Fachgebiet Maschinelles Lernen
Institut für Softwaretechnik und theoretische Informatik
Fakultät IV, Technische Universität Berlin
Prof. Dr. Klaus-Robert Müller
Email: klaus-robert.mueller@tu-berlin.de

Exercise Sheet 3

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.

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

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

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

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

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

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

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

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

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

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

Exercise 3: Fisher Discriminant (10 + 10 P)

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

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

(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).

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

Exercise 4: Programming (30 P)

Download the programming files on ISIS and follow the instructions.