Exercise 1: Fisher Discriminant (10 + 10 + 10 P)

The objective function to find the Fisher Discriminant has the form

$$\max_{\boldsymbol{w}} \frac{\boldsymbol{w}^\top \boldsymbol{S}_B \boldsymbol{w}}{\boldsymbol{w}^\top \boldsymbol{S}_W \boldsymbol{w}}$$

where $\boldsymbol{S}_B = (\boldsymbol{m}_2 - \boldsymbol{m}_1)(\boldsymbol{m}_2 - \boldsymbol{m}_1)^\top$ is the between-class scatter matrix and $\boldsymbol{S}_W$ is within-class scatter matrix, assumed to be positive definite. Because there are infinitely many solutions (multiplying $\boldsymbol{w}$ by a scalar doesn't change the objective), we can extend the objective with a constraint, e.g. that enforces $\boldsymbol{w}^\top \boldsymbol{S}_W \boldsymbol{w} = 1$.

(a) *Reformulate* the problem above as an optimization problem with a quadratic objective and a quadratic constraint.

$$\max_{w} \left(w^T S_B\, w \right) \quad \text{s.t.} \quad w^T S_W\, w = 1$$

(b) *Show* using the method of Lagrange multipliers that the solution of the reformulated problem is also a solution of the generalized eigenvalue problem:

$$\boldsymbol{S}_B \boldsymbol{w} = \lambda \boldsymbol{S}_W \boldsymbol{w}$$

$f(w) = w^T S_B\, w \qquad g(w) = w^T S_W\, w - 1$

$\mathcal{L}(w, \alpha) = w^T S_B w + \alpha \left(w^T S_W w - 1 \right)$, where α is a constant

$\nabla_w \mathcal{L}(w, \alpha) = 2 S_B\, w + 2\alpha S_W W = 0$

$\Rightarrow \; S_B W = -\alpha\, S_W W$

$\Rightarrow \; S_B W = \lambda\, S_W W$

(c) Show that the solution of this optimization problem is equivalent (up to a scaling factor) to

$$\boldsymbol{w}^\star = \boldsymbol{S}_W^{-1}(\boldsymbol{m}_1 - \boldsymbol{m}_2)$$

$S_B\, W = \lambda\, S_W\, W$

$\Leftrightarrow \; w = \frac{1}{\lambda} \cdot S_W^{-1} \cdot S_B \cdot w$

$\Leftrightarrow \; w = \frac{1}{\lambda} \cdot S_W^{-1} (m_2 - m_1) \underbrace{(m_2 - m_1)^T\, w}_{\text{scalar}}$

$\Leftrightarrow \; w = S_W^{-1} (m_2 - m_1)$

Exercise 2: Bounding the Error (10 + 10 P)

The direction learned by the Fisher discriminant is equivalent to that of an optimal classifier when the class-conditioned data densities are Gaussian with same covariance. In this particular setting, we can derive a bound on the classification error which gives us insight into the effect of the mean and covariance parameters on the error.

Consider two data generating distributions $P(\boldsymbol{x}|\omega_1) = \mathcal{N}(\boldsymbol{\mu}, \Sigma)$ and $P(\boldsymbol{x}|\omega_2) = \mathcal{N}(-\boldsymbol{\mu}, \Sigma)$ with $\boldsymbol{x} \in \mathbb{R}^d$. Recall that the Bayes error rate is given by:

$$P(\text{error}) = \int_{\boldsymbol{x}} P(\text{error}|\boldsymbol{x})\, p(\boldsymbol{x})\, d\boldsymbol{x}$$

(a) Show that the conditional error can be upper-bounded as:

$$P(\text{error}|\boldsymbol{x}) \leq \sqrt{P(\omega_1|\boldsymbol{x})P(\omega_2|\boldsymbol{x})}$$

$$P(\text{error}|x) = \min\{P(\omega_1|x), P(\omega_2|x)\} = M_{-\infty}(P(\omega_1|x), P(\omega_2|x))$$

$$M_0(P(\omega_1|x), P(\omega_2|x)) = \sqrt{\prod_{i=1}^{2} P(\omega_i|x)} = \sqrt{P(\omega_1|x)\cdot P(\omega_2 \cdot x)}$$

$$\Rightarrow \quad \min\{P(\omega_1|x), P(\omega_2|x)\} \leq \sqrt{P(\omega_1|x)\cdot P(\omega_2 \cdot x)}$$

