Lecture 16

Discriminant Analysis

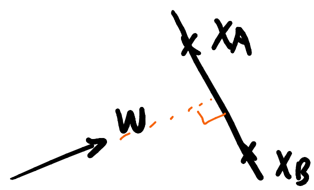
Linearly Separable Classes:
Data sets whose classes can be separated exactly by linear decision surfaces are said to be linearly separable

$$y(x) = w^T x + w_0 = 0 \text{ : decision boundary}$$

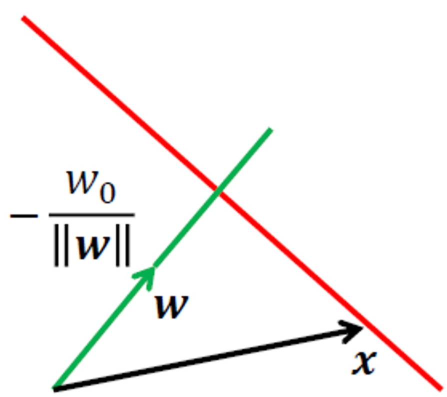
$$\left.\begin{array}{l} y(x_A) = 0 \Rightarrow w^T x_A + w_0 = 0 \\ y(x_B) = 0 \Rightarrow w^T x_B + w_0 = 0 \end{array}\right\} \Rightarrow w^T x_A = w^T x_B$$

$$\Downarrow$$

$$w^T (x_A - x_B) = 0$$



hence the vector $\boldsymbol{w}$ is orthogonal to every vector lying within the decision surface



if $\boldsymbol{x}$ is a point on the decision surface, then $\frac{\boldsymbol{w}^T \boldsymbol{x}}{\|\boldsymbol{w}\|}$ is the projection of the point $\boldsymbol{x}$ onto the weight vector $\boldsymbol{W}$. The projection remains the same regardless of the location of $\boldsymbol{x}$.

$$\frac{\boldsymbol{w}^T \boldsymbol{x}}{\|\boldsymbol{w}\|} = -\frac{w_0}{\|\boldsymbol{w}\|}$$

$$w^T x + w_0 = 0$$

Fisher's Linear Discriminant

Fisher's idea is to maximize a function that will give a large separation between the projected class means, while also giving a small variance within each class, thereby minimizing the class overlap

The projection $y = \mathbf{w}^{\mathrm{T}}\mathbf{x}$ transforms the set of labeled data points in $\mathbf{x}$ into a labeled set in the one-dimensional space y. The within-class variance of the transformed data from class C_k is therefore given by

$$s_k^2 = \sum_{n \in \mathcal{C}_k} (y_n - m_k)^2$$

where $y_n = \mathbf{w}^{\mathrm{T}}\mathbf{x}_{\mathrm{n}}$

We can define the total within-class variance for the whole data set to be s $s_1^2 + s_2^2$.

$$m_2 - m_1 = \mathbf{w}^{\mathrm{T}}(\mathbf{m}_2 - \mathbf{m}_1)$$

where m_k is the mean of the projected data from class C_k:

$$m_k = \mathbf{w}^{\mathrm{T}}\mathbf{m}_k$$

Fisher's Criterion

- The Fisher criterion is defined to be the ratio of the between-class variance to the within-class variance and is given by
$$J(\mathbf{w}) = \frac{(m_2 - m_1)^2}{s_1^2 + s_2^2}$$
- This Fisher criterion can be rewritten in matrix form as:
$$J(\mathbf{w}) = \frac{\overset{(1\times n)}{\mathbf{w}^{\mathrm{T}}} \overset{(n\times n)}{\mathbf{S}_{\mathrm{B}}} \overset{(n\times 1)}{\mathbf{w}}}{\mathbf{w}^{\mathrm{T}}\mathbf{S}_{\mathrm{W}}\mathbf{w}} \leftarrow 1\times 1$$

where $\mathbf{S}_{\mathrm{B}}$ is the *between-class* covariance matrix given by

$S_B = S_B^T \Leftarrow$
$$\mathbf{S}_{\mathrm{B}} = (\mathbf{m}_2 - \mathbf{m}_1)(\mathbf{m}_2 - \mathbf{m}_1)^{\mathrm{T}}$$

$$w^T S_B w = \underbrace{w^T(\vec{m}_2 - \vec{m}_1)}_{(m_2 - m_1)}\underbrace{(\vec{m}_2 - \vec{m}_1)^T w}_{(m_2 - m_1)} = (m_2 - m_1)^2$$

