# Week 7. Linear Discriminant Analysis (LDA) and logistic regression

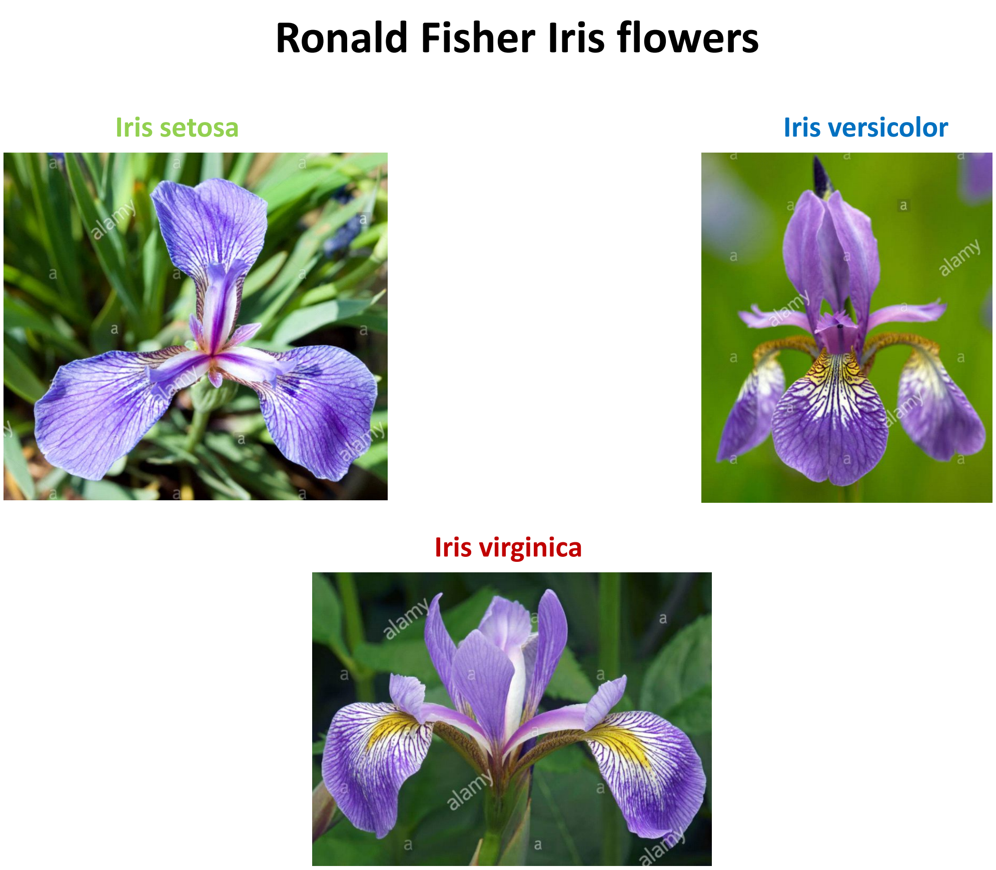
Previously, we learned:
(a) how to discriminate two populations using the ROC curve having a single feature/predictor
(b) how to compute the figure of merit of this discrimination using AUC
(c) how to find an optimal threshold.

What if we have several predictors? Answer: find the best linear combination of predictors and treat them as a univariate predictor.

