Optimal Decision Rule

When the two classes

- Are Gaussian
- Have the same covariance matrix

the optimal MAP rule is a linear classifier

We call it a linear classifier because the feature vector values are linearly combined as a weighted sum and compared to a threshold.

If we know the class densities and the prior probabilities, the solution is known

Geometric Interpretation of the Linear Classifier

The optimal MAP rule is to choose $l = 0$ if

$$\boxed{\underbrace{\underbrace{(\mathbf{m}_1-\mathbf{m}_0)^T}_{1\times p}\underbrace{\Sigma^{-1}}_{p\times p}}_{1\times p}}\underbrace{\mathbf{x}}_{p\times 1} < \boxed{(\mathbf{m}_1-\mathbf{m}_0)^T\Sigma^{-1}\frac{(\mathbf{m}_1+\mathbf{m}_0)}{2}+\log\frac{\pi_0}{\pi_1}}$$

A set of p weights

A scalar threshold

Note that if we know the prior probabilities and the class densities exactly, we know $\mathbf{m}_1, \mathbf{m}_0, \Sigma, \pi_0, \pi_1$, so that the weights and the threshold are known.

Geometry of the Linear Classifier

Define a function g as $g(\mathbf{x}) = \mathbf{w}^T \mathbf{x} + w_p$,

where $\mathbf{w}$ is a $(p \times 1)$ weight vector and w_p is the bias.

In terms of g, the decision rule is to choose $l = 0$ if $g(\mathbf{x}) < 0$.

This decision rule is linear because $g(\mathbf{x})$ is a linear combination

of the components of $\mathbf{x}$: $g(\mathbf{x}) = w_0 x_0 + w_1 x_1 + \cdots + w_{p-1} x_{p-1} + w_p$.

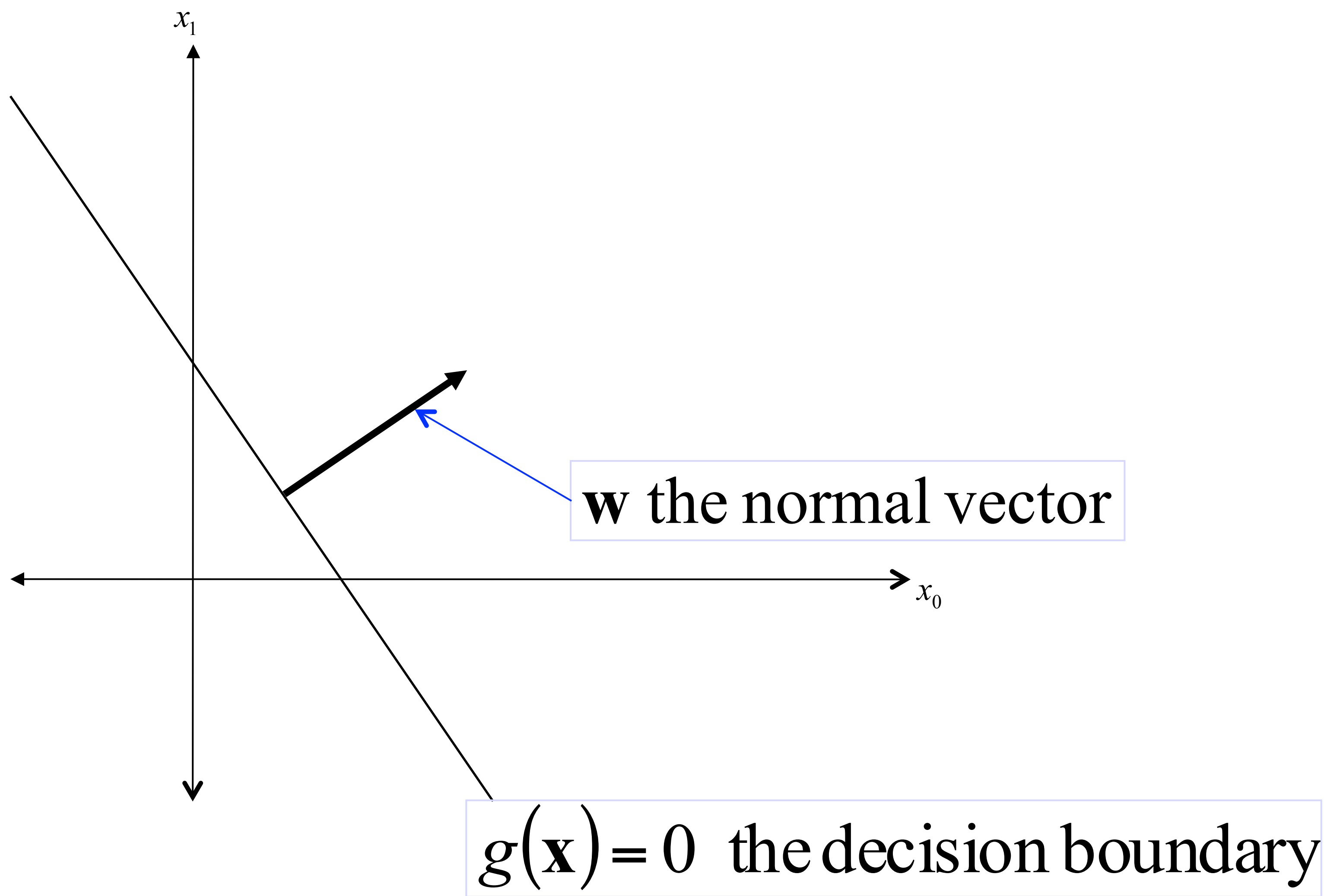
When $p = 2$, $g(\mathbf{x})$ is a line.

When $p = 3$, $g(\mathbf{x})$ is a plane.

