Recap - Linear Discriminants

Given \( \{x_i \mid x_i \in C_i, i = 1\ldots k \} \)

find the discriminant (decision function) \( d(x, C_i) \) such that

\[
d(x, C_i) > d(x, C_j) \ \forall \ x \in C_i, i \neq j
\]

Linear Discriminants \( d(x) = w^T x + w_0 \)
Decision boundary is a hyperplane (line in 2-feature space).

Perceptron Algorithm

Serial
\[
a_0 = 0 \quad a_{k+1} = a_k + y^k \quad \text{where} \quad a_k^T y^k \leq 0
\]

Parallel
\[
a_0 = 0 \quad a_{k+1} = a_k + \sum_{Y(a_k)} y_i
\]
Recap - Optimization Function

The perceptron discriminant is an example of a more general optimization approach.

Define a criterion function \( J(\alpha) \) which is a minimum if \( \alpha \) is a solution.

\[
J(\alpha) \triangleq - \sum_{y \in Y} a^T y \quad Y = \{ y | a^T y \leq 0 \}
\]

For this function, \( J(\alpha) \) is a measure of distance from \( y_i \) to the plane \( \alpha \cdot y_i = 0 \)

\[
\frac{\delta J(\alpha)}{\delta a} = - \sum_{y \in Y} y
\]

In the case where the classes are non-separable, the perceptron algorithm does not converge. How do we find a suitable linear discriminant?

**FIGURE 3.5.** Projection of the same set of samples onto two different lines in the directions marked \( w \). The figure on the right shows greater separation between the red and black projected points. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.
Optimization with a Criterion Function

Ideally, we want to find a discriminant that maximizes interclass (between class) variability and minimizes intraclass (within class) variability.

This will provide the best separation of classes while keeping them compact.

Assume we have samples from two classes...

\[ \text{We need to find } w, \text{ such that when we project the samples of these classes onto } w \text{ then the features are as separated as possible:} \]

\[ \text{var}[d(x)|C_1] \]
\[ \text{var}[d(x)|C_2] \]
\[ \text{E}[d(x)|C_1] \]
\[ \text{E}[d(x)|C_2] \]
Return to $n$ dimensions for optimizing (do not use $a$ and $y$).
\[ d(x) = w^T x + w_0 \]

Define:
\[ J(w) = \frac{(E[d(x)|C_1] - E[d(x)|C_2])^2}{N_1 \text{var}[d(x)|C_1] + N_2 \text{var}[d(x)|C_2]} \]

We want to maximize this value to find the optimal $w$. This is Fisher’s criterion.

\[
J(w) = \frac{(E[d(x)|C_1] - E[d(x)|C_2])^2}{N_1 \text{var}[d(x)|C_1] + N_2 \text{var}[d(x)|C_2]}
\]

We can solve for each piece as follows...

\[
E[d(x)|C_1] = E[w^T x + w_0|C_1] = \frac{1}{N} \sum_i (w^T (x_i - m_i))^2
\]

\[
\text{var}[d(x)|C_1] = \frac{1}{N} \sum_i (w^T (x_i - m_i))^2 = w^T S_1 w
\]

\[
E[d(x)|C_2] = w^T m_2 + w_0
\]

\[
\text{var}[d(x)|C_2] = w^T S_2 w
\]

So we can rewrite Fisher’s criterion as:
\[ J(w) = \frac{w^T (m_1 - m_2)(m_1 - m_2)^T w}{w^T N_1 S_1 w + w^T N_2 S_2 w} \]
Define between class scatter: \[ S_B \triangleq (m_1 - m_2)(m_1 - m_2)^T \]
and within class scatter: \[ S_W \triangleq N_1S_1 + N_2S_2 \]

Thus Fisher’s criterion becomes
\[ J(w) = \frac{w^T S_B w}{w^T S_W w} \]

We need to maximize this value to find the optimal value for \( w \).

\[
\begin{align*}
\frac{\partial J(w)}{\partial w} &= \frac{2S_B w}{w^T S_W w} - 2 \frac{w^T S_B w}{(w^T S_W w)^2} S_W w \\
\therefore \frac{\partial J(w)}{\partial w} &= 0 \rightarrow S_B w = J(w)S_W w \\
S_W^{-1}S_B w &= J(w)w \\
\Rightarrow w &= S_W^{-1}(m_1 - m_2)
\end{align*}
\]

So \( w \) is the maximum eigenvalue eigenvector of \( S_W^{-1}S_B \).

However, we don’t need to solve for the eigenvectors to find \( w \):
\[
S_B w = J(w)S_W w
\]

\[
(m_1 - m_2)(m_1 - m_2)^T w = J(w)S_W w
\]

\[
\Rightarrow w = S_W^{-1}(m_1 - m_2)
\]

Ignore scalars, because all we need is the direction.
Geometrically, $S_W^{-1}$ rotates the vector $(m_1 - m_2)$ between the means to take into account the shapes of the class distributions.

Full Fisher’s linear discriminant:

$$d_{FLD}(x) = w^T x + w_0 = (m_1 - m_2)^T S_W^{-1} (x - m)$$

where $w_0 = -w^T m$

and $m$ is the mean of all samples

Why is this a good choice for $w_0$?

It offsets the discriminant by the mean of all samples projected onto the line given by $w$. 
**Example of Calculating FLD**

Class 1

\[
m_1 = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad S_1 = \begin{bmatrix} 5 & 3 \\ 3 & 5 \end{bmatrix}
\]

Class 2

\[
m_2 = \begin{bmatrix} 5 \\ 0 \end{bmatrix} \quad S_2 = \begin{bmatrix} 5 & 3 \\ 3 & 5 \end{bmatrix}
\]

a. Sketch the unit standard deviation (USD) contours

b. Find fisher’s linear discriminant (FLD) and sketch the decision boundary