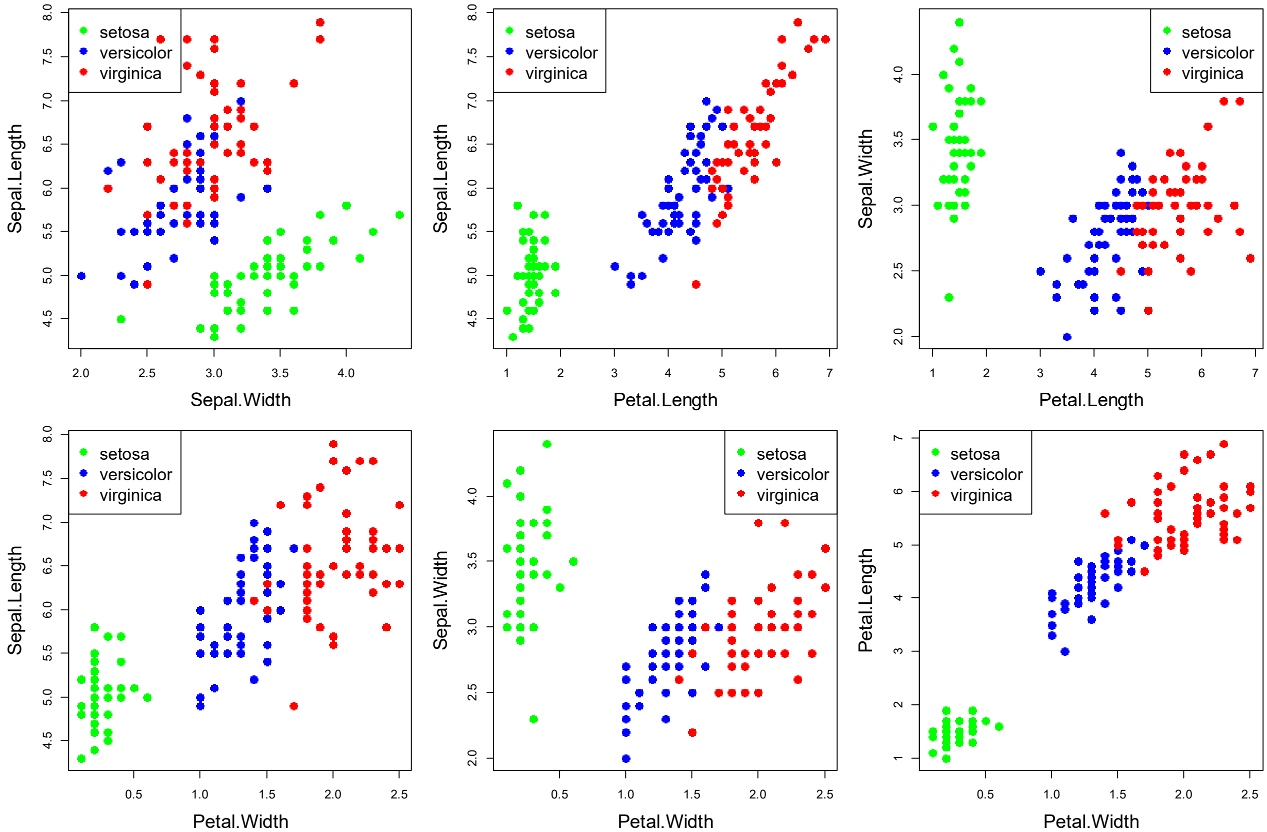
Several methods exist for the **supervised** binary classification: developing the linear rule for classification of the future observation:

1. Discriminant analysis (statistical model-based)
2. Logistic regression (statistical model-based)
3. SVM = Support Vector Machine (algorithm-based)

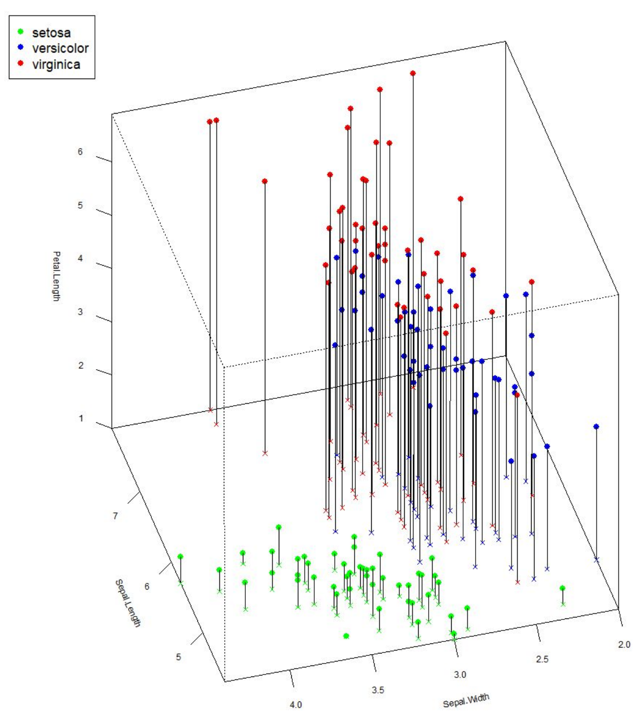
This week covers LDA.
Theorem 5.4
Ronald Fisher developed LDA.
Can we identify `virginica` having four measurements?
`iris.pptx`