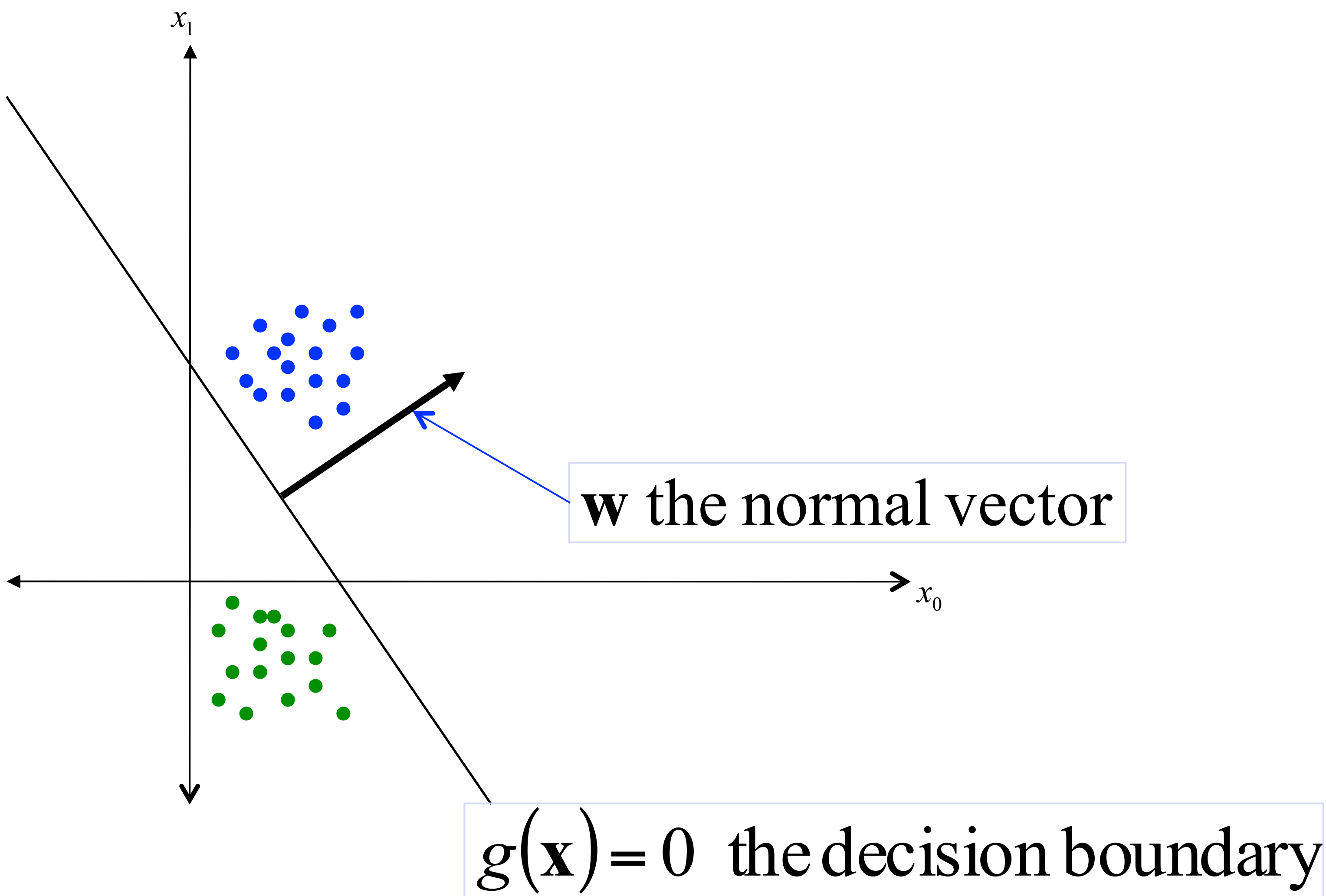
When $p > 3$, $g(\mathbf{x})$ is a hyperplane.

w is the normal vector

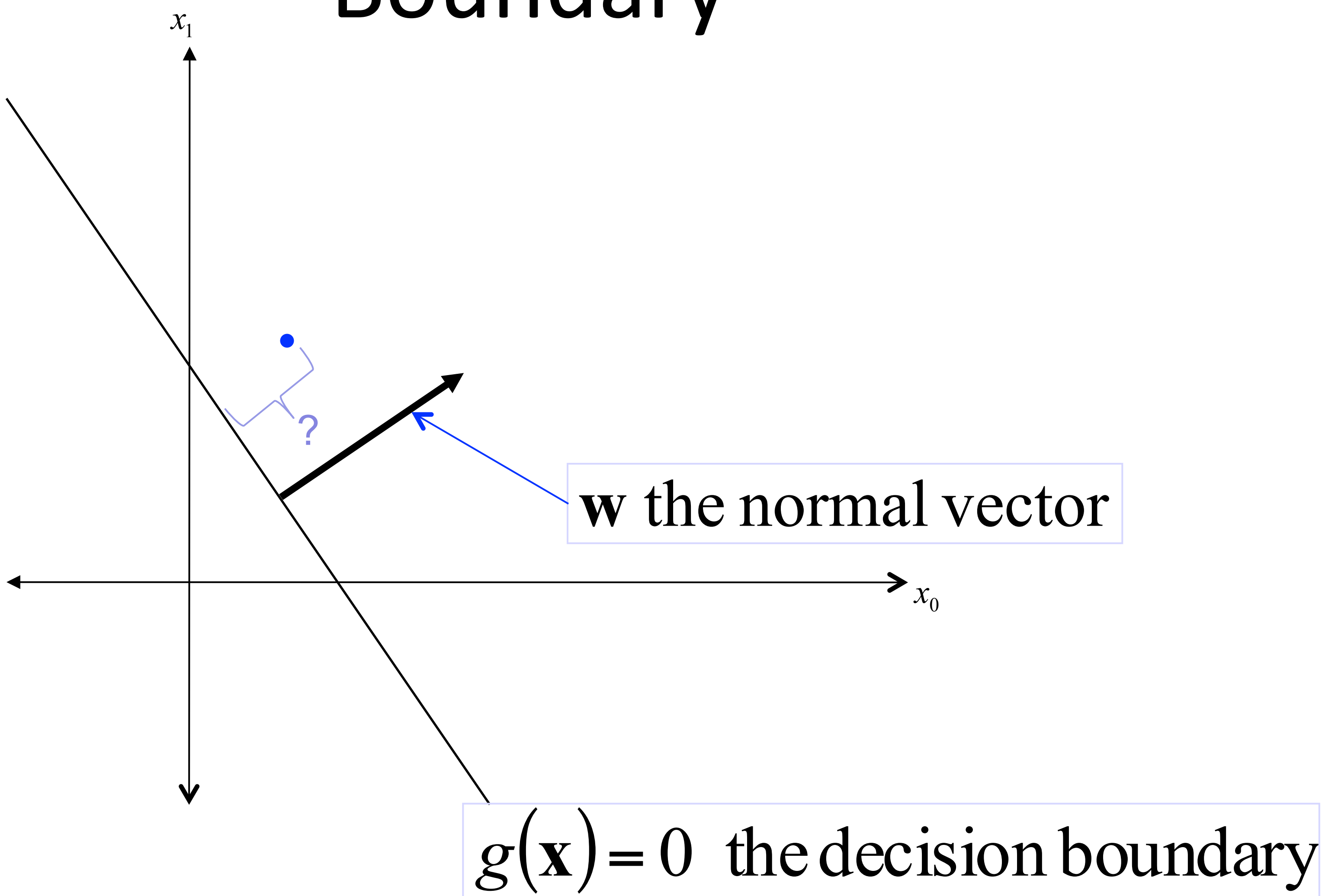
Linear Classifier



Linear Classifier



Distance of a Point to the Decision Boundary



Distance of a Point to the Decision Boundary

The decision boundary is $g(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + w_p = 0$.

