Linear Methods Of Classification

Discriminant Analysis

Sayan Dewanjee

Indian Statistical Institute

July 2025

Table of Contents

Table of Contents

Setting up the Framework

We have feature vector $X \in \mathbb{R}^p$.
And a class label $\mathbb{G} \in \{1, 2, ..., K\}$.
$\hat{G}(x)$ is the decision rule when we observed $X = x$

Zero-One Loss Function

Loss Function $L(k|l)$ is the price paid for classifying an observation to class k where true class is l.

$$L(k \mid l) = \begin{cases} 0 & \text{if } k = l \\ 1 & \text{if } k \neq l \end{cases}$$

This is zero-one loss function.

Risk Function

For zero-one loss

$$\begin{aligned} Risk(k \mid x) &= \sum_{\ell=1}^{K} L(k \mid \ell) \cdot P(G = \ell \mid X = x) \\ &= 1 - P(G = k \mid X = x) \end{aligned}$$

Bayes Decision Rule

Choose that class which minimizes conditional risk.

$$\begin{aligned}\hat{G}(x) &= \arg\min_k Risk(k|x) \\ &= \arg\min_k \{1 - P(G = k \mid X = x)\} \\ &= \arg\max_k \{P(G = k \mid X = x)\}\end{aligned}$$

Optimal decision is to choose that class with maximum posterior probability.

Table of Contents

Linear Regression of Indicator Matrix

For classification, we use output vector y whose *ith* element is 1 if true class label is i else 0. Hence we have an indicator response matrix Y of dimension $N \times K$ for N training instances.

We fit linear regression to all columns of Y simultaneously.

$$\begin{aligned}\hat{Y} &= X(X^\top X)^{-1}X^\top Y \\ &= X\hat{B}\end{aligned}$$

Prediction

For new input vector x,

- Fitted output $f(\hat{x})^{\top} = (1, x^{\top})\hat{B}$
- $f(\hat{x})$ is a vector of order $K \times 1$.
- Find largest element of this vector. Let's say *jth* element is largest.
- Classify input vector to class j.

Problems

For number of classes $K \geq 3$ there are serious problems with this method.

- Masking
- Encoding class labels
- Bad estimates of Posterior Probabilities
- Assumption of Homoscedasticity and Linearity

Masking

We draw samples of size 50 each from $Uniform(1,2)$, $Uniform(3,4)$, $Uniform(5,6)$ with random noise as 50 observations from each Class A,B and C.We fitted regression with Indicator matrix as stated previously.

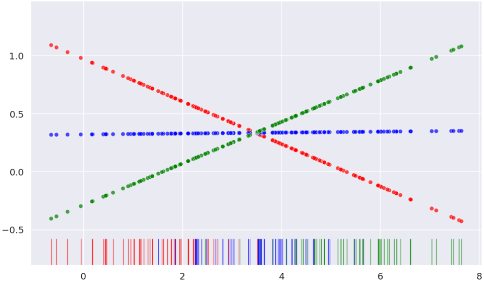


Figure: Predicted Value for each class with different colours

Table of Contents

Discriminant Analysis

- Consider we have two classes with label 1 and 2.
- $\pi_i = P(G = i)$ for $i = 1, 2$ is prior probability.
- $f_i(x)$ is the conditional probability density function of multivariate set of features x given it arises from π_i

Then,

$$
\begin{aligned}
P(G = 1 \mid x) &= \frac{P(x \cap G = 1)}{P(x)} \\
&= \frac{P(x \mid G = 1)P(G = 1)}{P(x \mid G = 1)P(G = 1) + P(x \mid G = 2)P(G = 2)} \\
&= \frac{\pi_1 f_1(x)}{\pi_1 f_1(x) + \pi_2 f_2(x)}
\end{aligned}
$$

Classification Rule

Classify an observation x to class label 1 if

$$P(G=1 \mid x) \geq P(G=2 \mid x)$$

$$implies, \frac{\pi_1 f_1(x)}{\pi_1 f_1(x)+\pi_2 f_2(x)} \geq \frac{\pi_2 f_2(x)}{\pi_1 f_1(x)+\pi_2 f_2(x)}$$

$$implies, \pi_1 f_1(x) \geq \pi_2 f_2(x)$$

Remark

If $\pi_1=\pi_2$ then classify an observation x to that population whose density is maximum i.e. classify to class label 1 if $f_1(x) \geq f_2(x)$. This is maximum likelihood problem.