Measuring petals and setals (botanical study)



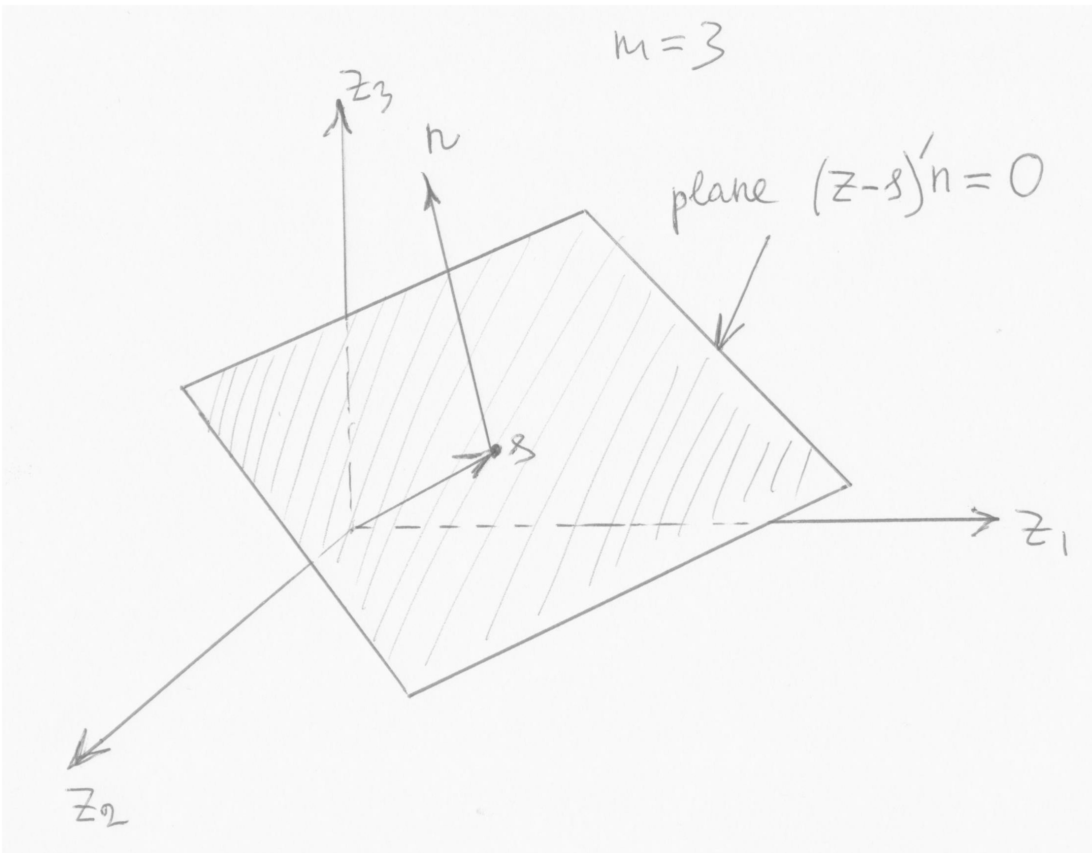

See the R function `iris3D`



**LDA problem set up.** There are two Gaussian multivariate populations $\mathcal{N}(\boldsymbol{\mu}_x, \boldsymbol{\Omega})$ and $\mathcal{N}(\boldsymbol{\mu}_y, \boldsymbol{\Omega})$ in $R^m$. Note, $\boldsymbol{\mu}_x \neq \boldsymbol{\mu}_y$, but the two populations share the same covariance matrix $\boldsymbol{\Omega}$. Given observation $\mathbf{z} \in R^m$, develop a linear discrimination rule: what population $\mathbf{z}$ belongs to?

