Linear Discriminant Function

Dripta Mj
Department of Mathematics
Ramakrishna Mission Vivekananda Educational and Research Institute
Belur Math, India

Machine Learning

CS230

Sem 3, 2018-19

Discriminant Function

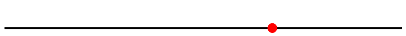
- A discriminant is a function (say $f(\mathbf{x})$) used to check the class of data points.

- For a two class classifier
  - if $f(\mathbf{x}) > 0$, then data point $\mathbf{x}$ is assigned to class c_1.
  - if $f(\mathbf{x}) < 0$, then data point $\mathbf{x}$ is assigned to class c_2.

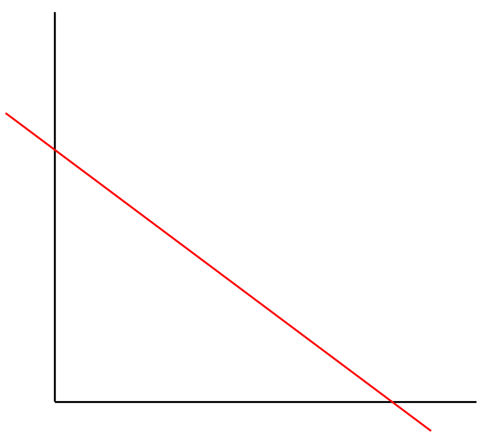
- $f(\mathbf{x}) = 0$ is the discriminant surface.

Hyperplane

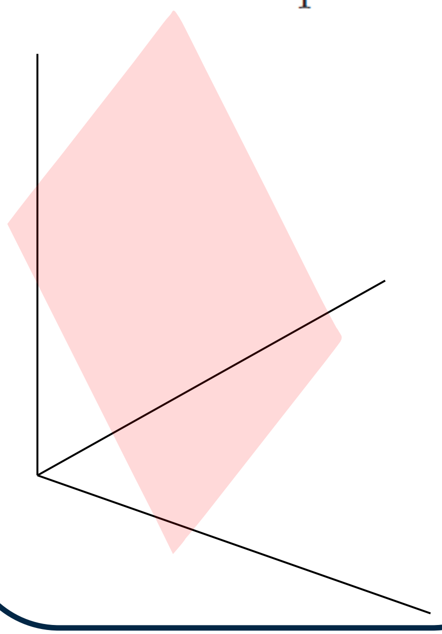
- The decision surface separates points assigned to class c_1 from those assigned to class c_2.
- If the function $f(\mathbf{x})$ is linear, then the decision surface is a hyperplane.
- A hyperplane
  - in **1D** is a point



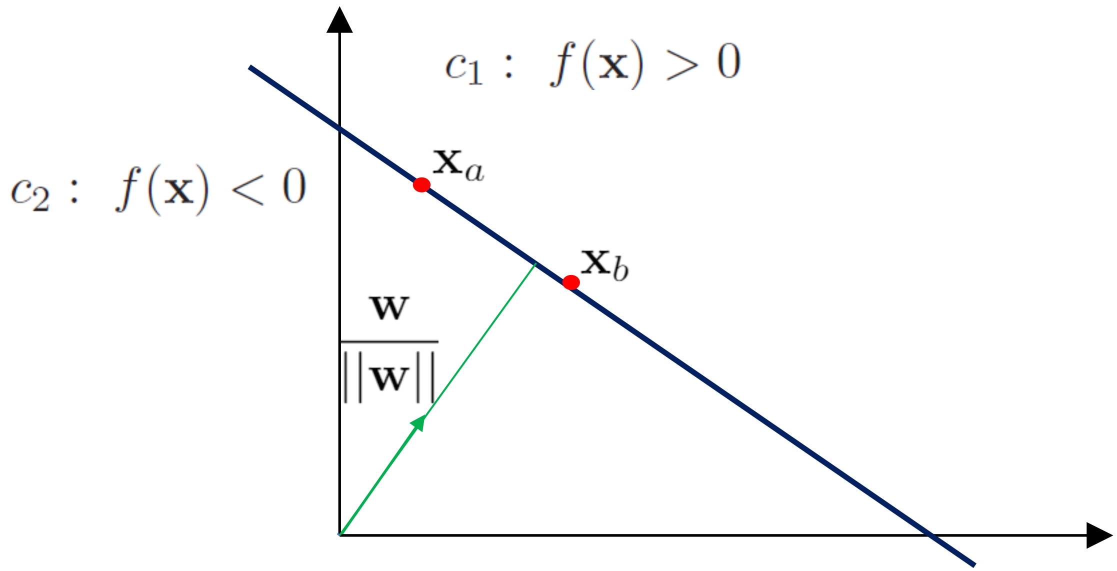
  - in **2D** is a line



  - in **3D** is a plane



Two classes – linear discriminant



- Linear discriminant function can written in the form:

$$f(\mathbf{x}) = \mathbf{w}^{\mathrm{T}}\mathbf{x} + w_0$$

- Consider two points – $\mathbf{x}_a$ and $\mathbf{x}_b$ – on the decision surface $f(\mathbf{x}) = 0$.

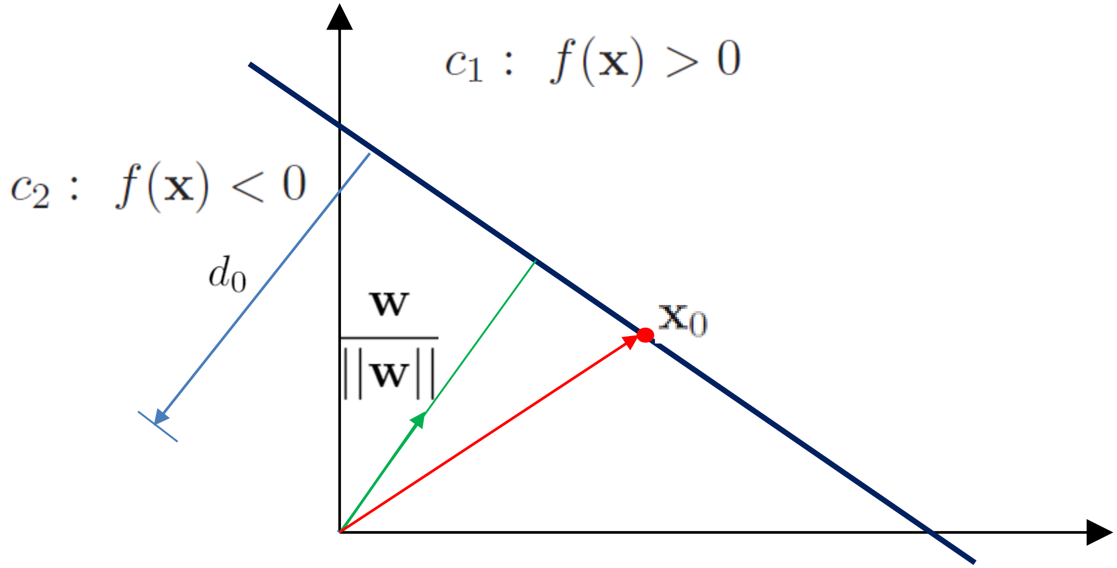
$$f(\mathbf{x}_a) = 0 \;\Rightarrow \mathbf{w}^{\mathrm{T}}\mathbf{x}_a + w_0 = 0$$

$$f(\mathbf{x}_b) = 0 \;\Rightarrow \mathbf{w}^{\mathrm{T}}\mathbf{x}_b + w_0 = 0$$

$$\overline{\mathbf{w}^{\mathrm{T}}(\mathbf{x}_a - \mathbf{x}_b) = 0}$$

- Therefore the vector $\mathbf{w}$ is orthogonal to all vectors lying on the decision surface.