LDA

Assume probability density function of x follows multivariate normal i.e. $f_i(x) \sim MVN(\mu_i, \Sigma_i)$.
Assuming $\Sigma_i = \Sigma \ \forall \ i = 1, 2$ and using logarithmic transformation(monotonic),

$$\begin{aligned}\delta(x) &= \log \frac{P(G = 1 \mid x)}{P(G = 2 \mid x)} \\ &= \log \frac{f_1(x)}{f_0(x)} + log \frac{\pi_1}{\pi_2} \\ &= \log \frac{\pi_1}{\pi_2} - \frac{1}{2}(\mu_1 + \mu_2)^\top \Sigma^{-1}(\mu_1 - \mu_2) + x^\top \Sigma^{-1}(\mu_1 - \mu_2)\end{aligned}$$

For a new observation x, classify it to class label 1 if $\delta(x) \geq 0$. Now parameters $\mu_1, \mu_2, \Sigma, \pi_1, \pi_2$ are unknown. Hence we have to estimate it from sample.

- μ_i is estimated with sample mean of all samples with label i
- Σ is estimated with pooled sample variance
- π_i is estimated with ratio of number of samples with class label i

As the decision boundary is linear in x we call it Linear Discriminant Analysis

Remark

Coefficients derived from regression is proportional to LDA. If both class has same sample size, then both rule will be identical.

Multi-Class Classification

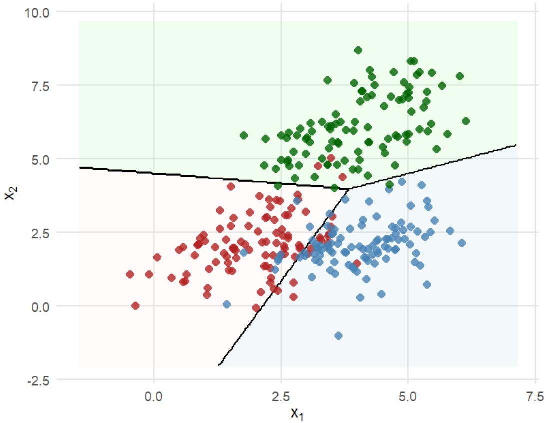
Consider we have K classes with density $f_i(x) \sim MVN(\mu_i, \Sigma)$.
Define Linear Score Function for class i

$$\begin{aligned}\delta_i(x) &= \log \pi_i f_i(x) \\ &= \log \pi_i - \frac{1}{2}\mu_i^\top \Sigma^{-1}\mu_i + \mu_i^\top \Sigma^{-1} x\end{aligned}$$

Classify an observation x to the class which has largest linear score.
We can estimate all parameters from samples here in similar way.

Visualisation

100 samples from 3 Bivariate normal with different means and equal covariance matrix - indicating 3 different classes.



QDA

Now consider we drop the assumptions of same variance - covariance structure of every class. We can compute Quadratic Score Functions in previous manner.

$$
\begin{aligned}
\delta_i(x) &= \log \pi_i f_i(x) \\
&= \log \pi_i - \frac{1}{2} \log |\Sigma_i| - \frac{1}{2}(x - \mu_i)^\top \Sigma_i^{-1}(x - \mu_i)
\end{aligned}
$$

Parameters μ_i, Σ_i have to estimate with sample mean and sample variance with class label i.
Classify an observation x to the class which has largest quadratic score function.

Visualisation

100 samples from 3 Bivariate normal with different means and different covariance matrix - indicating 3 different classes.

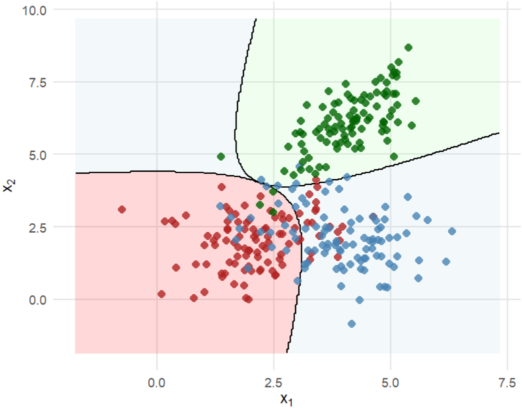


Table of Contents

Regularized Discriminant Analysis

Why RDA?