Let $\mathbf{x}$ be an arbitrary point in the feature space and let δ be the distance of $\mathbf{x}$ from the decision boundary. What is δ?

Let $\mathbf{x}_\perp$ be the point on the decision boundary that is closest to $\mathbf{x}$,

so that $\mathbf{x} = \mathbf{x}_\perp + \delta \dfrac{\mathbf{w}}{\|\mathbf{w}\|}$. Perform the inner product of $\mathbf{w}$ with both sides

of the equation and add the bias term to both inner products. We have

$$\underbrace{\mathbf{w}^T\mathbf{x} + w_p}_{\boxed{g(\mathbf{x})}} = \underbrace{\mathbf{w}^T\mathbf{x}_\perp + w_p}_{\boxed{g(\mathbf{x}_\perp)}} + \delta \frac{\mathbf{w}^T\mathbf{w}}{\|\mathbf{w}\|}, \text{ so that } \delta = \frac{g(\mathbf{x})}{\|\mathbf{w}\|}.$$

$g(\mathbf{x}_\perp) = 0$ because by definition $\mathbf{x}_\perp$ sits on the decision boundary

Distance of a Point to the Decision Boundary

The decision boundary is $g(\mathbf{x}) = \mathbf{w}^T \mathbf{x} + w_p = 0.$

Let $\mathbf{x}$ be an arbitrary point in the feature space.

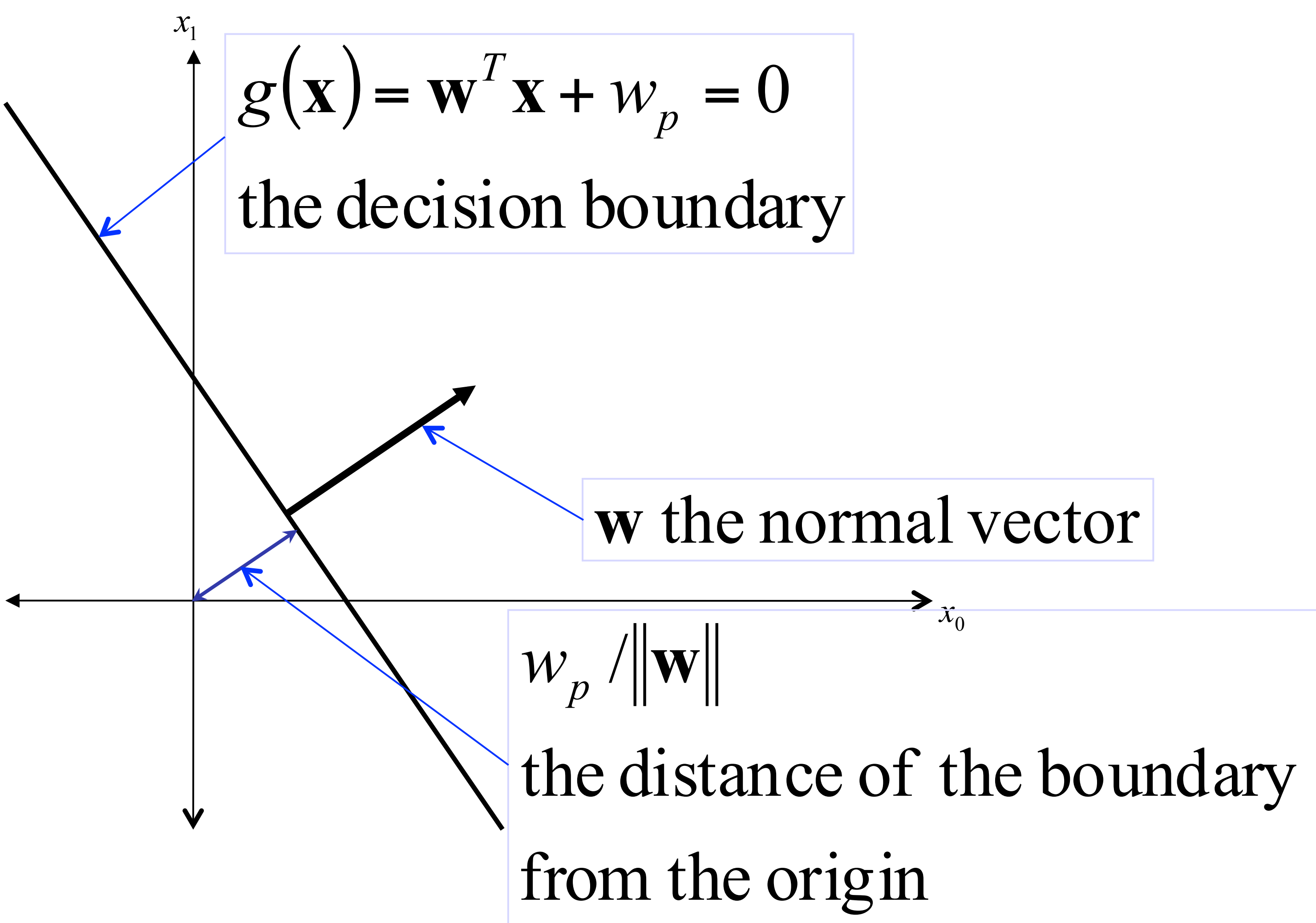
The distance of $\mathbf{x}$ from the decision boundary is $\delta(\mathbf{x}) = \dfrac{g(\mathbf{x})}{\|\mathbf{w}\|}.$

What is the distance of the decision boundary from the origin?

