27 Linear Discriminant Analysis: Pushing Classes Apart

27.1 Introduction: From Chaos to Clarity

27.1.1 What is LDA?

Linear Discriminant Analysis (LDA) seeks a projection direction that best separates different classes. It finds a line (or, in higher dimensions, a projection subspace) onto which data points are mapped so that sample points from the same class cluster tightly together, while different classes lie as far apart as possible. In this sense, LDA performs dimensionality reduction and classification at the same time.

Unlike unsupervised methods such as Principal Component Analysis (PCA), LDA makes explicit use of class labels, which often leads to better class separation when labels are available. LDA assumes that each class follows a Gaussian distribution and that all classes share the same covariance matrix. Under this shared-covariance assumption, LDA produces a linear decision boundary, making the model easy to interpret and computationally efficient. However, when the true data distribution deviates strongly from Gaussian assumptions or when the classes are not linearly separable, LDA may perform sub-optimally.

In fact, LDA can be viewed as a special case of Gaussian Discriminant Analysis (GDA). GDA allows each class to have its own covariance matrix, which leads to more flexible (possibly nonlinear) decision boundaries. LDA corresponds to the simplified case where all covariance matrices are identical, resulting in a linear boundary. In later chapters, we will develop GDA in detail and show how probabilistic modeling leads naturally to LDA as a constrained version.

27.1.2 Visual Intuition

In Figure 1, the blue and red samples overlap in 2D space. LDA tries to find a projection line (represented by a unit vector $\boldsymbol{v}$) such that, after projecting all points onto this line, the two classes are well separated.

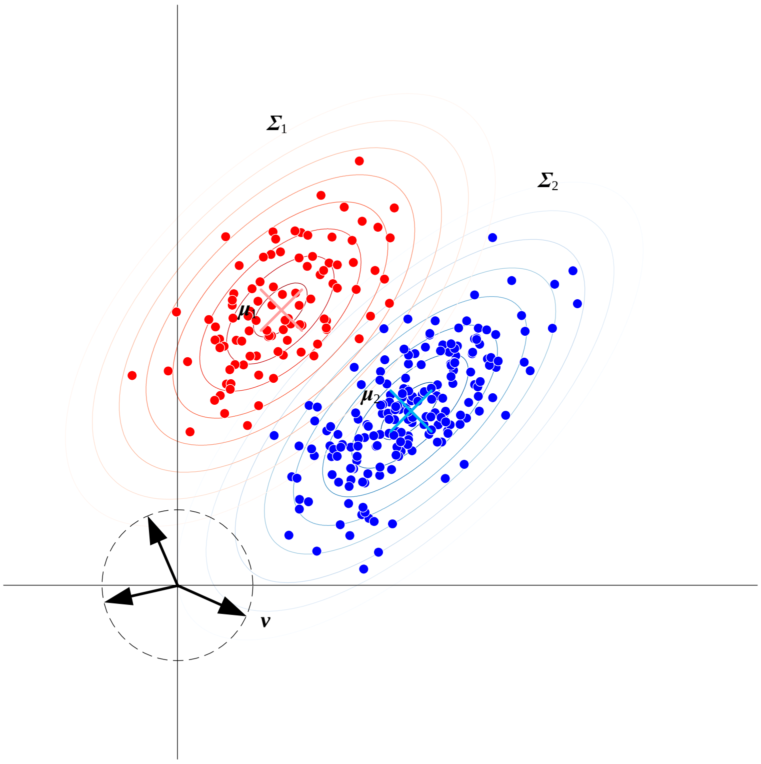

Figure 1. Two Classes Mixed in the Original Space

LDA measures "good separation" using two intuitive goals: (1) the projected class means should be as far apart as possible (large between-class distance), and (2) samples within the same class should lie close together on the line (small within-class variance).

If the projection direction is poorly chosen, as in Figure 2, the projected points from both classes overlap heavily and are difficult to classify.