- Consider $N_k < p$ holds. Then Σ_k will be singular and can not be invertible.
- If $N_k < p$ holds, then all parameters will not be identifiable.
- QDA has low bias,LDA has high bias. QDA may overfit with more parameters. RDA balances this trade-off by shrinking the covariance estimate toward a shared structure.

Regularized Covariance Matrix

Regularized covariance matrix has the form :

$$\hat{\Sigma}_k(\alpha) = \alpha\hat{\Sigma}_k + (1-\alpha)\hat{\Sigma}$$

We can further regularized $\hat{\Sigma}$ with

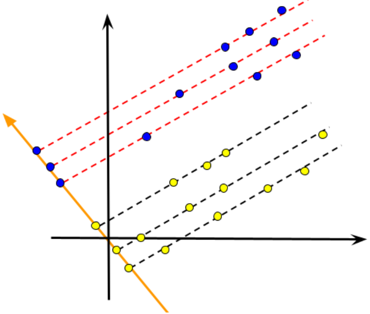
$$\hat{\Sigma}(\gamma) = \gamma\hat{\Sigma} + (1-\gamma)\hat{\sigma}^2 \mathrm{I}$$

Here, $\alpha, \gamma \in [0, 1]$.

- If $\alpha = 1$ then it will lead us to QDA.
- If $\alpha = 0$ and $\gamma = 1$, it will lead us to LDA.

LDA:Fisher's Approach

Consider we have $x_1, x_2, \ldots, x_k \in \mathbb{R}^2$. Now if we take projection of all those points over two parallel lines then projections onto those two lines will have same amount of seperation.



Now we have to find a best line or best direction which seperates all the class.
With mathematical notation,

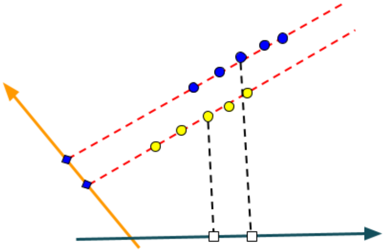
$$v_i = a^\top x_i$$

We have to find optimal a.

How do we quantify separation so that we can choose optimal direction a?

For binary classification, corresponding class-means after 1D projection should be different. Hence we can quantify separation with differrence of projected class-means.

But that will not work always. Consider the follwoing example:



Hence we have to choose optimal a such that both class has well-separated means and small class-variance.

Mathematical Notation

Consider $v_{11}, v_{12}, \ldots, v_{1n_1}$ are projections which has class labels 1 and $v_{21}, v_{22}, \ldots, v_{2n_2}$ are projections which has class labels 2.

μ_i is average mean of $v_{i1}, v_{i2}, \ldots, v_{in_i}$.

$S_i^2 = \sum_{j=1}^{n_i} (v_{ij} - \mu_i)^2$

Hence we want to find optimal a :

$$\max_{a:||a||=1} \frac{(\mu_1 - \mu_2)^2}{S_v^2}$$

where, $S_v^2 = \frac{1}{n_1+n_2-2}(S_1^2 + S_2^2)$

If m_1 and m_2 are class-means of original data of class 1 and class 2 respectively, we can write

$$\mu_1 = a^\top m_1$$
$$\mu_2 = a^\top m_2$$
$$S_v^2 = a^\top S_x^2 a$$

Then the optimization problem is :

$$\max_{a:||a||=1} \frac{(\mu_1 - \mu_2)^2}{S_v^2} = \max_{a:||a||=1} \frac{a^\top (m_1 - m_2)(m_1 - m_2)^\top a}{a^\top {S_x}^2 a}$$
$$= \max_{a:||a||=1} \frac{a^\top S_b a}{a^\top S_w a}$$

where, S_b is Between class scatter matrix and S_w is Within class scatter matrix.