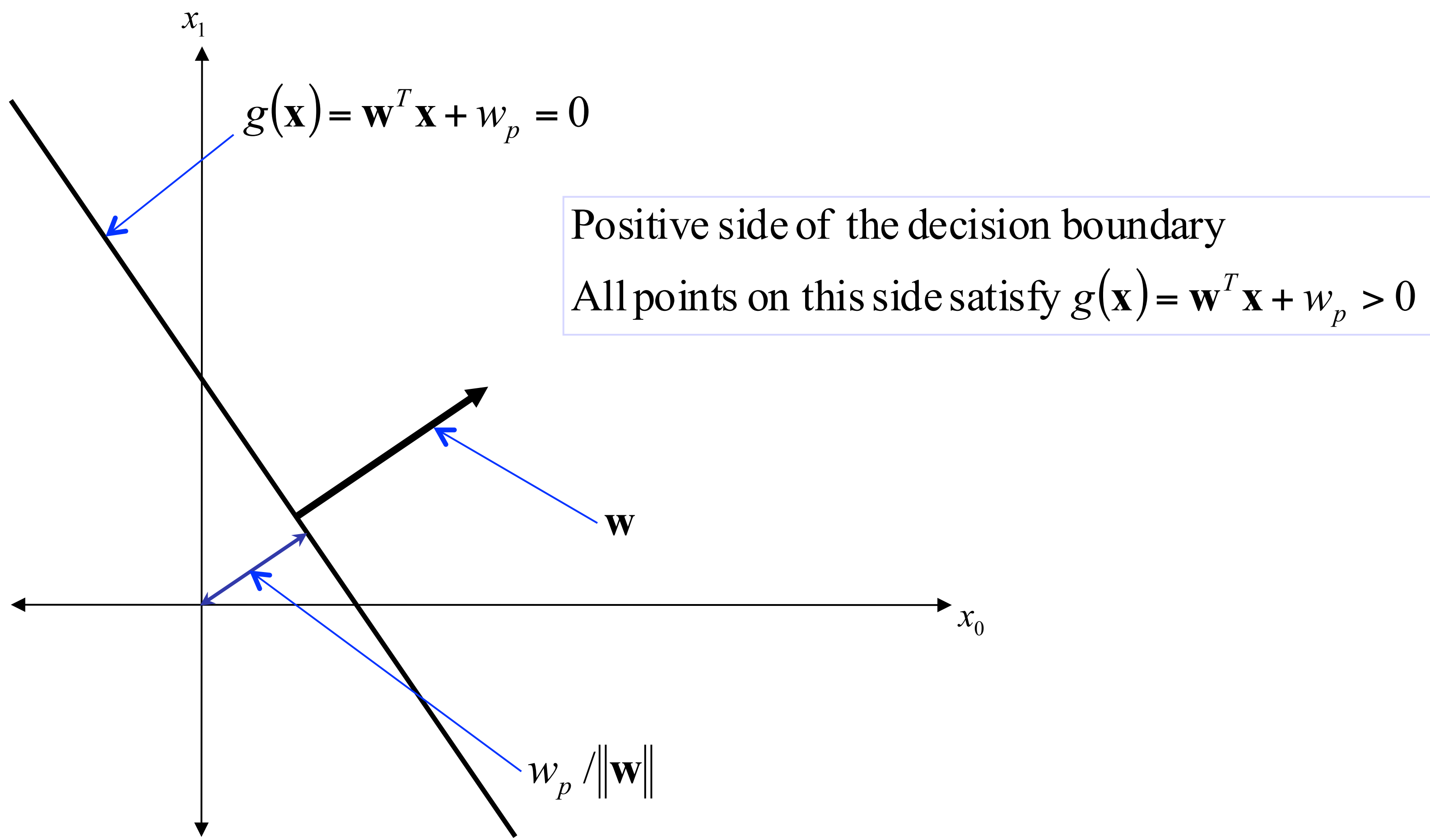
Let $\mathbf{x} = \mathbf{0}$, the origin. Then

$$\delta(\mathbf{0}) = \frac{g(\mathbf{0})}{\|\mathbf{w}\|} = \frac{\mathbf{w}^T \mathbf{0} + w_p}{\|\mathbf{w}\|} = \frac{w_p}{\|\mathbf{w}\|}.$$

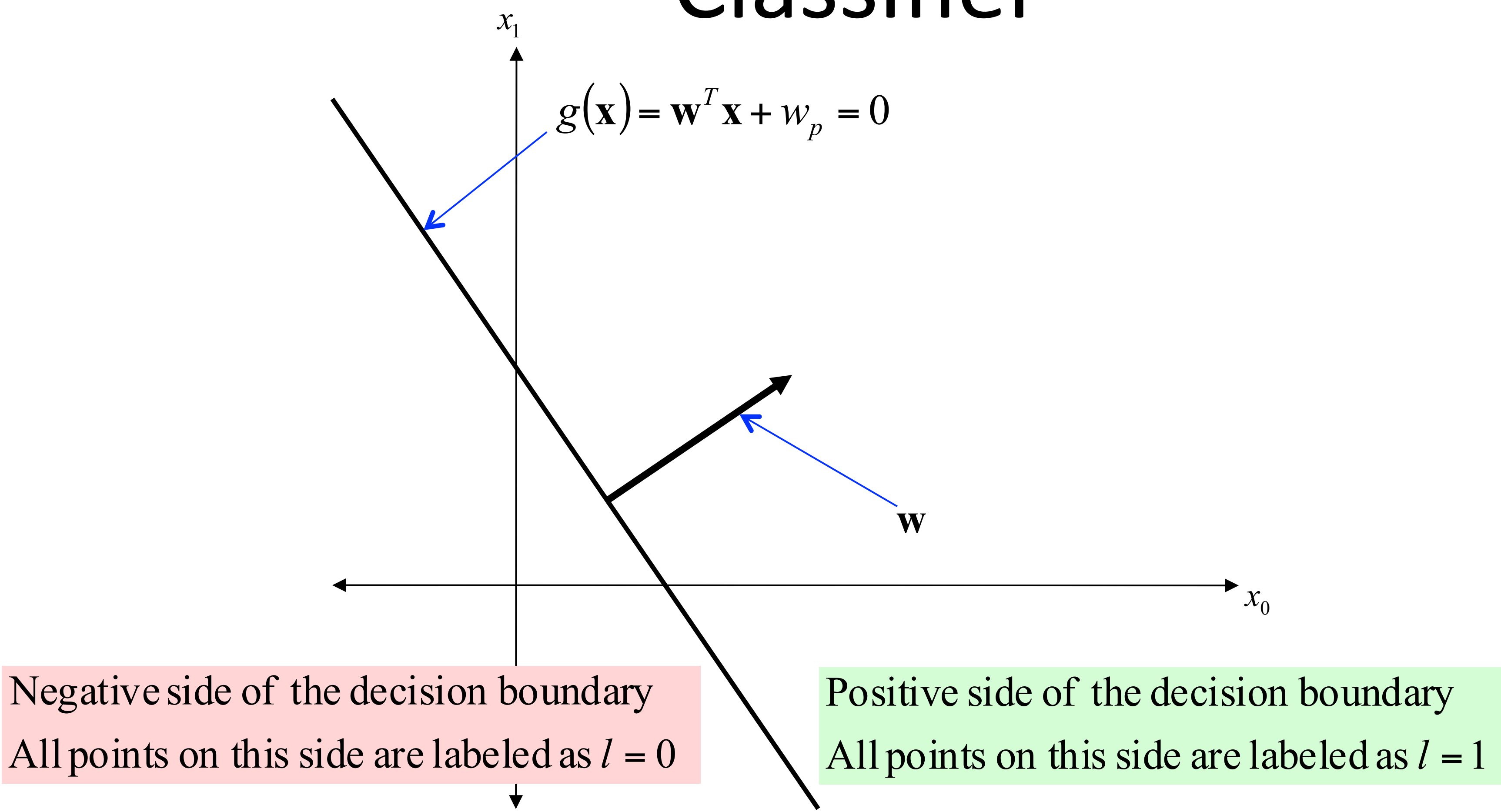
Linear Classifier



Linear Classifier



Decision Regions of a Linear Classifier



Linear Classifier

Suppose $K = 2$.

