Exercises for the course

Machine Learning 1

Winter semester 2024/25

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

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

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 $P(\boldsymbol{x} \mid \omega_1) = \mathcal{N}(\boldsymbol{\mu}, \Sigma)$ 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.

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.

$$\underset{w}{\operatorname{argmax}}\ w^T S_B w \qquad \text{with} \quad w^T S_W w = 1$$

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

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

$$\mathcal{L}(w, \lambda) = w^T S_B w + \lambda (1 - w^T S_W w)$$

$$\frac{\partial}{\partial \lambda} \mathcal{L}(w, \lambda) = 1 - w^T S_W w = 0$$

$$\nabla_w \mathcal{L}(w, \lambda) = 2 S_B w - 2 \cdot \lambda S_W \cdot w = 0 \quad \Rightarrow \quad S_B \cdot w = \lambda S_W \cdot w$$

Why $\nabla_w (w^T S_B w) = 2 S_B w$ and $\nabla_w (w^T S_W w) = 2 S_W \cdot w$?

$$w = \underbrace{\begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_d \end{bmatrix}}_{d} \qquad S_B = \overbrace{\begin{bmatrix} S_{11} & \cdots & \\ \vdots & \ddots & \vdots \\ & \cdots & S_{dd} \end{bmatrix}}^{d}$$

$$w^T S_B w = [w_1\ w_2\ \cdots\ w_d] \begin{bmatrix} S_{11} & S_{21} & \cdots & S_{d1} \\ S_{12} & & \ddots & \\ \vdots & & & \\ S_{1d} & & & S_{dd} \end{bmatrix} \begin{bmatrix} w_1 \\ \vdots \\ w_d \end{bmatrix}$$

$$= \underbrace{[w_1 S_{11} + w_2 S_{12} + \cdots + w_d S_{1d} \quad \cdots\cdots \quad w_1 S_{d1} + w_2 S_{d2} + \cdots + w_d S_{dd}]}_{d} \begin{bmatrix} w_1 \\ \vdots \\ w_d \end{bmatrix}$$

$$= [\, w_1 \cdot (w_1 S_{11} + w_2 S_{12} + \cdots + w_d S_{1d}) + w_2 (\cdots) + \cdots + w_d \cdot (w_1 S_{d1} + w_2 S_{d2} + \cdots + w_d S_{dd})$$

$$\underbrace{\nabla (w^T S_B w)}_{= f} = \begin{bmatrix} \frac{\partial f}{\partial w_1} \\ \vdots \\ \frac{\partial f}{\partial w_d} \end{bmatrix} = \begin{bmatrix} 2 S_{11} \cdot w_1 + 2 w_2 S_{21} + \cdots + 2 w_d S_{d1} \\ \cdots \\ 2 w_1 S_{1d} + \cdots + 2 w_d S_{dd} \end{bmatrix}$$

$$= 2 S_B \cdot w$$

A more formal proof is provided below:

Why $\nabla_w(w^T S_B w) = 2 S_B w$ and $\nabla_w(w^T S_W w) = 2 S_W \cdot w$:

formal Proof:

We show that $\nabla_w(w^T S w) = 2 S w$ if $S = S^T$. $w^T S w$ can be rewritten as

$$w^T S w = \sum_i w_i \sum_j S_{ij} \cdot w_j = \sum_i \sum_j w_i S_{ij} w_j$$

This is derived with respect to some w_k. Each $w_i S_{ij} w_j$ is looked at separately. If $i=j$, then $w_i S_{ij} w_j = w_i \cdot S_{ii} w_i = S_{ii} w_i^2$. This square is only non-zero when derived, if $i=j=k$; in this case we got $\frac{\partial}{w_k}(S_{kk} w_k^2) = 2 \cdot S_{kk} w_k^2$. If $i \neq j$ all terms where neither i nor j are k will be zero if derived with respect to x_k. Only pairs $\frac{\partial}{w_k}(w_i \cdot S_{ik} \cdot w_k) = S_{ik} \cdot w_i$ and $\frac{\partial}{w_k}(w_k \cdot S_{kj} w_j) = S_{kj} w_j$ are non-zero, because the others are constants. Because S is symmetric, we get $S_{ik} \cdot w_i = S_{ki} \cdot w_i$. Therefore the following holds

$$\frac{\partial}{w_k}\left(\sum_i \sum_j w_i S_{ij} w_j\right) = \sum_i 2 \cdot S_{ki} w_i = 2 \cdot \sum_i S_{ki} w_i = 2 \cdot S w$$

Therefore $\nabla_w(w^T S w) = 2 S w$ if $S = S^T$.