The optimal direction will be :

$$\begin{aligned} a &= \text{largest eigenvector of } S_w^{-1} S_b \\ &= S_w^{-1}(m_1 - m_2) \end{aligned}$$

Decision rule

Classify observation$x$ to population 1 if
$(m_1 - m_2)^\top S_w^{-1} x \geq \frac{1}{2}(m_1 - m_2)^\top S_w^{-1}(m_1 + m_2)$

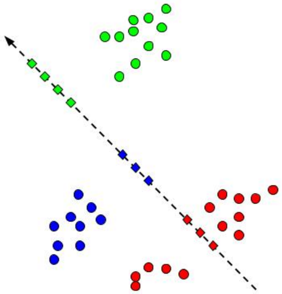
Note,

- If prior probability is same,then lda decision rule and above decision rule is same.
- Fisher's approach did not need the gaussian assumption.

Multi-class Classification

We can extend the same idea for multi-class classification. We will find optimal decision in such a way

- projected class is as tight as possible
- centroids are as far from each other as possible



With previous notation, we can define Within class scatter matrix S_w as follows :

$$\begin{aligned}\sum_j s_j^2 &= \sum_j \mathrm{a}^\top S_j \mathrm{a} \\ &= \mathrm{a}^\top \Big(\sum_j S_j\Big)\mathrm{a} \\ &= \mathrm{a}^\top S_w \mathrm{a}\end{aligned}$$

where,

$$S_j = \sum_{x \in \mathrm{C}_j} (\mathrm{x} - m_j)(\mathrm{x} - m_j)^\top$$

Define Between class scatter matrix with $\sum_{j=1}^{K} n_j(\mu_j - \mu)^2$ where

$$\mu = \frac{1}{n}\sum_{j=1}^{K} n_j\mu_j$$

We can write,

$$\begin{aligned}\sum_{j=1}^{K} n_j(\mu_j - \mu)^2 &= \sum_{j=1}^{K} n_j(v^\top m_j - v^\top m)^2 \\ &= v^\top \Big(\sum_{j=1}^{K} n_j(m_j - m)(m_j - m)^\top\Big)v \\ &= v^\top S_b v\end{aligned}$$

S_b is Between-class scatter matrix.

Hence again we have to find optimal direction a:

$$\max_{a:||a||=1} \frac{a^\top S_b a}{a^\top S_w a}$$

Optimal solution is :

$$a = \text{largest eigenvector of } S_w^{-1} S_b$$

Fisher's Approach as Dimensionality Reduction

Consider $v_1, v_2, ...v_s$ is non-zero eigenvectors of $S_w^{-1}S_b$. Here v_1 is the eigenvector corresponding to largest eigenvalue, v_2 is the eigenvector corresponding to second largest eigenvalue and so on.

- Note $v_1^\top x$ projects with most discriminating power, then $v_2^\top x$ projects with less discriminating power and so on.
- Assuming S_w is inverse, $s \leq c - 1$ holds. Hence we can get at most $c - 1$ discriminant direction.

Now consider the representation :

$$y = \begin{bmatrix} v_1^\top \\ v_2^\top \\ \vdots \\ v_s^\top \end{bmatrix} x$$

It is a transformation from $\mathbb{R}^p$ to $\mathbb{R}^s$. If s is much smaller than p, it will reduce dimension to a good extent.
We obtain a low-dimensional representation of data,that represents data as much as possible.

Decision Rule

Can we use this representation for classification?

Yes, we can modify our decision rule for Binary class appropriately.

- Centroids of all classes in p-dimensional input space lies on an affine subspace of dimension $\leq C - 1$. This space is same as subspace spanned by $v_1, v_2, \ldots, v_s$.
- Like binary-class classification, for a new observation x we will project it in above-mentioned subspace.
- Find the distance between centroids and projection of x.
- Classify x to that class for which distance is minimum.

Visualisation - Iris Data

We considered Iris dataset for visualisation.

