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.

Solution:

The problem can be formulated as

$$\max_{w} \; w^T S_B w \quad \text{s.t.} \quad w^T S_w w = 1 \quad \Longrightarrow \quad \max_{w} \; w^T S_B w \quad \text{s.t.} \quad 1 - w^T S_w w = 0$$

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

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

Solution.

We can formulate the Lagrange function:

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

$$\frac{\partial \mathcal{L}}{\partial w} = S_B w - \lambda S_w w$$

$$\Downarrow \text{ let } \frac{\partial \mathcal{L}}{\partial w} = 0$$

$\therefore \quad S_B w = \lambda S_w w \longrightarrow$ generalized Eigenvalue problem.

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

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

Solution:

日期:

From above we have:

$$S_B w = \lambda S_w w$$

$\Downarrow$ left multiplied by S_w^{-1}

$\therefore$ $$S_w^{-1} S_B w = \lambda S_w^{-1} S_w w$$

$$\Downarrow$$

$$S_w^{-1} S_B w = \lambda I w$$

$\Downarrow$ $S_B = (m_2 - m_1)(m_2 - m_1)^T$

$$S_w^{-1}(m_2 - m_1)\underbrace{(m_2 - m_1)^T w}_{\text{scaler}} = \lambda w$$

$$\text{scaler} \Downarrow \beta$$

$$\Downarrow$$

$$S_w^{-1}(m_2 - m_1) \cdot \beta = \lambda w$$

$\therefore$ suppose the scaling factor is $\lambda^* = \frac{\lambda}{\beta}$

$$\Downarrow$$

$$w^* = \frac{1}{\lambda^*} S_w^{-1}(m_2 - m_1)$$

$\Downarrow$

α (scaling factor)

$\therefore$ $$w^* = \alpha S_w^{-1}(m_2 - m_1)$$

$$\Downarrow$$

$$w^* = S_w^{-1}(m_2 - m_1)$$

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

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

Solution:

From sheet 1 we have the definition:

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

And we know that there is an inequality called generalized mean inequality:

$$\min(x_1, \cdots, x_n) \leq M_p(x_1, \cdots, x_n) \leq \max(x_1, \cdots, x_n)$$

where $M_p(x_1, \cdots, x_n) = \left(\frac{1}{n}\sum_{i=1}^{n} x_i^p\right)^{\frac{1}{p}}$ $\quad p>0$

$$M_0(x_1, \cdots, x_n) = \left(\prod_{i=1}^{n} x_i\right)^{\frac{1}{n}}$$

$\Rightarrow$ equal to geometric mean

$\therefore$ we can derive:

$$\min(P(\omega_1|x), P(\omega_2|x)) \leq M_0(P(\omega_1|x), P(\omega_2|x))$$

$$\Downarrow$$

$$\min(P(\omega_1|x), P(\omega_2|x)) \leq (P(\omega_1|x)P(\omega_2|x))^{\frac{1}{2}}$$

$$\Downarrow$$

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

$$\Downarrow$$

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

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

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

Solution:

$$P(\text{error}) \le \int_x P(\text{error}|x)\, P(x)\, dx$$

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

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

$$= \int_x \sqrt{P(x|\omega_1)P(\omega_1)\, P(x|\omega_2)P(\omega_2)}\, dx$$

$$= \sqrt{P(\omega_1)P(\omega_2)} \int_x \sqrt{P(x|\omega_1)P(x|\omega_2)}\, dx$$

Since $P(x|\omega_1)$ and $P(x|\omega_2)$ have opposite mean and same variance Gaussian.

$$\therefore \int_x \sqrt{P(x|\omega_1)P(x|\omega_2)}\, dx$$

$$= \int_x \sqrt{\frac{1}{{}^{2d}\!\sqrt{2\pi}\,(\det(\Sigma))^2} \exp\left(-\frac{1}{2}\left((x-\mu)^T\Sigma^{-1}(x-\mu) + (x+\mu)^T\Sigma^{-1}(x+\mu)\right)\right)}\, dx$$

$$= \int_x \frac{1}{{}^{d}\!\sqrt{2\pi}\, \det(\Sigma)} \sqrt{\exp\left(-\frac{1}{2}(2x^T\Sigma^{-1}x + 2\mu^T\Sigma^{-1}\mu)\right)}\, dx$$

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

$$= \exp\left(-\frac{1}{2}\mu^T\Sigma^{-1}\mu\right) \cdot \underbrace{\int_x \frac{1}{{}^{d}\!\sqrt{2\pi}\, \det(\Sigma)} \exp\left(-\frac{1}{2}x^T\Sigma^{-1}x\right) dx}_{N(0,\Sigma)}$$

日期: /

$$\therefore = \exp\left(-\frac{1}{2}\mu^T\Sigma^{-1}\mu\right)\cdot 1$$ (integrate the pdf on the whole interval) $=1$

$$\therefore p(\text{error}) \le \sqrt{P(\omega_1)P(\omega_2)}\ \exp\left(-\frac{1}{2}\mu^T\Sigma^{-1}\mu\right)$$

$\therefore$ proofed !

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).

Solution:

$$w^* = S_W^{-1}(m_2 - m_1)$$

$$= (\Sigma_1 + \Sigma_2)^{-1}(m_2 - m_1)$$

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

$$= \begin{bmatrix}4 & 0\\0 & 2\end{bmatrix}^{-1} \cdot \begin{bmatrix}2\\2\end{bmatrix}$$

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

$\therefore$ w^* could be $\begin{bmatrix}\frac{1}{2}\\1\end{bmatrix}$

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

Solution:

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

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

$\therefore$ $S_W^{-1} S_B = \begin{bmatrix}1 & 1\\2 & 2\end{bmatrix} = A$

$\therefore$ $Aw = \lambda w$

$\therefore$ $|\lambda I - A| = (\lambda - 1)(\lambda - 2) - 2 = \lambda^2 - 3\lambda = 0$

$\therefore$ ① $\lambda_1 = 0 \longrightarrow w$ could be $\begin{bmatrix}2\\-2\end{bmatrix}$

② $\lambda_2 = 3 \longrightarrow w$ could be $\begin{bmatrix}1\\2\end{bmatrix}$ or $\begin{bmatrix}2\\4\end{bmatrix}$ ....