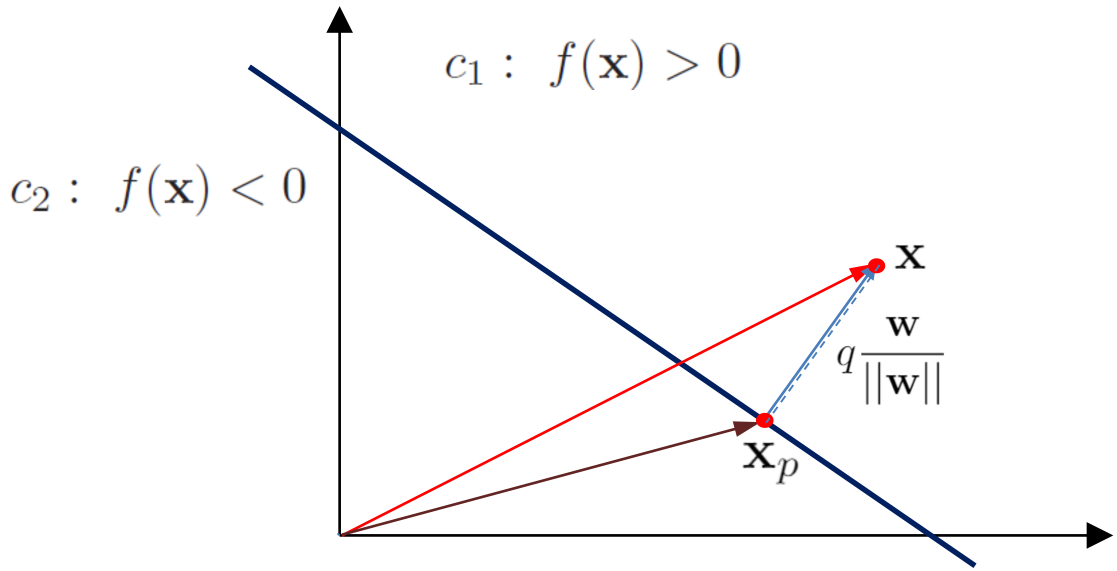
Distance from origin



- Want to compute the distance d_0 between the decision surface and the origin.
- Consider a point (say $\mathbf{x}_0$) on the decision surface, then d_0 can be computed as

$$\begin{aligned} d_0 &= \frac{\mathbf{w}^{\mathrm{T}}}{||\mathbf{w}||}(\mathbf{x}_0 - \mathbf{0}) \\ &= -\frac{w_0}{||\mathbf{w}||} \qquad \text{(since } f(\mathbf{x}_0) = 0\text{)} \end{aligned}$$

Modelling distance from an arbitrary point



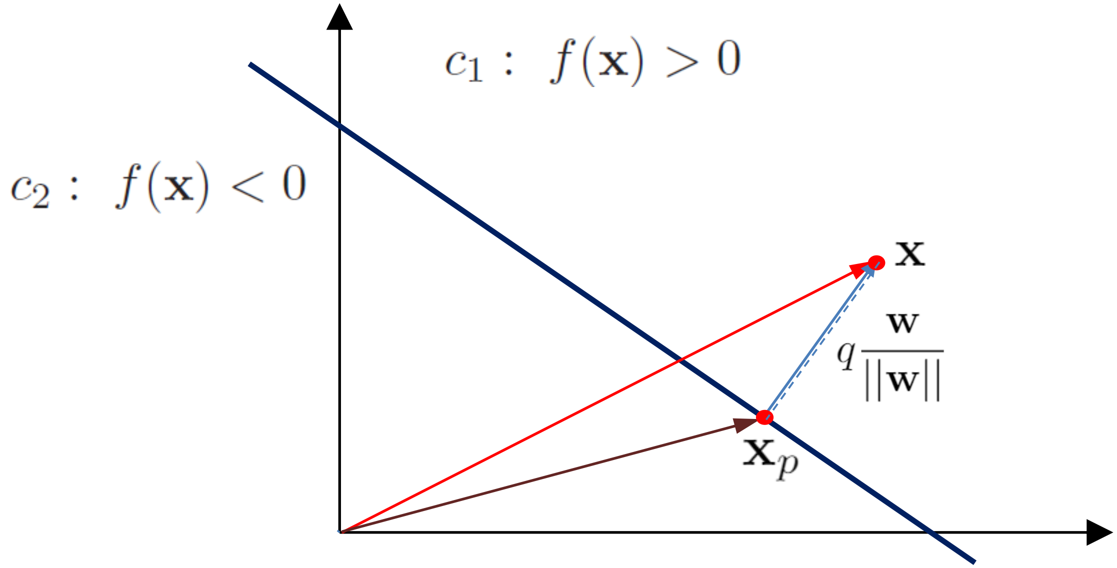
- Consider an arbitrary point $\mathbf{x}$ in the feature space.
- Suppose $\mathbf{x}_p$ is the orthogonal projection of the point $\mathbf{x}$ on the decision surface, which means