Let the class densities be Gaussian that differ only in their means $\mathbf{m}_1$ and $\mathbf{m}_0$; let the covariance matrix be $\boldsymbol{\Sigma}$.

Let $\mathbf{w} = \boldsymbol{\Sigma}^{-T}(\mathbf{m}_1 - \mathbf{m}_0)$ and $w_{\mathrm{p}} = -(\mathbf{m}_1 - \mathbf{m}_0)^T \boldsymbol{\Sigma}^{-1} \frac{(\mathbf{m}_1 + \mathbf{m}_0)}{2} + \log \frac{\pi_0}{\pi_1}$.

Define $g(\mathbf{x}) = \mathbf{w}^T \mathbf{x} + w_p$. The hyperplane $g(\mathbf{x}) = 0$ is the decision boundary.

The optimal decision rule is to choose $l = 0$ if $g(\mathbf{x}) < 0$ and $l = 1$ if $g(\mathbf{x}) > 0$.

$$g(\mathbf{x}) = \mathbf{w}^T \mathbf{x} + w_p$$

Controls the orientation of the decision boundary

Controls the placement of the decision boundary (*ie* how far is it away from the origin)

Classification when the priors are the same

Choose $l = k$ to minimize

$$\frac{1}{2}(\mathbf{x}-\mathbf{m}_k)^T \mathbf{\Sigma}^{-1}(\mathbf{x}-\mathbf{m}_k) - \log \pi_k .$$

When all classes have the same prior probability, then the decision is choose $l = k$ to minimize

$$\frac{1}{2}(\mathbf{x}-\mathbf{m}_k)^T \mathbf{\Sigma}^{-1}(\mathbf{x}-\mathbf{m}_k), \text{ or, equivalently, } (\mathbf{x}-\mathbf{m}_k)^T \mathbf{\Sigma}^{-1}(\mathbf{x}-\mathbf{m}_k).$$

Let $d_M^2(\mathbf{x},\mathbf{m}_k) = (\mathbf{x}-\mathbf{m}_k)^T \mathbf{\Sigma}^{-1}(\mathbf{x}-\mathbf{m}_k)$; the term $d_M(\mathbf{x},\mathbf{m}_k)$ is called the Mahalanobis distance between $\mathbf{x}$ and $\mathbf{m}_k$.

Classification when the priors are the same

Choose $l = k$ to minimize

$$\left(\mathbf{x} - \mathbf{m}_k\right)^T \mathbf{\Sigma}^{-1}\left(\mathbf{x} - \mathbf{m}_k\right),$$

the squared Mahalanobis distance between $\mathbf{x}$ and $\mathbf{m}_k$.

Mahalanobis Distance

The Mahalanobis distance $d_M(\mathbf{x}, \mathbf{y})$ between $\mathbf{x}$ and $\mathbf{y}$ is given by

$$d_M^2(\mathbf{x}, \mathbf{y}) = (\mathbf{x} - \mathbf{y})^T \mathbf{\Sigma}^{-1} (\mathbf{x} - \mathbf{y}).$$

When the components are pairwise uncorrelated so that

$$\mathbf{\Sigma} = \begin{bmatrix} \sigma_0^2 & & \\ & \ddots & \\ & & \sigma_{p-1}^2 \end{bmatrix},$$ the Mahalanobis distance is

$$d_M^2(\mathbf{x}, \mathbf{y}) = (\mathbf{x} - \mathbf{y})^T \mathbf{\Sigma}^{-1} (\mathbf{x} - \mathbf{y}) = (\mathbf{x} - \mathbf{y})^T \begin{bmatrix} \frac{1}{\sigma_0^2} & & \\ & \ddots & \\ & & \frac{1}{\sigma_{p-1}^2} \end{bmatrix} (\mathbf{x} - \mathbf{y}) = \sum_{i=0}^{p-1} \frac{(x_i - y_i)^2}{\sigma_i^2}.$$

Mahalanobis Distance

The Mahalanobis distance $d_M(\mathbf{x}, \mathbf{y})$ between $\mathbf{x}$ and $\mathbf{y}$ is given by

$$d_M^2(\mathbf{x}, \mathbf{y}) = (\mathbf{x} - \mathbf{y})^T \mathbf{\Sigma}^{-1} (\mathbf{x} - \mathbf{y}).$$

When the components are pairwise uncorrelated and each has unit variance, so that $\mathbf{\Sigma} = \mathbf{I}$, the Mahalanobis distance is the same as the Euclidean distance

$$d_M^2(\mathbf{x}, \mathbf{y}) = (\mathbf{x} - \mathbf{y})^T \mathbf{I}^{-1} (\mathbf{x} - \mathbf{y}) = (\mathbf{x} - \mathbf{y})^T (\mathbf{x} - \mathbf{y}) = \sum_{i=0}^{p-1} (x_i - y_i)^2.$$

Covariance Matrix

Let the covariance matrix be $\mathbf{\Sigma} = E[(\mathbf{x}-\mathbf{m})(\mathbf{x}-\mathbf{m})^T]$.

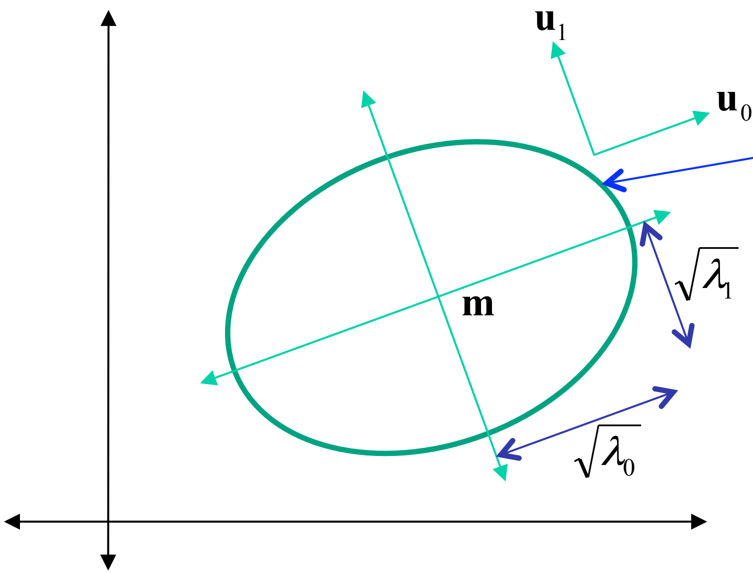
The ijth element is defined as $\sigma_{ij} = E\left[(x_i - m_i)(x_j - m_j)\right]$.