First, we assume that $\boldsymbol{\mu}_x$, $\boldsymbol{\mu}_y$, and $\boldsymbol{\Omega}$ are known, and then apply to the case when we have data (estimate the means and the common covariance matrix).

Any linear discrimination rule is equivalently to finding by a plane $(\mathbf{z}-\mathbf{s})'\mathbf{n}$, where $\mathbf{s}$ is called the translation vector and $\mathbf{n}$ is the normal vector.




**Theorem 1** *The optimal linear discrimination rule is as follows:* $\mathbf{z}$ *belongs to population* $\mathbf{x}$ *if*

$$(\mathbf{z}-\mathbf{s})'\mathbf{n} > 0,$$

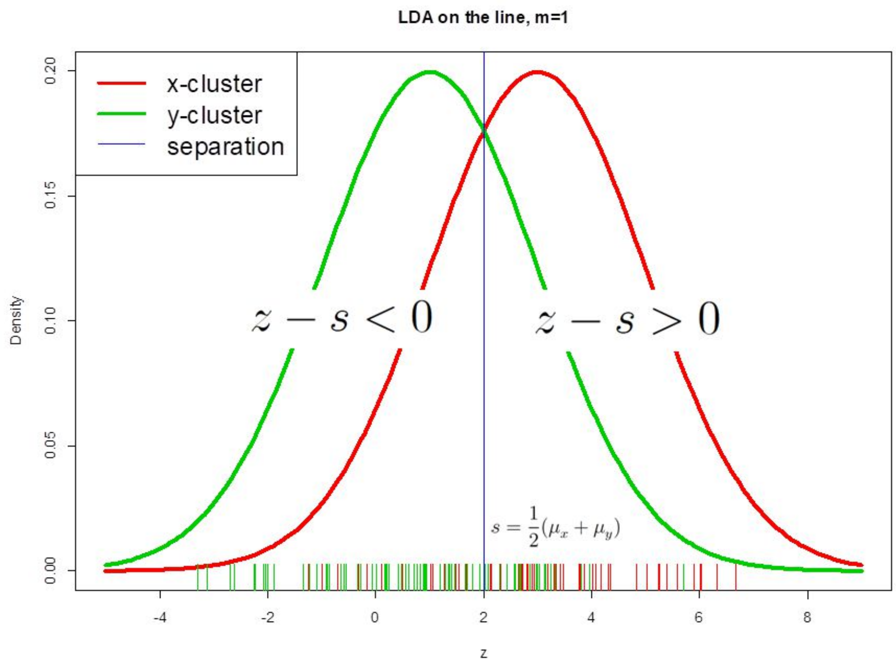
*and, otherwise, to population* $\mathbf{y}$, *where*

$$\mathbf{s} = \frac{1}{2}(\boldsymbol{\mu}_x + \boldsymbol{\mu}_y), \ \mathbf{n} = \boldsymbol{\Omega}^{-1}(\boldsymbol{\mu}_x - \boldsymbol{\mu}_y).$$