$$f(\mathbf{x}_p) = \mathbf{w}^{\mathrm{T}}\mathbf{x}_p + w_0 = 0$$

- Let q be the distance between $\mathbf{x}$ and $\mathbf{x}_p$, then can write

$$\mathbf{x} = \mathbf{x}_p + q\frac{\mathbf{w}}{||\mathbf{w}||}$$

Signed orthogonal distance



- Multiplying both sides of the equation by $\mathbf{w}^{\mathrm{T}}$, we have

$$\mathbf{w}^{\mathrm{T}}\mathbf{x} = \mathbf{w}^{\mathrm{T}}\mathbf{x}_p + q\frac{\mathbf{w}^{\mathrm{T}}\mathbf{w}}{||\mathbf{w}||}$$

$$f(\mathbf{x}) - w_0 = -w_0 + q\frac{||\mathbf{w}||^2}{||\mathbf{w}||}$$

$$\Rightarrow \quad q = \frac{f(\mathbf{x})}{||\mathbf{w}||}$$

Key points (2 classes)

- Linear discriminant function divides the feature space using hyperplane decision surface.
- The vector $\mathbf{w}$ is orthogonal to the decision surface and indicates its orientation.
- The bias parameter w_0 determines the location of the decision surface.
- For an arbitrary point $\mathbf{x}$, the value $f(\mathbf{x})/||\mathbf{w}||$ yields a signed measure of the the orthogonal distance form the point $\mathbf{x}$ to the decision surface.

Multiple classes

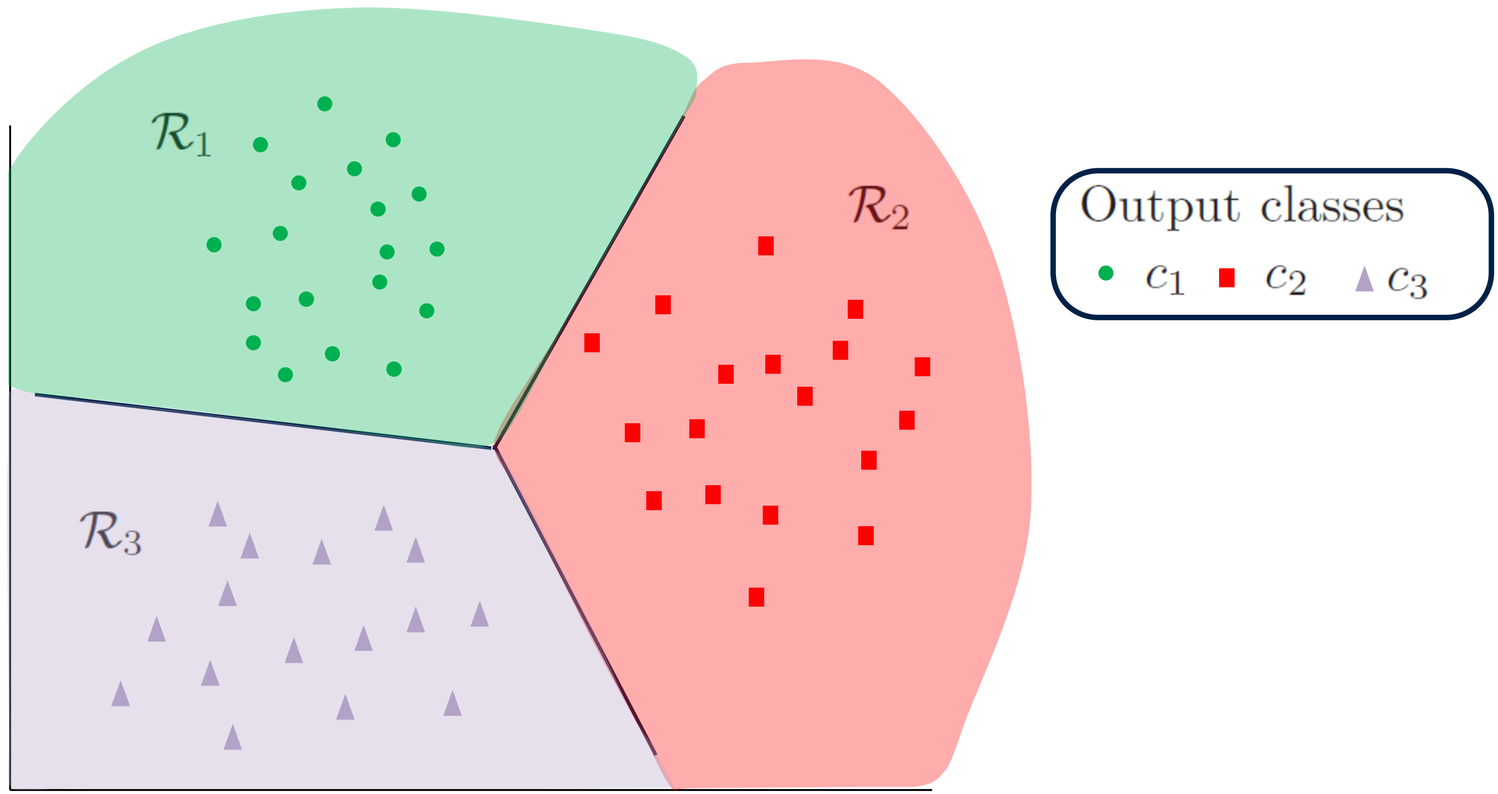
- Consider a problem with J output classes: $\{\mathcal{C}_1, \mathcal{C}_2, ..., \mathcal{C}_J\}$.