Since $\sigma_{ij} = \sigma_{ji}$, the covariance matrix $\mathbf{\Sigma}$ is symmetric.

Covariance Matrix of a Gaussian Density

The density of a Gaussian vector with mean $\mathbf{m}$ and covariance matrix $\boldsymbol{\Sigma}$ is

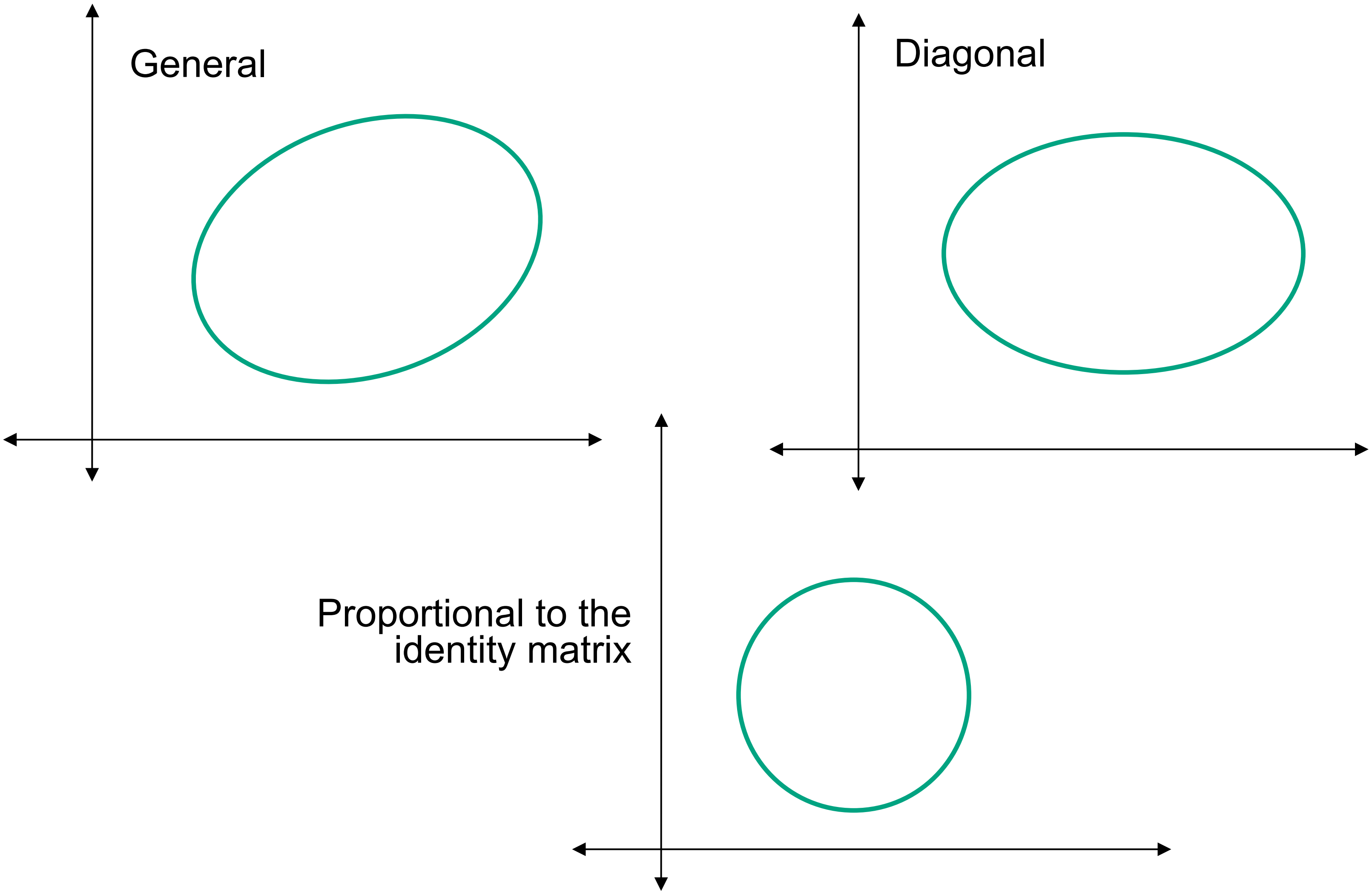
$$p(\mathbf{x}) = \frac{1}{(2\pi)^{p/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} e^{-\frac{1}{2}(\mathbf{x}-\mathbf{m})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\mathbf{m})}.$$



The surface of constant density on which $p(\mathbf{x})$ is $e^{-1/2}$ of $p(\mathbf{m})$

The covariance matrix $\boldsymbol{\Sigma}$ has eigenvectors $\mathbf{u}_0$ and $\mathbf{u}_1$ and corresponding eigenvalues λ_0 and λ_1.

Covariance Matrices



Examples

- Class densities are Gaussian that differ only in their means

Two classes

Class means:

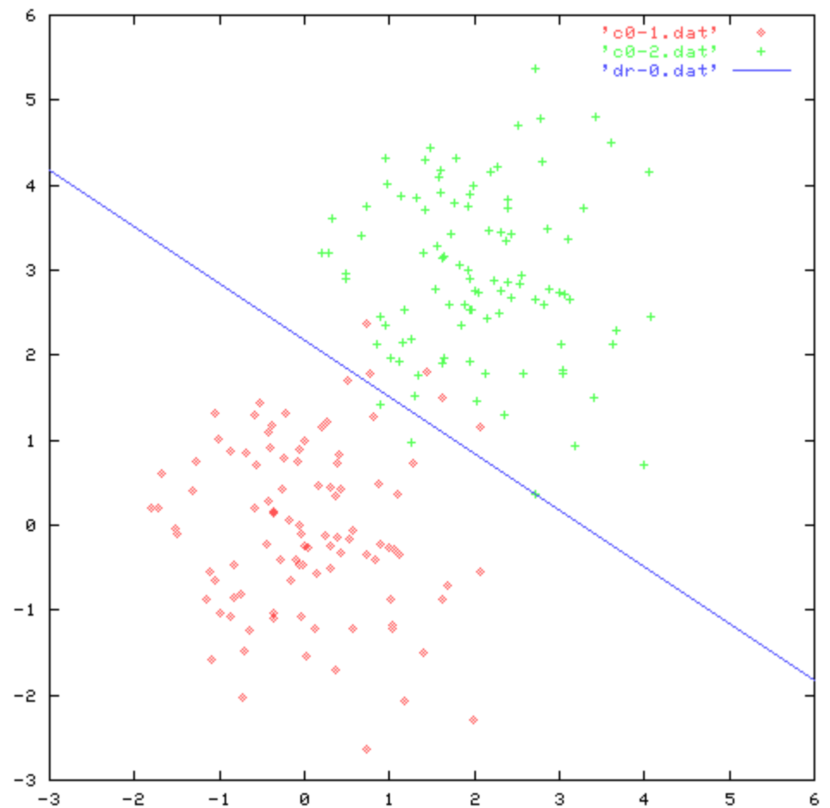
$$\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 2 \\ 3 \end{bmatrix}$$