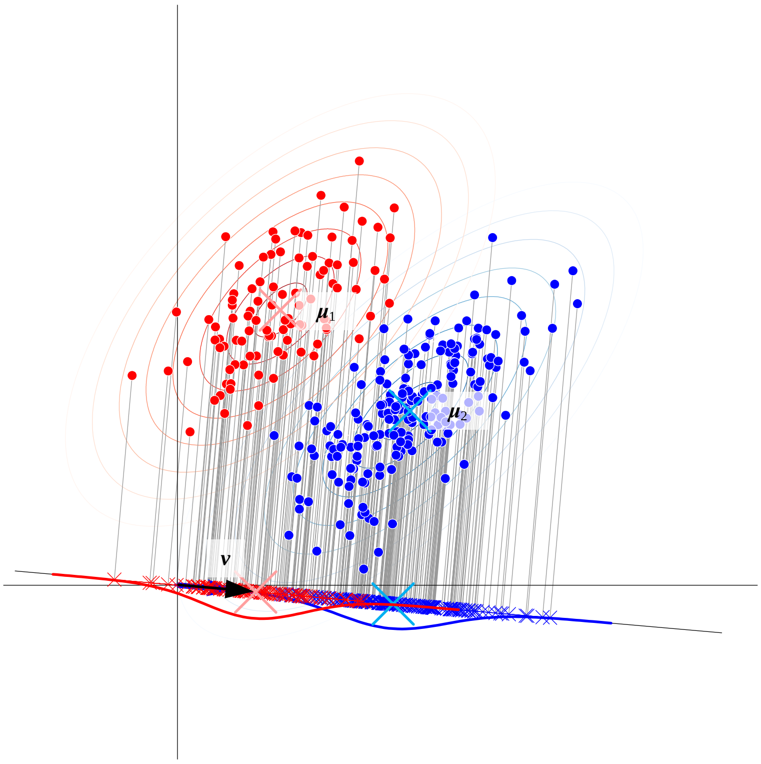

Figure 2. A Poor Projection Direction: Heavy Overlap After Projection

In contrast, the direction in Figure 3 separates the class means more clearly and produces more compact clusters after projection. In simple terms, LDA looks for a direction that pushes different classes apart and squeezes same-class points together.

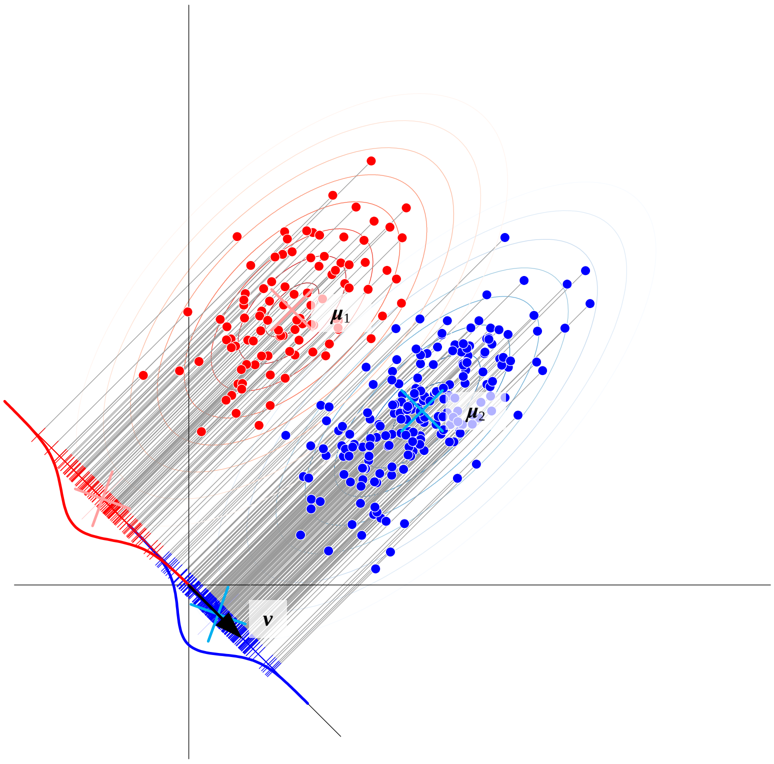

Figure 3. A Good Projection Direction: Large Class Separation and Small Within-Class Spread

27.2 The Math Behind LDA: Finding the Perfect Line

27.2.1 Class-Conditional Densities

Linear Discriminant Analysis works by finding a projection direction that maps high-dimensional data into a lower-dimensional space while keeping different classes as distinguishable as possible. The goal is to maximize the distance between class centers after projection and minimize the spread of samples within each class. This creates the best separation along a single line, which makes classification easier.

To achieve this, LDA assumes that each class follows a Gaussian distribution. As illustrated in Figure 4, the class-conditional densities $f_{Y,\chi}(C_1, \boldsymbol{x})$ and $f_{Y,\chi}(C_2, \boldsymbol{x})$ can be described by four sets of parameters: the means $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$, and the covariance matrices $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$.

Under this model, LDA seeks a decision boundary that separates the two Gaussian distributions, leading to a linear classifier.

The boundary (black line in Figure 4) can be expressed in the form

$$\boldsymbol{v}^{\mathrm{T}}\boldsymbol{x} + b = 0 \tag{1}$$

where $\boldsymbol{v}$ is a unit normal vector that determines the orientation of the boundary.