Both $S_B = (m_2 - m_1)(m_2 - m_1)^T$ and $S_W = S_1 + S_2$. S_W are symmetric, because the scatter matrix is symmetric:

$$\begin{aligned}
S_k &= \sum_{i \in C_k} (x_i - \mu_k)(x_i - \hat{\mu}_k)^T \\
&= \sum_{i \in C_k} \left((x_i - \mu_k)^T (x_i - \hat{\mu}_k)\right)^T \\
&= \left(\sum_{i \in C_k} (x_i - \mu_k)^T (x_i - \mu_k)\right)^T \\
&= \left(\sum_{i \in C_k} (x_i - \mu_k)(x_i - \mu_k)^T\right)^T = S_k^T
\end{aligned}$$

Using the lemma that was proven above results in $\nabla_w(w^T S_B w) = 2 S_B w$ and $\nabla_w(w^T S_W w) = 2 S_W w$.

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

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

$$S_B \cdot w = \lambda S_W \cdot w$$

$$(m_2 - m_1)(m_2 - m_1)^T \cdot w = \lambda S_W \cdot w$$

$$\Leftrightarrow \quad S_W^{-1}(m_2 - m_1)(m_2 - m_1)^T \cdot w = \lambda w$$

(S_W^{-1} exists because S_W is assumed to be positive definite)

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

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

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 $P(\boldsymbol{x} \mid \omega_1) = \mathcal{N}(\boldsymbol{\mu}, \Sigma)$ 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:

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

$$P(\text{error} \mid x) = \min(P(W_1|x), P(W_2|x))$$

Case 1. if $P(W_1|x) \leq P(W_2|x)$

$$P(\text{error} \mid x) = P(W_1|x) = \sqrt{P(W_1|x)\, P(W_1|x)}$$
$$\leq \sqrt{P(W_1|x)\, P(W_2|x)}$$

case 2. if $P(W_1|x) \geq P(W_2|x)$

$$P(\text{error} \mid x) = P(W_2|x) = \sqrt{P(W_2|x)\, P(W_2|x)}$$
$$\leq \sqrt{P(W_1|x)\, P(W_2|x)}$$

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

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

$$P(\text{error}) = \int_x P(\text{error} \mid x)\, P(x)\, dx$$
$$\leq \int_x \sqrt{P(W_1|x) P(W_2|x)}\, P(x)\, dx$$
$$= \int_x \sqrt{\frac{P(x|W_1)P(W_1)}{\cancel{P(x)}} \frac{P(x|W_2)P(W_2)}{\cancel{P(x)}}}\, \cancel{P(x)}\, dx$$
$$= \sqrt{P(W_1)P(W_2)} \cdot \int_x \sqrt{P(x|W_1)\, P(x|W_2)}\, dx$$
$$= \sqrt{P(W_1)P(W_2)} \int_x \sqrt{\left(\frac{1}{\sqrt{(2\pi)^d \det(\Sigma)}}\right)^2 e^{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)} \cdot e^{-\frac{1}{2}(x+\mu)^T\Sigma^{-1}(x+\mu)}}\, dx$$
$$= \sqrt{P(W_1)P(W_2)}\, \frac{1}{\sqrt{(2\pi)^d \det(\Sigma)}} \int_x \sqrt{e^{-\frac{1}{2}\left[(x-\mu)^T\Sigma^{-1}(x-\mu) + (x+\mu)^T\Sigma^{-1}(x+\mu)\right]}}\, dx$$

specifically we have: $(x-\mu)^T\Sigma^{-1}(x-\mu) + (x+\mu)^T\Sigma^{-1}(x+\mu)$

$$= x^T\Sigma^{-1}x + \mu^T\Sigma^{-1}\mu - \cancel{x^T\Sigma^{-1}\mu} - \cancel{\mu^T\Sigma^{-1}x} + x^T\Sigma^{-1}x + \mu^T\Sigma^{-1}\mu + \cancel{x^T\Sigma^{-1}\mu} + \cancel{\mu^T\Sigma^{-1}x}$$

$$= 2\left(x^T\Sigma^{-1}x + \mu^T\Sigma^{-1}\mu\right)$$

$$P(error) = \sqrt{P(w_1)P(w_2)} \frac{1}{\sqrt{(2\pi)^d \det(\Sigma)}} \int_x \sqrt{e^{-\frac{1}{2} \cdot 2 (x^T\Sigma^{-1}x + \mu^T\Sigma^{-1}\mu)}} \, dx$$

$$= \sqrt{P(w_1)P(w_2)} \cdot e^{-\frac{1}{2}(\mu^T\Sigma^{-1}\mu)} \cdot \underbrace{\int_x \frac{1}{\sqrt{(2\pi)^d \det(\Sigma)}} e^{-\frac{1}{2}x^T\Sigma^{-1}x} \, dx}_{=1}$$

the density of normal distribution $\mathcal{N}(0, \Sigma)$ integrate over the entire space.

$$= \sqrt{P(w_1)P(w_2)} \cdot e^{-\frac{1}{2}(\mu^T\Sigma^{-1}\mu)}$$

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).