$$\mathbf{\Sigma} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

$$\mathbf{\Sigma}^{-1} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

Decision boundary:

$$2x_0 + 3x_1 = 6.5$$

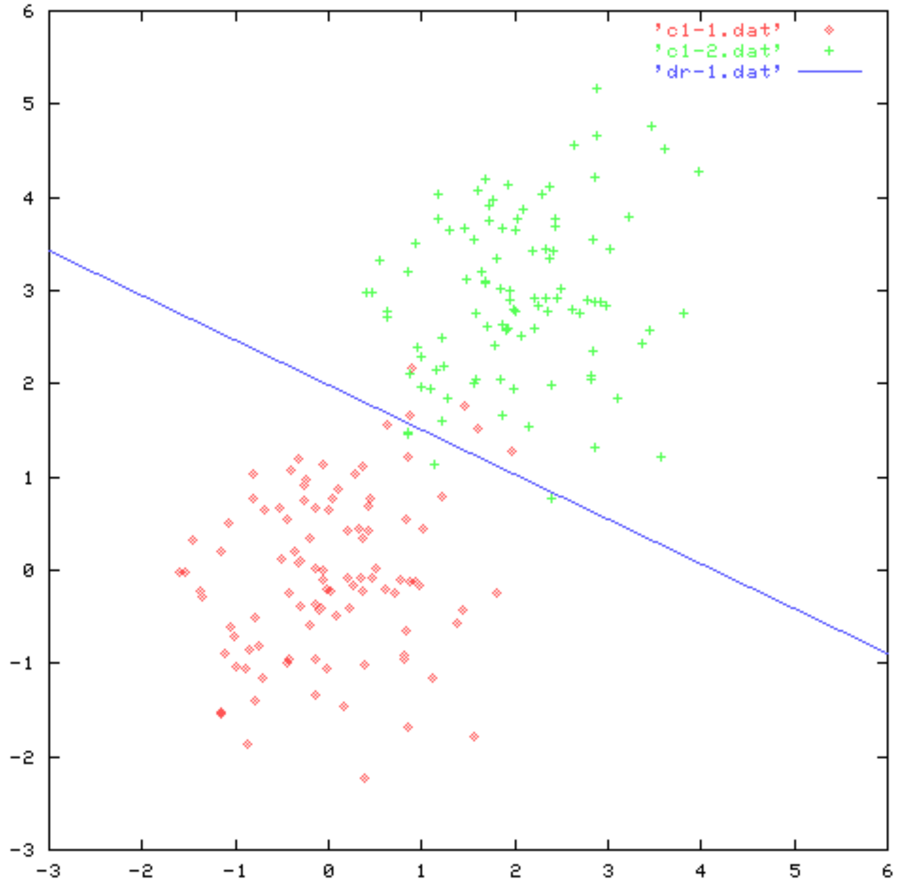


$$\Sigma = \begin{bmatrix} 0.82 & 0.20 \\ 0.20 & 0.79 \end{bmatrix}$$

$$\Sigma^{-1} = \begin{bmatrix} 1.30 & -0.32 \\ -0.32 & 1.35 \end{bmatrix}$$

Decision boundary:

$$1.63x_0 + 3.40x_1 = 6.73$$

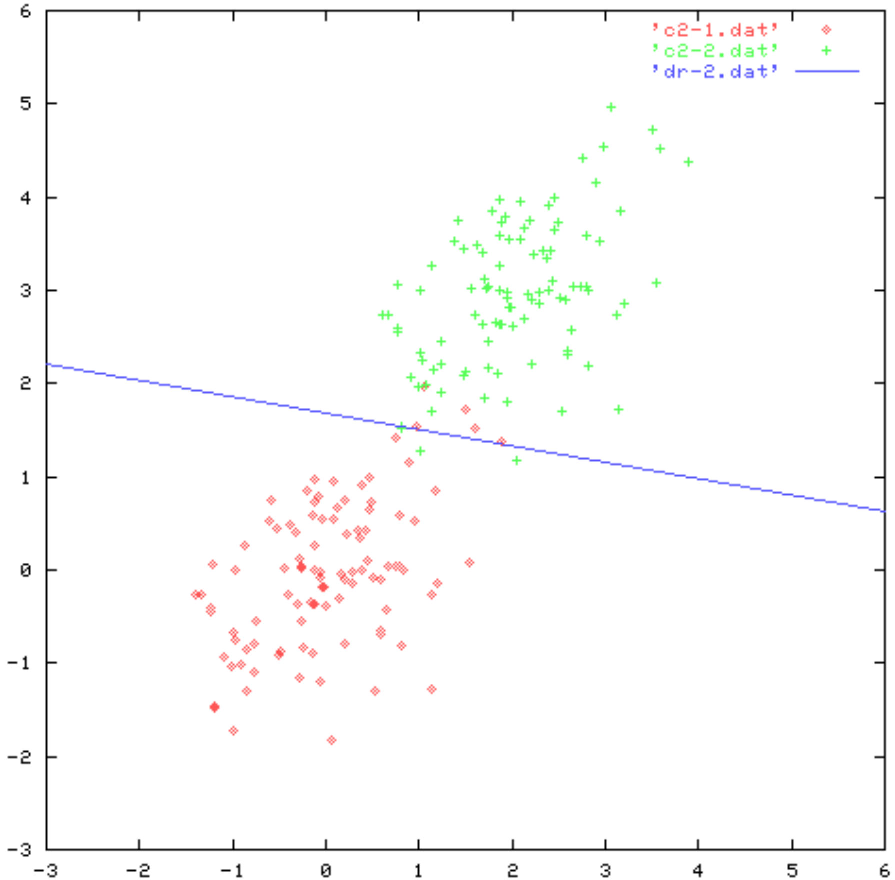


$$\Sigma = \begin{bmatrix} 0.68 & 0.34 \\ 0.34 & 0.64 \end{bmatrix}$$

$$\Sigma^{-1} = \begin{bmatrix} 2.03 & -1.10 \\ -1.10 & 2.17 \end{bmatrix}$$

Decision boundary :