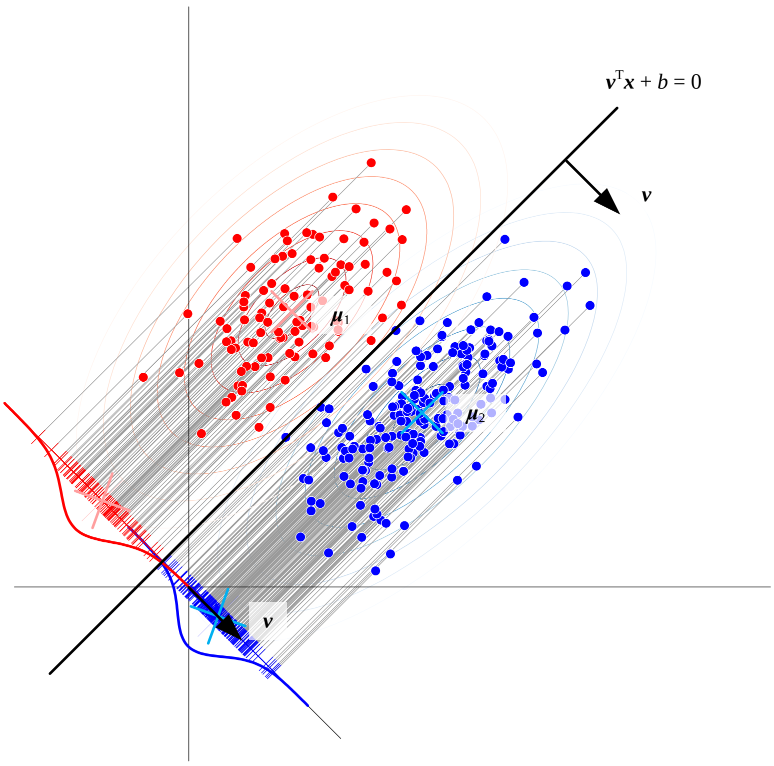

Figure 4. Linear Discriminant Boundary Between Two Gaussian Classes

27.2.2 Projection of Means and Variances

To understand the learned direction $\boldsymbol{v}$, we project the class means and covariance matrices onto $\boldsymbol{v}$.

Projecting $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ gives

$$\mu_{1,\boldsymbol{v}} = \boldsymbol{v}^{\mathrm{T}}\boldsymbol{\mu}_1, \quad \mu_{2,\boldsymbol{v}} = \boldsymbol{v}^{\mathrm{T}}\boldsymbol{\mu}_2 \tag{2}$$

which represent the class centers on the projection line.

The projected variances of the two classes are

$$\sigma_{1,\boldsymbol{v}}^2 = \boldsymbol{v}^{\mathrm{T}}\boldsymbol{\Sigma}_1\boldsymbol{v}, \quad \sigma_{2,\boldsymbol{v}}^2 = \boldsymbol{v}^{\mathrm{T}}\boldsymbol{\Sigma}_2\boldsymbol{v} \tag{3}$$

27.2.3 The Objective Function

From the projection viewpoint, LDA maximizes the following objective:

$$\arg\max_{\boldsymbol{v}} \frac{\left(\mu_{1,\boldsymbol{v}} - \mu_{2,\boldsymbol{v}}\right)^2}{\sigma_{1,\boldsymbol{v}}^2 + \sigma_{2,\boldsymbol{v}}^2} \tag{4}$$

Rewrite above and we get

$$\arg\max_{\boldsymbol{v}} \frac{\boldsymbol{v}^{\mathrm{T}}\left(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\right)\left(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\right)^{\mathrm{T}}\boldsymbol{v}}{\boldsymbol{v}^{\mathrm{T}}\left(\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2\right)\boldsymbol{v}} \tag{5}$$

where the numerator measures between-class separation and the denominator measures within-class compactness. This criterion is a generalized Rayleigh quotient, a concept introduced in linear algebra to describe ratios of quadratic forms.

27.3 Intuition Meets Algebra: Rayleigh Quotient in Action

27.3.1 From Quadratic Forms to Projection

As introduced earlier in this book, the Rayleigh quotient is expressed as a ratio of two quadratic forms, with both the numerator and the denominator being matrix quadratic forms

$$R\left(\boldsymbol{x}\right) = \frac{\boldsymbol{x}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{x}}{\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x}} \tag{6}$$

And the generalized Rayleigh quotient extends this to

$$R(\boldsymbol{x}) = \frac{\boldsymbol{x}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{x}}{\boldsymbol{x}^{\mathrm{T}}\boldsymbol{B}\boldsymbol{x}} \tag{7}$$

where $\boldsymbol{A}$ is symmetric and $\boldsymbol{B}$ is positive definite.

In LDA, this form naturally arises because we want to compare two quadratic quantities: the class-separation term versus the within-class variance term.

For a 2-D example, the generalized Rayleigh quotient forms a curved surface. Figure 5 shows its 3-D shape and contour plot. The maximum of this surface corresponds to the best projection direction.

