Linear Discriminant Analysis for Breast Cancer Classification*, 2017 2nd

International Conference on Communication and Electronics Systems (ICCES), Coimbatore, 2017, pp. 266-269, doi: 10.1109/CESYS.2017.8321279.

[17] N. Tollenaar, P. G. M. van der Heijden. Which method predicts recidivism best?: a comparison of statistical, machine learning and data mining predictive models, *Journal of the Royal Statistical Society*. Series A (Statistics in Society), Vol. 176, No. 2 (FEBRUARY 2013), pp. 565-584.

[18] E. I. Altman. Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy, *The Journal of Finance*, Vol. 23, No. 4 (Sep., 1968), pp. 589-609.

[19] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, Cambridge, 2006. p.181-192.

[20] G. James, D. Witten, T. Hastie, R. Tibshirani. *An Introduction to Statistical Learning with Applications in R*, Springer, New York, 2013, pp.142-143.

[21] M. Zikeba, S. K.Tomczak, J. M.Tomczak. *Ensemble Boosted Trees with Synthetic Features Generation in Application to Bankruptcy Prediction*, Expert Systems with Applications, Elsevier, 2016.