$$0.76x_0 + 4.31x_1 = 7.23$$

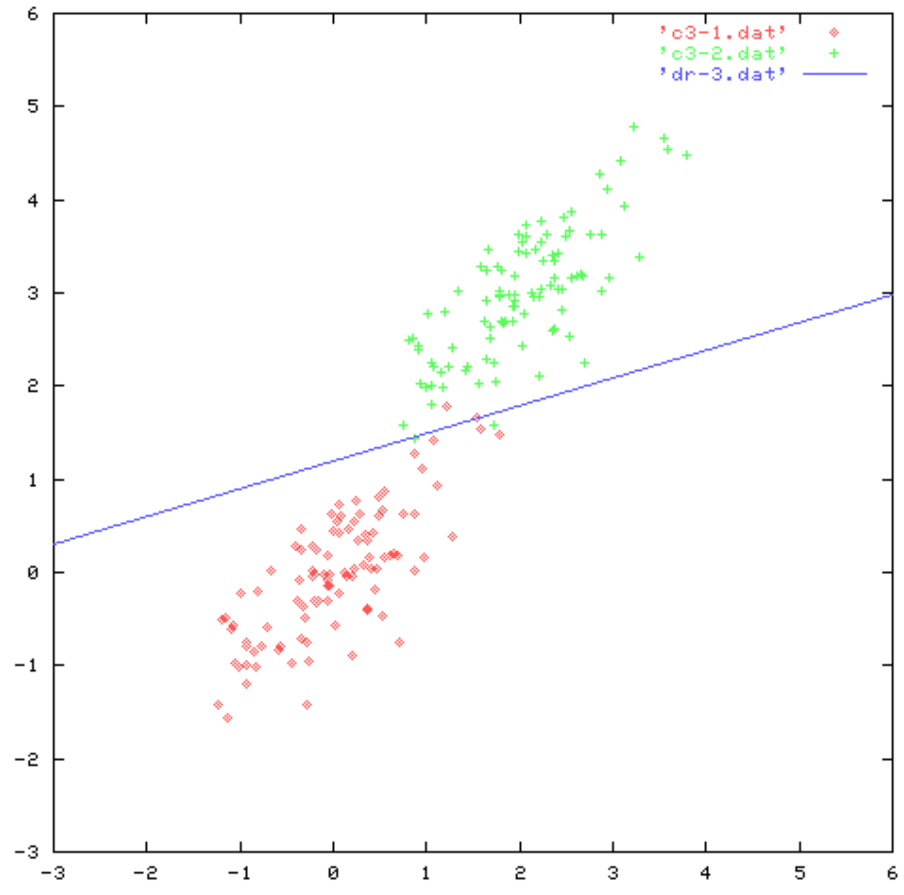


$$\Sigma = \begin{bmatrix} 0.58 & 0.44 \\ 0.44 & 0.54 \end{bmatrix}$$

$$\Sigma^{-1} = \begin{bmatrix} 4.66 & -3.84 \\ -3.84 & 5.02 \end{bmatrix}$$

Decision boundary :

$$-2.19x_0 + 7.37x_1 = 8.86$$

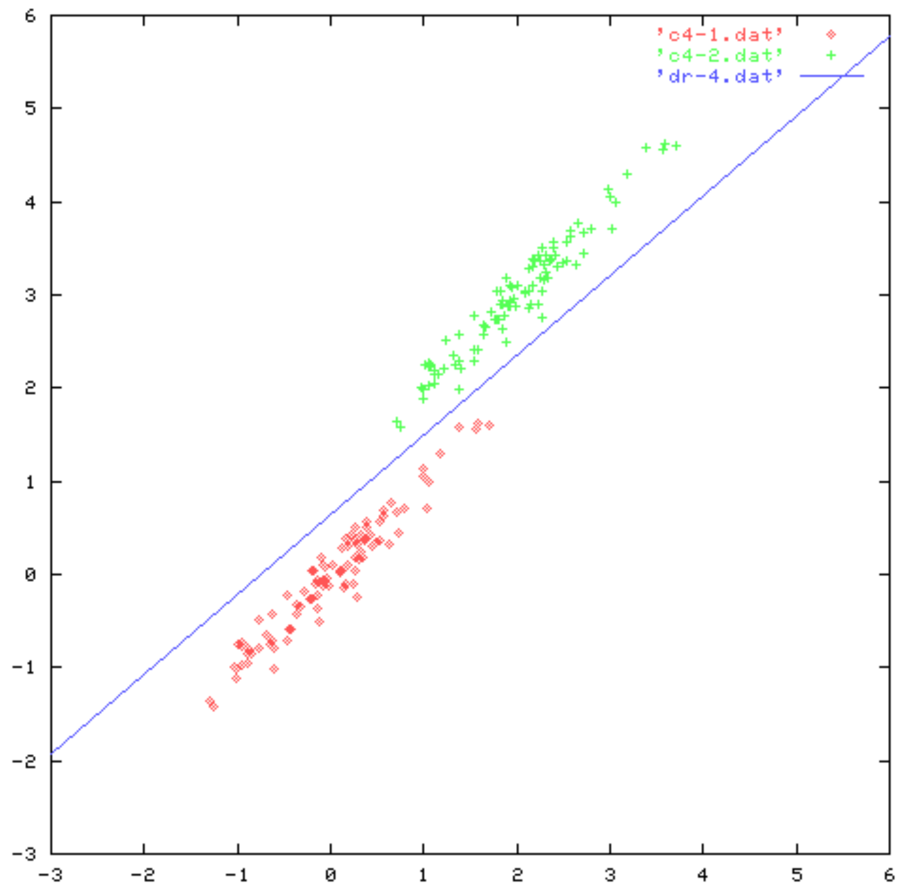


$$\Sigma = \begin{bmatrix} 0.52 & 0.50 \\ 0.50 & 0.50 \end{bmatrix}$$

$$\Sigma^{-1} = \begin{bmatrix} 34.78 & -34.44 \\ -34.44 & 36.11 \end{bmatrix}$$

Decision boundary:

$$-33.78x_0 + 39.44x_1 = 25.39$$



Linear Classifier

- Optimal when the two classes are Gaussian and differ only in their means
- The decision boundary is linear