Primer on Linear Discriminant Analysis (LDA)

What is it?

It's a classification method that identifies the combination of features that distinguish between two or more groups of datapoints - specifically for data where the groups are known, and the predictors are continuous numeric variables.

For example, if our data has 4 known groups, LDA can plot the clusters on multiple dimensions in a way that clearly separates these groups. The dimensions are the "features" in the model.

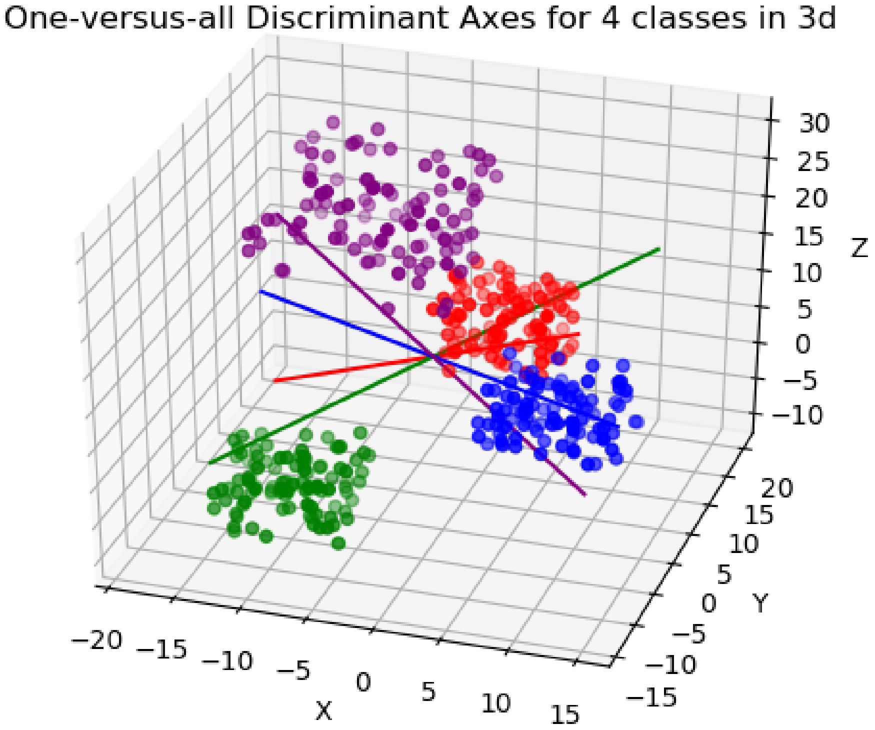

Figure 1: plot of 3-D axes with 4 coloured clusters of points

Image by Amélia Oliveira Freitas da Silva - Own work, CC BY-SA 4.0, https://commons.wikimedia.org/w/index.php?curid=104693008

It's like logistic regression, but where the outcome is categorical and all the predictors are continuous numeric. It's also like Principal Components Analysis, except in LDA the group labels are known (in PCA they are not).

As well as being a classifier, LDA is useful for dimension reduction - e.g. if our original dataset with 4 groups has 20 variables, LDA has found the 3 or 4 main features that distinguish between these groups, without us needing to have 20 predictors in our model. (Note: The dimensions from LDA may not be the raw continuous variables.)

It can also classify groups even when the clusters are interspersed, e.g.:

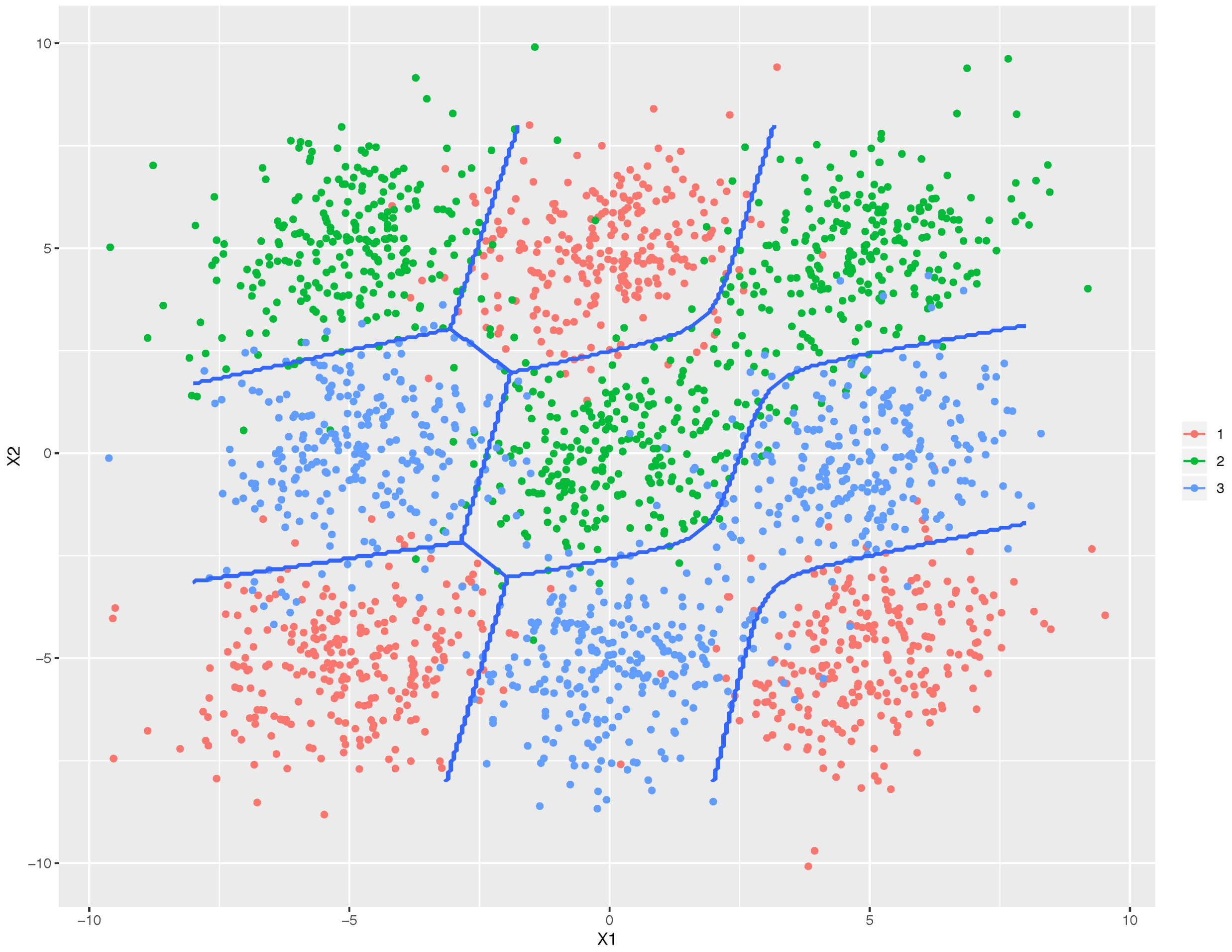

Figure 2: plot with coloured datapoints separated by boundary lines

An extension of LDA can be used to find the boundaries between intermingled gorups of data. Source: Yang Xiaozhou, https://yangxiaozhou.github.io/data/2019/10/02/linear-discriminant-analysis.html.

What is it used for?

A classic use case is for predicting bankruptcy v financial survival - Altman's model (1968) used LDA to derive a score for likelihood of business bankruptcy from only 4 or 5 predictors (business functions). It was trained on a small dataset (66 companies that went bankrupt) and was pretty accurate (but is no longer used today).

LDA is also used in facial recognition programs, to identify the main features by which to distinguish between faces.

Advantages and pitfalls

- LDA can be used with small sample sizes but is sensitive to outliers
- LDA is a very robust modelling technique *if* its assumptions are met. However, it relies on several assumptions that are often not met. Assumptions include:
  - predictor variables are normally distributed
  - homoskedasticity - same level of variance across all predictors
  - multicolinearity - if two or more predictors are correlated, LDA's predictive power is weakened
  - all observations are independent from each other (i.e. no paired data, no people before/after groups)
  - must have sample size greater than the number of variables

References

- https://en.wikipedia.org/wiki/Linear_discriminant_analysis

- https://yangxiaozhou.github.io/data/2019/10/02/linear-discriminant-analysis.html
- https://www.analyticsvidhya.com/blog/2021/08/a-brief-introduction-to-linear-discriminant-analysis/
- https://en.wikipedia.org/wiki/Altman_Z-score