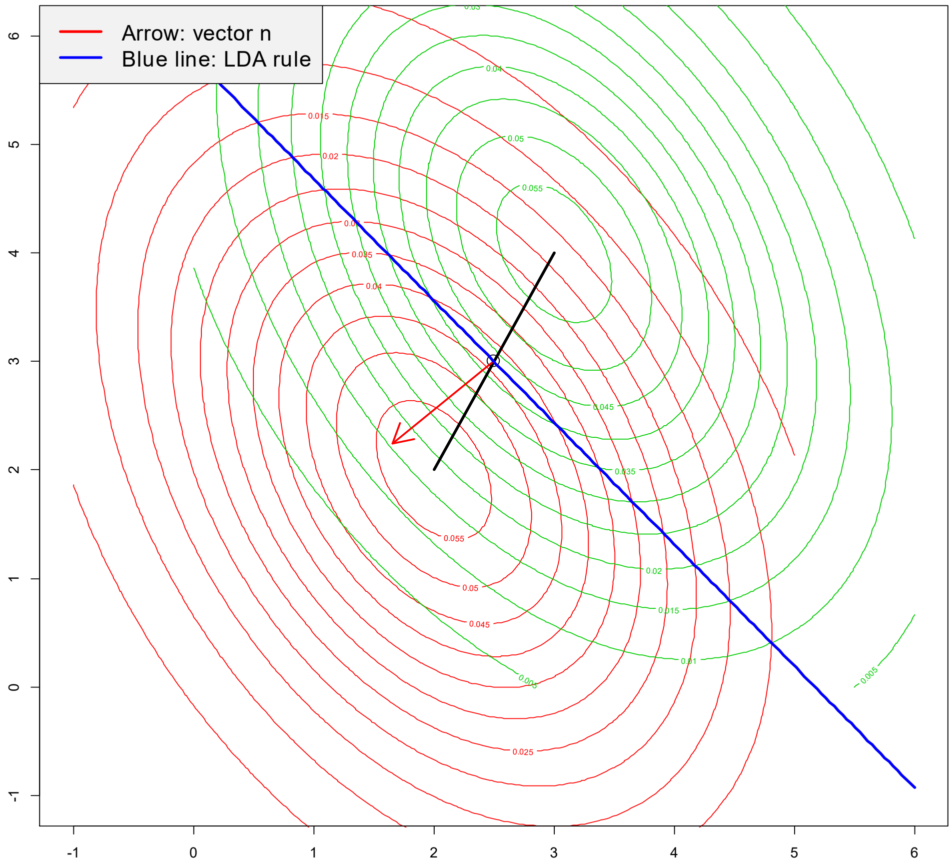
**Consider two cases:**

(a) $m = 1$: the point of separation is $s = (\mu_x + \mu_y)/2$

(b) $m = 2$ and $\mathbf{\Omega} = \sigma^2\mathbf{I}$: then $\mathbf{n}$ is parallel to $\boldsymbol{\mu}_x - \boldsymbol{\mu}_y$ and the separation line goes through $\mathbf{s} = (\boldsymbol{\mu}_x + \boldsymbol{\mu}_y)/2$ and orthogonal to $\boldsymbol{\mu}_x - \boldsymbol{\mu}_y$.



```
mah(job=0)
```




**Theorem 2** *The classification rule defined by the plane* $(\mathbf{z}-\mathbf{s})'\mathbf{n} = 0$ *minimizes the total sum of classification error.*

**Proof.** Let $\mathbf{z}$ be assigned to cluster $\mathbf{x}$ if $(\mathbf{z}-\mathbf{s})'\mathbf{n} > 0$ and to cluster $\mathbf{y}$ otherwise. The total classification error is

$$\begin{aligned} & \Pr\left((\mathbf{z}-\mathbf{s})'\mathbf{n} > 0 | \mathbf{z} \sim \mathcal{N}(\boldsymbol{\mu}_y, \boldsymbol{\Omega})\right) + \Pr\left((\mathbf{z}-\mathbf{s})'\mathbf{n} \leq 0 | \mathbf{z} \sim \mathcal{N}(\boldsymbol{\mu}_x, \boldsymbol{\Omega})\right) \\ = \; & 1 - \Phi\left(\frac{(\boldsymbol{\mu}_y - \mathbf{s})'\mathbf{n}}{\sqrt{\mathbf{n}'\boldsymbol{\Omega}\mathbf{n}}}\right) + \Phi\left(\frac{(\boldsymbol{\mu}_x - \mathbf{s})'\mathbf{n}}{\sqrt{\mathbf{n}'\boldsymbol{\Omega}\mathbf{n}}}\right). \end{aligned}$$

