Exercises for the course

Machine Learning 1

Winter semester 2021/22

Abteilung 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 4

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

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

(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

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

(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).

(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.

Sheet 4

$\overline{x}$: vector

$\underline{X}$: matrixe

1

a) $$J = \max_{w} \frac{\overline{w}^T \underline{S}_B \overline{w}}{\overline{w}^T \underline{S}_w \overline{w}}$$ with the constraint 1) $\overline{w}^T \underline{S}_w \overline{w} = 1$

$$\stackrel{1)}{=} \max_{w} \frac{\overline{w}^T \underline{S}_B \overline{w}}{1} = \max_{w} \overline{w}^T \underline{S}_B \overline{w}$$

b) $$J(\overline{w}, \lambda) = \overline{w}^T \underline{S}_B \overline{w} - \lambda(\overline{w}^T \underline{S}_w \overline{w} - 1)$$

$$\frac{\partial J}{\partial \overline{w}} = 2 \cdot \underline{S}_B \cdot \overline{w} - 2\lambda \underline{S}_w \cdot \overline{w} \stackrel{!}{=} 0$$

$$\Leftrightarrow \underline{S}_B \cdot \overline{w} = \lambda \underline{S}_w \overline{w}$$

c) $$\Leftrightarrow \frac{1}{\lambda} \underline{S}_B \overline{w} = \underline{S}_w \overline{w}$$

$$\Leftrightarrow \frac{1}{\lambda} \underline{S}_w^{-1} \underline{S}_B \overline{w} = \overline{w}$$

$$\Leftrightarrow \frac{1}{\lambda} \underline{S}_w^{-1} (\overline{m}_1 - \overline{m}_2)(\overline{m}_1 - \overline{m}_2)^T \overline{w} = \overline{w}$$

$$\Leftrightarrow \overline{w} = \text{const.} \cdot \underline{S}_w^{-1} (\overline{m}_1 - \overline{m}_2)$$

$$\Leftrightarrow \overline{w} \propto \underline{S}_w^{-1} (\overline{m}_1 - \overline{m}_2)$$

$\boxed{2}$

a) According to Sheet 1 exercise 1, the Bayes error $P(\text{error}|\bar{x})$ is $\min[P(w_1|\bar{x}), P(w_2|\bar{x})]$ for two classes w_1 and w_2:

$$P(\text{error}|\bar{x}) = \min\left[P(w_1|\bar{x}), P(w_2|\bar{x})\right]$$

$$= M_{-\infty}\left[P(w_1|\bar{x}), P(w_2|\bar{x})\right]$$

with $M_{-\infty}$ being the generalized mean (see solution 1a) of sheet 1)

$$\leq M_0\left[P(w_1|\bar{x}), P(w_2|\bar{x})\right]$$

because $M_p \leq M_q$ if $p < q$

$$= \sqrt{P(w_1|\bar{x})\, P(w_2|\bar{x})}$$

b) According to Eq. 1 in sheet 1:

$$P(\text{error}) = \int P(\text{error}|\bar{x})\, p(\bar{x})\, d\bar{x}$$

according to 2a):

$$\leq \int \sqrt{P(w_1|\bar{x})P(w_2|\bar{x})}\; p(\bar{x})\, d\bar{x}$$

according to Bayes' rule:

$$= \int \sqrt{\frac{P(\bar{x}|w_1)P(w_1)}{\cancel{P(\bar{x})}} \frac{P(\bar{x}|w_2)P(w_2)}{\cancel{P(\bar{x})}}}\; \cancel{p(\bar{x})}\, d\bar{x}$$

$$= \sqrt{P(w_1)P(w_2)} \int \sqrt{P(\bar{x}|w_1)\, P(\bar{x}|w_2)}\, d\bar{x}$$

$$= \sqrt{P(w_1)P(w_2)} \int \sqrt{\mathcal{N}(\bar{\mu}, \underline{\Sigma}) \cdot \mathcal{N}(-\bar{\mu}, \underline{\Sigma})}\, d\bar{x}$$

$$= \sqrt{P(w_1)P(w_2)} \int \sqrt{\frac{e^{-\frac{1}{2}(\bar{x}-\bar{\mu})^T \underline{\Sigma}^{-1}(\bar{x}-\bar{\mu})}}{\sqrt{(2\pi)^d|\underline{\Sigma}|}} \cdot \frac{e^{-\frac{1}{2}(\bar{x}+\bar{\mu})^T \underline{\Sigma}^{-1}(\bar{x}+\bar{\mu})}}{\sqrt{(2\pi)^d|\underline{\Sigma}|}}}\; d\bar{x}$$

$$= \sqrt{P(w_1)P(w_2)} \int \frac{e^{-\bar{x}^T\underline{\Sigma}^{-1}\bar{x}}\; e^{-\bar{\mu}^T\underline{\Sigma}^{-1}\bar{\mu}}}{\sqrt{(2\pi)^d|\underline{\Sigma}|}}\, d\bar{x} = \sqrt{P(w_1)P(w_2)}\, e^{-\frac{1}{2}\bar{\mu}^T\underline{\Sigma}^{-1}\bar{\mu}} \underbrace{\int \frac{e^{-\frac{1}{2}\bar{x}^T\underline{\Sigma}^{-1}\bar{x}}}{\sqrt{(2\pi)^d|\underline{\Sigma}|}}\, d\bar{x}}_{=1}$$

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

3

a) The within-class scatter matrix $\underline{S_W}$ is:

$$\underline{S_W} = \begin{pmatrix} 2 & 0 \\ 0 & 1 \end{pmatrix}$$

The difference between the means is:

$$\bar{\mu}_2 - \bar{\mu}_1 = \begin{pmatrix} 1 \\ 1 \end{pmatrix} - \begin{pmatrix} -1 \\ -1 \end{pmatrix} = \begin{pmatrix} 2 \\ 2 \end{pmatrix}$$

According to 1c):

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

b)

The projection $w = \begin{pmatrix} -1 \\ 1 \end{pmatrix}$ minimizes the ratio

because $w^T \mu_1 = (-1 \;\; 1) \begin{pmatrix} -1 \\ -1 \end{pmatrix} = 0$

and $w^T \mu_2 = (-1 \;\; 1) \begin{pmatrix} 1 \\ 1 \end{pmatrix} = 0$.