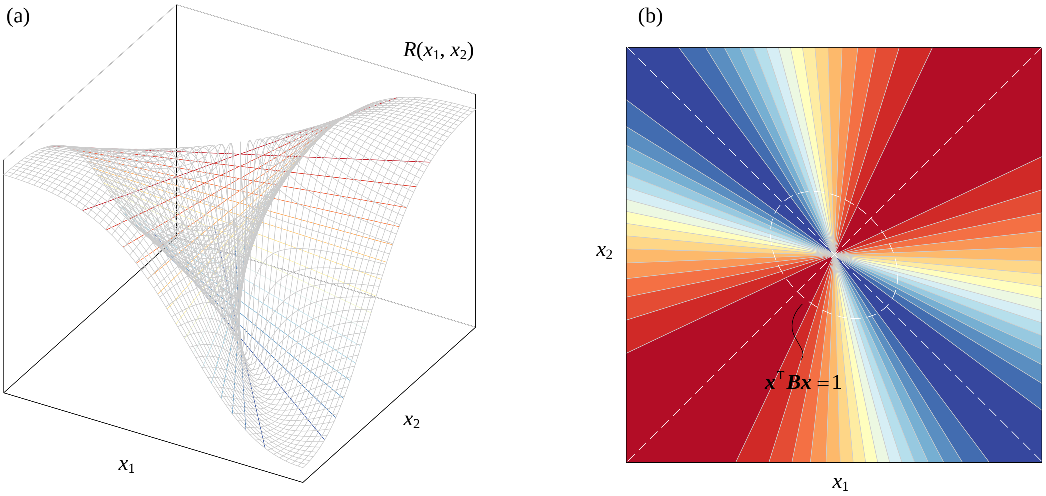

Figure 5. Surface and Contour of a 2-D Generalized Rayleigh Quotient

27.3.2 Optimal Projection Direction

Since a 2-D direction vector $\boldsymbol{v}$ can be written as

$$\boldsymbol{v} = \begin{bmatrix} \cos\theta \\ \sin\theta \end{bmatrix} \tag{8}$$

the objective function of (5) becomes a function of θ.

As shown in Figure 6, this function reaches two symmetric maxima 180 degrees apart. These two directions are equivalent, as reversing the direction only flips the projection without changing class separation.

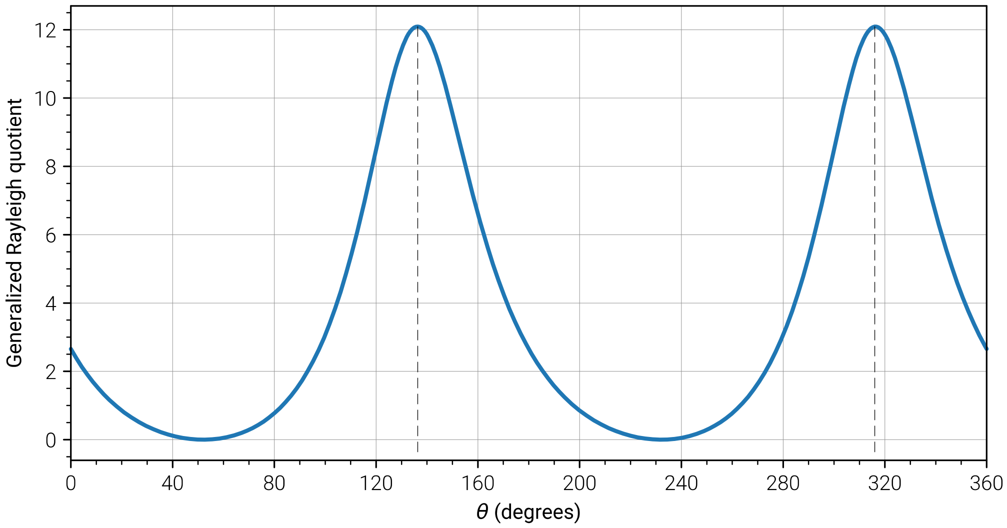

Figure 6. The LDA Objective as a Function of θ

27.4 Conclusion

Linear Discriminant Analysis (LDA) is a supervised method that simultaneously reduces dimensionality and enhances class separability. Its main goal is to find a projection direction that maps high-dimensional data onto a line or subspace, such that samples from the same class cluster tightly while different classes are spread far apart. Unlike unsupervised methods like PCA, LDA uses class labels, which improves separation when label information is available.

LDA assumes that each class follows a Gaussian distribution with a shared covariance matrix, producing a linear decision boundary that is computationally efficient and easy to interpret. Conceptually, LDA can be seen as a special case of Gaussian Discriminant Analysis (GDA), where identical covariance matrices simplify the decision boundary to a linear form.

The algorithm works by projecting class means and variances onto a direction vector $\boldsymbol{v}$, then maximizing the ratio of between-class variance to within-class variance—a generalized Rayleigh quotient. This ratio ensures that classes are pushed apart while samples within each class remain compact.