We want to find $\mathbf{s}$ and $\mathbf{n}$ such that the total error is minimum. Differentiating with respect to $\mathbf{s}$ we obtain

$$\frac{1}{\sqrt{\mathbf{n}'\boldsymbol{\Omega}\mathbf{n}}}\left(\phi\left(\frac{(\boldsymbol{\mu}_y - \mathbf{s})'\mathbf{n}}{\sqrt{\mathbf{n}'\boldsymbol{\Omega}\mathbf{n}}}\right) - \phi\left(\frac{(\boldsymbol{\mu}_x - \mathbf{s})'\mathbf{n}}{\sqrt{\mathbf{n}'\boldsymbol{\Omega}\mathbf{n}}}\right)\right)\mathbf{n} = \mathbf{0}$$

leading to $(\boldsymbol{\mu}_y - \mathbf{s})'\mathbf{n} = -(\boldsymbol{\mu}_x - \mathbf{s})'\mathbf{n}$. This implies a solution

$$\mathbf{s} = \frac{1}{2}(\boldsymbol{\mu}_x + \boldsymbol{\mu}_y).$$

Differentiation with respect to $\mathbf{n}$ gives $\mathbf{n} = \boldsymbol{\Omega}^{-1}(\boldsymbol{\mu}_x - \boldsymbol{\mu}_y)$.

**Probability of misclassification**

Misclassification: assign $\mathbf{z}$ to cluster $\mathbf{x}$, i.e. apply the rule $(\mathbf{z}-\mathbf{s})'\mathbf{n} > 0$ but in fact $\mathbf{z}$ belongs to cluster $\mathbf{y}$:

$$\Pr((\mathbf{z}-\mathbf{s})'\mathbf{n} > 0 | \mathbf{z} \sim \mathcal{N}(\boldsymbol{\mu}_y, \boldsymbol{\Omega})).$$

But

$$(\mathbf{z}-\mathbf{s})'\mathbf{n} \sim \mathcal{N}\left(-\frac{1}{2}\delta^2, \delta^2\right),$$

where

$$\delta^2 = (\boldsymbol{\mu}_x - \boldsymbol{\mu}_y)'\boldsymbol{\Omega}^{-1}(\boldsymbol{\mu}_x - \boldsymbol{\mu}_y).$$

**Definition 3** *The Mahalanobis distance between normal populations is defined as*

$$\delta = \sqrt{(\boldsymbol{\mu}_x - \boldsymbol{\mu}_y)'\boldsymbol{\Omega}^{-1}(\boldsymbol{\mu}_x - \boldsymbol{\mu}_y)}.$$

The probability of misclassifying a point from cluster $x$ to cluster $y$ is given by

$$\Phi\left(-\frac{1}{2}\delta\right).$$

Indeed, denote

$$Z = (\mathbf{z}-\mathbf{s})'\mathbf{n} \sim \mathcal{N}\left(-\frac{1}{2}\delta^2, \delta^2\right).$$

Then

$$\begin{aligned}\Pr(Z &> 0) = 1 - \Pr(Z < 0) = 1 - \Phi\left(\frac{0-\mu_Z}{\sigma_Z}\right) \\ &= 1 - \Phi\left(\frac{\frac{1}{2}\delta^2}{\delta}\right) = 1 - \Phi\left(\frac{1}{2}\delta\right) = \Phi\left(-\frac{1}{2}\delta\right).\end{aligned}$$

The same probability of misclassification of a point from cluster $y$ to assign to cluster $x$:

$$\Phi\left(-\frac{1}{2}\delta\right).$$

Thus the **total misclassification probability** is

$$2\Phi\left(-\frac{1}{2}\delta\right).$$

When you have data given by matrices $\mathbf{X}_1^{n_1 \times m}$ and $\mathbf{X}_2^{n_2 \times m}$, how to find all parameters? Estimate

$$\widehat{\boldsymbol{\mu}}_1 = \overline{\mathbf{x}}_1, \quad \widehat{\boldsymbol{\mu}}_2 = \overline{\mathbf{x}}_2$$

in R as `mu1=colMeans(X1)` and `mu2=colMeans(X2)`. The common/pooled covariance matrix is estimated as

$$\widehat{\boldsymbol{\Omega}} = \frac{1}{n_1+n_2-2}[(n_1-1)\mathrm{var}(\mathbf{X}_1) + (n_2-1)\mathrm{var}(\mathbf{X}_2)].$$