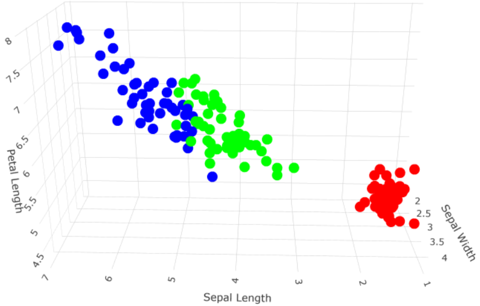
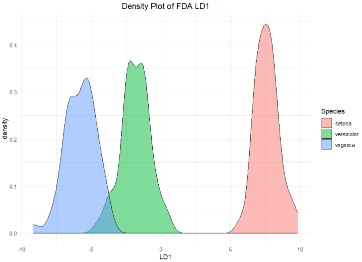
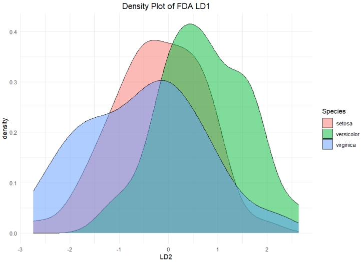


Figure: Red: Setosa, Green: Versicolor, Blue: Virginica

Discriminant Projection coordinates



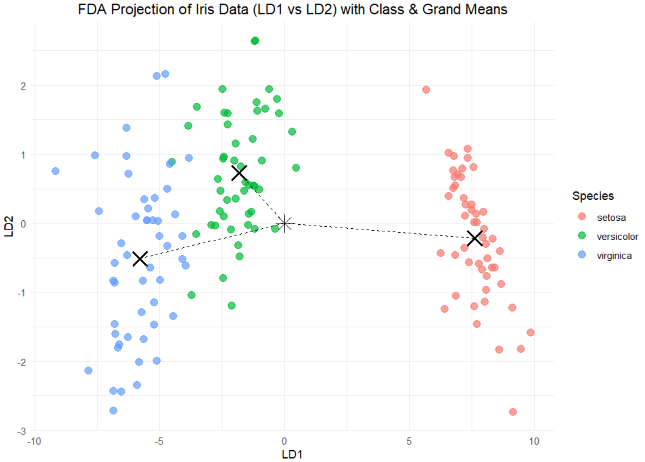
(a) First Discriminant Projection



(b) Second Discriminant Projection

Figure: Comparison of discriminant projections of Iris data

Projection on 2-d Plane



Principal Component Analysis

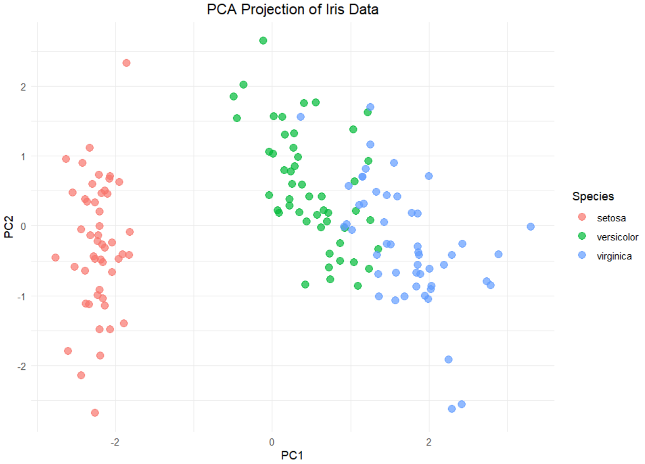


Table of Contents

When to Use What?

Choosing the Right Method for Linear Classification :

- **LDA (Linear Discriminant Analysis)**
  - Classes are Gaussian-distributed with **equal covariance**.
  - Number of observations per class is **moderate to large**.
  - You need a **simple and interpretable model**.
- **QDA (Quadratic Discriminant Analysis)**
  - Classes are Gaussian-distributed with **different covariances**.
  - You suspect **nonlinear boundaries** between classes.
  - Class-specific modeling is needed, and you have **enough data**.
- **FDA (Fisher's Discriminant Analysis)**
  - Goal is **dimensionality reduction** while preserving class separation.
  - Best when input dimension is high ($p \gg n$).

Reference

- The Elements of Statistical Learning
- Regularized Discriminant Analysis by Friedman(1989)
- Linear Discriminant Analysis

Thank You!

Questions or Feedback Welcome

sayandewanjee23@example.com
Indian Statistical Institute