$\uparrow \uparrow$ Scalars

$$= \left[w^T(m_2 - m_1)\right]^T = \underbrace{(m_2 - m_1)^T}_{\text{Scalar}}$$

and $\mathbf{S}_{\mathrm{W}}$ is the total *within-class* covariance matrix, given by

$$\mathbf{S}_{\mathrm{W}} = \sum_{n\in\mathcal{C}_1}(\mathbf{x}_n - \mathbf{m}_1)(\mathbf{x}_n - \mathbf{m}_1)^{\mathrm{T}} + \sum_{n\in\mathcal{C}_2}(\mathbf{x}_n - \mathbf{m}_2)(\mathbf{x}_n - \mathbf{m}_2)^{\mathrm{T}}.$$

$$J(\mathbf{w}) = \frac{\mathbf{w}^{\mathrm{T}}\mathbf{S}_{\mathrm{B}}\mathbf{w}}{\mathbf{w}^{\mathrm{T}}\mathbf{S}_{\mathrm{W}}\mathbf{w}}$$

$$\left(\frac{u}{v}\right)' = \frac{u'v - uv'}{v^2}$$

- Determine the value of $\mathbf{w}$ such that $J(\mathbf{w})$ is maximized, by differentiating $J(\mathbf{w})$ with respect to $\mathbf{w}$:

$$J'(\mathbf{w}) = \frac{(\mathbf{w}^{\mathrm{T}}\mathbf{S}_{\mathrm{B}}\mathbf{w})'(\mathbf{w}^{\mathrm{T}}\mathbf{S}_{\mathrm{W}}\mathbf{w}) - (\mathbf{w}^{\mathrm{T}}\mathbf{S}_{\mathrm{B}}\mathbf{w})(\mathbf{w}^{\mathrm{T}}\mathbf{S}_{\mathrm{W}}\mathbf{w})'}{(\mathbf{w}^{\mathrm{T}}\mathbf{S}_{\mathrm{W}}\mathbf{w})^2}$$
$$= \frac{2\mathbf{S}_{\mathrm{B}}\mathbf{w}(\mathbf{w}^{\mathrm{T}}\mathbf{S}_{\mathrm{W}}\mathbf{w}) - 2\mathbf{S}_{\mathrm{W}}\mathbf{w}(\mathbf{w}^{\mathrm{T}}\mathbf{S}_{\mathrm{B}}\mathbf{w})}{(\mathbf{w}^{\mathrm{T}}\mathbf{S}_{\mathrm{W}}\mathbf{w})^2} = 0$$

Thus

$$\mathbf{S}_{\mathrm{B}}\mathbf{w}(\mathbf{w}^{\mathrm{T}}\mathbf{S}_{\mathrm{W}}\mathbf{w}) = \mathbf{S}_{\mathrm{W}}\mathbf{w}(\mathbf{w}^{\mathrm{T}}\mathbf{S}_{\mathrm{B}}\mathbf{w})$$
$$\frac{\mathbf{S}_{\mathrm{W}}^{-1}\mathbf{S}_{\mathrm{B}}\mathbf{w}(\mathbf{w}^{\mathrm{T}}\mathbf{S}_{\mathrm{W}}\mathbf{w})}{(\mathbf{w}^{\mathrm{T}}\mathbf{S}_{\mathrm{B}}\mathbf{w})} = \mathbf{w}$$

For a scalar α given by a quadratic form: $\alpha = \mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{x}$

For the special case $\mathbf{A}^{\mathrm{T}} = \mathbf{A}$, then $\underbrace{\frac{\partial[\mathbf{x}^{\mathrm{T}}(\mathbf{A}+\mathbf{A}^{\mathrm{T}})]}{\partial\mathbf{x}}}_{\text{column vector}} = (2\mathbf{x}^{\mathrm{T}}\mathbf{A})^{\mathrm{T}} = 2A^T x$

$$(\mathbf{w}^T S_B \mathbf{w})' = 2\underbrace{S_B^T}_{= S_B}\mathbf{w} = 2S_B\mathbf{w}$$

w: $n\times 1$

$\Rightarrow w^T$: $1\times n$

$$\mathbf{w} = \frac{\mathbf{S}_{\mathrm{W}}^{-1}\mathbf{S}_{\mathrm{B}}\mathbf{w}(\mathbf{w}^{\mathrm{T}}\mathbf{S}_{\mathrm{W}}\mathbf{w})}{(\mathbf{w}^{\mathrm{T}}\mathbf{S}_{\mathrm{B}}\mathbf{w})}$$, where

$$\begin{aligned}\mathbf{S}_{\mathrm{B}}\mathbf{w} &= (\mathbf{m}_2 - \mathbf{m}_1)(\mathbf{m}_2 - \mathbf{m}_1)^{\mathrm{T}}\mathbf{w} \\ &= (\mathbf{m}_2 - \mathbf{m}_1)\left(\mathbf{w}^{\mathrm{T}}(\mathbf{m}_2 - \mathbf{m}_1)\right)^{\mathrm{T}} \\ &= (\mathbf{m}_2 - \mathbf{m}_1)(m_2 - m_1)\end{aligned}$$

Since $(\mathbf{w}^{\mathrm{T}}\mathbf{S}_{\mathrm{W}}\mathbf{w})$, $(\mathbf{w}^{\mathrm{T}}\mathbf{S}_{\mathrm{B}}\mathbf{w})$ and $(m_2 - m_1)$ are all scalar factors, we can drop them if we care only about the direction of the weight vector $\mathbf{w}$, instead of its magnitude. Thus we can obtain

$$\mathbf{w} \propto \mathbf{S}_{\mathrm{W}}^{-1}(\mathbf{m}_2 - \mathbf{m}_1)$$

$\mathbf{w} \propto \mathbf{S}_{\mathrm{W}}^{-1}(\mathbf{m}_2 - \mathbf{m}_1)$ is known as *Fisher's linear discriminant.*