- Can use J linear discriminants: $\{f_1(\mathbf{x}), f_2(\mathbf{x}), ..., f_J(\mathbf{x})\}$.

- Assign an example to class $\mathcal{C}_j$ if $f_j(\mathbf{x}) > f_i(\mathbf{x})$, for all $j \neq i$.

- Decision boundaries divide the feature space into decision regions – $\{\mathcal{R}_1, \mathcal{R}_2, ..., \mathcal{R}_J\}$.
  In the jth region $\mathcal{R}_j$ we have $f_j(\mathbf{x}) > f_i(\mathbf{x})$, for all $j \neq i$.

3 classes



Gaussian distribution

- Class conditional probability distribution $p(\mathbf{x}|c_j)$ is taken to be Gaussian:

$$p(\mathbf{x}|c_j) = \frac{1}{(2\pi)^{D/2}|\Sigma|^{1/2}} \exp\Big[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_j)^{\mathrm{T}}\Sigma_j^{-1}(\mathbf{x}-\boldsymbol{\mu}_j)\Big]$$

  where $\boldsymbol{\mu}_j$ is the mean vector and Σ_j is the covariance matrix of the features corresponding to class c_j.

- A linear decision boundary is obtained when the covariance of the classes are the same.

- The posterior probability can be computed using Bayes rule:

$$\begin{aligned} P(c_j|\mathbf{x}) &= \frac{p(\mathbf{x}|c_j)P(c_j)}{p(\mathbf{x})} \\ &= \frac{p(\mathbf{x}|c_j)p(c_j)}{\sum_{j=1}^{J} p(\mathbf{x}|c_j)p(c_j)} \end{aligned}$$

Discriminant function

- Taking ln of the posterior distribution gives

$$\begin{aligned}\ln P(c_j|\mathbf{x}) &= \ln p(\mathbf{x}|c_j) + \ln P(c_j) + \text{const.}\\ &= f_j(\mathbf{x})\end{aligned}$$

  where $f_j(\mathbf{x})$ is the discriminant function corresponding to the jth class.

- So we have a set of discriminant functions – $\{f_1(\mathbf{x}), f_2(\mathbf{x}), ..., f_J(\mathbf{x})\}$, one for each class.

- For a Gaussian conditional distribution we obtain the discriminant function of the jth class to be

$$\begin{aligned}f_j(\mathbf{x}) &= \ln\left(\frac{1}{(2\pi)^{D/2}|\Sigma_j|^{1/2}}\exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_j)^{\mathrm{T}}\Sigma_j^{-1}(\mathbf{x}-\boldsymbol{\mu}_j)\right]\right) + \ln P(c_j) + \text{const.}\\ &= -\frac{D}{2}\ln(2\pi) - \frac{1}{2}\ln|\Sigma_j| - \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_j)^{\mathrm{T}}\Sigma_j^{-1}(\mathbf{x}-\boldsymbol{\mu}_j) + \ln P(c_j) + \text{const.}\\ &= -\frac{D}{2}\ln(2\pi) - \frac{1}{2}\ln|\Sigma_j| - \frac{1}{2}(\mathbf{x}^{\mathrm{T}}\Sigma_j^{-1}\mathbf{x} - \mathbf{x}^{\mathrm{T}}\Sigma_j^{-1}\boldsymbol{\mu}_j - \boldsymbol{\mu}_j^{\mathrm{T}}\Sigma_j^{-1}\mathbf{x} + \boldsymbol{\mu}_j^{\mathrm{T}}\Sigma_j^{-1}\boldsymbol{\mu}_j)\\ &\qquad + \ln P(c_j) + \text{const.}\end{aligned}$$