$$S_W = S_1 + S_2 = \begin{bmatrix}2 & 0\\0 & 1\end{bmatrix} + \begin{bmatrix}2 & 0\\0 & 1\end{bmatrix} = \begin{bmatrix}4 & 0\\0 & 2\end{bmatrix}$$

$$A = \begin{bmatrix}a & b\\c & d\end{bmatrix}$$

$$A^{-1} = \frac{1}{ad-bc}\begin{bmatrix}d & -b\\-c & a\end{bmatrix}$$

$$S_W^{-1} = \frac{1}{8}\begin{bmatrix}2 & 0\\0 & 4\end{bmatrix} = \begin{bmatrix}\frac{1}{4} & 0\\0 & \frac{1}{2}\end{bmatrix}$$

$$w = S_W^{-1}(\mu_2 - \mu_1) = \begin{bmatrix}\frac{1}{4} & 0\\0 & \frac{1}{2}\end{bmatrix}\begin{bmatrix}1-(-1)\\1-(-1)\end{bmatrix} = \begin{bmatrix}\frac{1}{4} & 0\\0 & \frac{1}{2}\end{bmatrix}\begin{bmatrix}2\\2\end{bmatrix} = \begin{bmatrix}\frac{1}{2}\\1\end{bmatrix}$$

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

$$S_B = (\mu_2 - \mu_1)(\mu_2 - \mu_1)^T = \begin{bmatrix}2\\2\end{bmatrix}\begin{bmatrix}2 & 2\end{bmatrix} = \begin{bmatrix}4 & 4\\4 & 4\end{bmatrix}$$

$$S_B w = \lambda S_W \cdot w$$

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

$$\begin{bmatrix}\frac{1}{4} & 0\\0 & \frac{1}{2}\end{bmatrix}\begin{bmatrix}4 & 4\\4 & 4\end{bmatrix} w = \lambda w$$

$$\begin{bmatrix}1 & 1\\2 & 2\end{bmatrix} w = \lambda w$$

$$\det\left(\begin{bmatrix}1-\lambda & 1\\2 & 2-\lambda\end{bmatrix}\right) = (1-\lambda)(2-\lambda) - 2 = \lambda^2 - 3\lambda + 2 - 2 = 0$$

$$\lambda_1 = 0 \quad \lambda_2 = 3$$

① eigenvector for $\lambda_1 = 0$:

$$\begin{bmatrix}1-0 & 1\\2 & 2-0\end{bmatrix}\begin{bmatrix}v_1\\v_2\end{bmatrix} = \begin{bmatrix}0\\0\end{bmatrix}$$

$$v_1 + v_2 = 0$$
$$2v_1 + 2v_2 = 0$$

$$\Rightarrow \vec{v_1} = \begin{pmatrix}v_1\\-v_1\end{pmatrix} \stackrel{v_1=1}{=\!=} \begin{pmatrix}1\\-1\end{pmatrix}$$

② eigenvector for $\lambda_2 = 3$:

$$\begin{bmatrix}1-3 & 1\\2 & 2-3\end{bmatrix}\begin{bmatrix}v_1\\v_2\end{bmatrix} = \begin{bmatrix}0\\0\end{bmatrix}$$

$$\begin{bmatrix}-2 & 1\\2 & -1\end{bmatrix}\begin{bmatrix}v_1\\v_2\end{bmatrix} = \begin{bmatrix}0\\0\end{bmatrix} \Rightarrow \vec{v_2} = \begin{pmatrix}v_1\\2v_1\end{pmatrix} \stackrel{v_1=\frac{1}{2}}{=\!=} \begin{pmatrix}\frac{1}{2}\\1\end{pmatrix}$$

① $w = \vec{v}_1 = \begin{pmatrix}1\\-1\end{pmatrix}$

$$J_1 = \frac{w^T S_B w}{w^T S_W w} = w^T S_B w = (1 \;\; -1)\begin{bmatrix}8 & 8\\ 8 & 8\end{bmatrix}\begin{bmatrix}1\\-1\end{bmatrix}$$

$$= [0 \;\; 0]\begin{bmatrix}1\\-1\end{bmatrix} = 0$$

② $w = \vec{v}_2 = \begin{pmatrix}\frac{1}{2}\\1\end{pmatrix}$

$$J_2 = [\tfrac{1}{2} \;\; 1]\begin{bmatrix}8 & 8\\ 8 & 8\end{bmatrix}\begin{bmatrix}\frac{1}{2}\\1\end{bmatrix}$$

$$= [12 \;\; 12]\begin{bmatrix}\frac{1}{2}\\1\end{bmatrix} = 18$$

$J_1 < J_2$

→ When we choose the corresponding eigenvalue $\vec{V}_1 = \begin{pmatrix}1\\-1\end{pmatrix}$ as w,

the ratio is minimized