(b) Show that the Bayes error rate can then be upper-bounded by:

$$P(\text{error}) \leq \sqrt{P(\omega_1)P(\omega_2)} \cdot \exp\left(-\frac{1}{2}\boldsymbol{\mu}^\top\Sigma^{-1}\boldsymbol{\mu}\right)$$

$$P(\text{error}) \leq \int_x \sqrt{P(\omega_1|x)\cdot P(\omega_2 \cdot x)} \cdot p(x)\, dx$$

$$\leq \int_x \sqrt{\frac{P(x|\omega_1)\cdot P(\omega_1)}{P(x)} \cdot \frac{P(x|\omega_2)\cdot P(\omega_2)}{P(x)}} \cdot p(x)\, dx$$

$$\leq \sqrt{P(\omega_1)P(\omega_2)} \int_x \sqrt{P(x|\omega_1)\cdot P(x|\omega_2)} \cdot p(x)\, dx$$

$$\leq \sqrt{P(\omega_1)P(\omega_2)} \int_x \sqrt{\frac{\exp(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu))}{(2\pi)^d \det(\Sigma)} \cdot \exp(-\frac{1}{2}(x+\mu)^T\Sigma^{-1}(x+\mu))}\, dx$$

$$\leq \sqrt{P(\omega_1)\cdot P(\omega_2)} \int_x \sqrt{\frac{\exp(-x^T\Sigma^{-1}x - \mu^T\Sigma^{-1}\mu)}{(2\pi)^d \det(\Sigma)}}\, dx$$

$$\leq \sqrt{P(\omega_1)P(\omega_2)} \cdot \exp(-\frac{1}{2}\mu^T\Sigma^{-1}\mu) \cdot \int_x \sqrt{\frac{\exp(-x^T\Sigma^{-1}x)}{(2\pi)^d \det(\Sigma)}}\, dx$$

$$= \sqrt{P(\omega_1)P(\omega_2)} \cdot \exp(-\frac{1}{2}\mu^T\Sigma^{-1}\mu)$$

$\square$

Exercise 3: Fisher Discriminant (10 + 10 P)

Consider the case of two classes ω_1 and ω_2 with associated data generating probabilities

$$p(\boldsymbol{x}|\omega_1) = \mathcal{N}\left(\begin{pmatrix}-1\\-1\end{pmatrix}, \begin{pmatrix}2 & 0\\0 & 1\end{pmatrix}\right) \qquad \text{and} \qquad p(\boldsymbol{x}|\omega_2) = \mathcal{N}\left(\begin{pmatrix}+1\\+1\end{pmatrix}, \begin{pmatrix}2 & 0\\0 & 1\end{pmatrix}\right)$$

(a) Find for this dataset the Fisher discriminant $\boldsymbol{w}$ (i.e. the projection $y = \boldsymbol{w}^\top \boldsymbol{x}$ under which the ratio between inter-class and intra-class variability is maximized).

$$S_W = \begin{pmatrix}2 & 0\\0 & 1\end{pmatrix} \qquad \omega_1 - \omega_2 = \begin{pmatrix}-1\\-1\end{pmatrix} - \begin{pmatrix}1\\1\end{pmatrix} = \begin{pmatrix}-2\\-2\end{pmatrix}$$

$$S_W^{-1}: \left(\begin{array}{cc|cc}2 & 0 & 1 & 0\\0 & 1 & 0 & 1\end{array}\right) = \left(\begin{array}{cc|cc}1 & 0 & \frac{1}{2} & 0\\0 & 1 & 0 & 1\end{array}\right)$$

$$w = S_W^{-1}(\omega_1 - \omega_2) = \begin{pmatrix}\frac{1}{2} & 0\\0 & 1\end{pmatrix}\begin{pmatrix}-2\\-2\end{pmatrix} = \begin{pmatrix}-1\\-2\end{pmatrix}$$

(b) Find a projection for which the ratio is minimized.

$$w = \begin{pmatrix}-1\\1\end{pmatrix}$$