$\Sigma_j = \Sigma$

- Suppose all the class conditional Gaussian distributions have the same covariance matrix Σ. Then we have

$$f_j(\mathbf{x}) = -\frac{D}{2}\ln(2\pi) - \frac{1}{2}\ln|\Sigma| - \frac{1}{2}(\mathbf{x}^{\mathrm{T}}\Sigma^{-1}\mathbf{x} - \mathbf{x}^{\mathrm{T}}\Sigma^{-1}\boldsymbol{\mu}_j - \boldsymbol{\mu}_j^{\mathrm{T}}\Sigma^{-1}\mathbf{x} + \boldsymbol{\mu}_j^{\mathrm{T}}\Sigma^{-1}\boldsymbol{\mu}_j) + \ln P(c_j) + \text{const.}$$

- The terms that are independent of j are constant and common to all discriminant functions $f_1(\mathbf{x}), f_2(\mathbf{x}), \ldots., f_J(\mathbf{x})$, and so can be ignored.

- The simplification yields

$$f_j(\mathbf{x}) = -\frac{1}{2}(-\mathbf{x}^{\mathrm{T}}\Sigma^{-1}\boldsymbol{\mu}_j - \boldsymbol{\mu}_j^{\mathrm{T}}\Sigma^{-1}\mathbf{x} + \boldsymbol{\mu}_j^{\mathrm{T}}\Sigma^{-1}\boldsymbol{\mu}_j) + \ln P(c_j)$$

- Now Σ is a symmetric matrix, and so we have $\mathbf{x}^{\mathrm{T}}\Sigma^{-1}\boldsymbol{\mu}_j = \boldsymbol{\mu}_j^{\mathrm{T}}\Sigma^{-1}\mathbf{x}$.

Linear discriminant

- Finally we obtain

$$f_j(\mathbf{x}) = \boldsymbol{\mu}_j^{\mathrm{T}}\Sigma^{-1}\mathbf{x} - \frac{1}{2}\boldsymbol{\mu}_j^{\mathrm{T}}\Sigma^{-1}\boldsymbol{\mu}_j + \ln P(c_j)$$
$$= \mathbf{w}_j^{\mathrm{T}}\mathbf{x} + \mathbf{w}_{j,0}$$

where

$$\mathbf{w}_j^{\mathrm{T}} = \boldsymbol{\mu}_j^{\mathrm{T}}\Sigma^{-1}$$

$$\mathbf{w}_{j,0} = -\frac{1}{2}\boldsymbol{\mu}_j^{\mathrm{T}}\Sigma^{-1}\boldsymbol{\mu}_j + \ln P(c_j)$$

- Therefore $f_j(\mathbf{x})$ is linear discriminant as it is a linear function as it is a linear function of $\mathbf{x}$.

Decision boundary

- The decision boundary between two classes $\mathcal{C}_j$ and $\mathcal{C}_i$ is given as

$$f_j(\mathbf{x}) = f_i(\mathbf{x})$$

$$\Rightarrow \mathbf{w}_j^{\mathrm{T}}\mathbf{x} + \mathbf{w}_{j,0} = \mathbf{w}_i^{\mathrm{T}}\mathbf{x} + \mathbf{w}_{i,0}$$

$$\Rightarrow \left(\mathbf{w}_j - \mathbf{w}_i\right)^{\mathrm{T}}\mathbf{x} + \left(\mathbf{w}_{j,0} - \mathbf{w}_{i,0}\right) = 0$$

where

$$\left(\mathbf{w}_j - \mathbf{w}_i\right)^{\mathrm{T}} = \left(\boldsymbol{\mu}_j - \boldsymbol{\mu}_i\right)^{\mathrm{T}}\Sigma^{-1}$$

$$\left(\mathbf{w}_{j,0} - \mathbf{w}_{i,0}\right) = -\frac{1}{2}\left(\boldsymbol{\mu}_j^{\mathrm{T}}\Sigma^{-1}\boldsymbol{\mu}_j - \boldsymbol{\mu}_i^{\mathrm{T}}\Sigma^{-1}\boldsymbol{\mu}_i\right) + \ln\left(\frac{P(c_j)}{P(c_i)}\right)$$