Exercises for the course
Machine Learning 1
Winter semester 2018/19

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 6

Exercise 1: Fisher Discriminant (10 + 10 P)

We know from Duda et al., that the Fisher linear discriminant of a two classes dataset with means $\boldsymbol{m}_1, \boldsymbol{m}_2$ and scatter matrices $\boldsymbol{S}_1, \boldsymbol{S}_2$ can be found by maximizing the objective

$$J(\boldsymbol{w}) = \frac{\boldsymbol{w}^\top \boldsymbol{S}_B\, \boldsymbol{w}}{\boldsymbol{w}^\top \boldsymbol{S}_W\, \boldsymbol{w}}$$

where $\boldsymbol{S}_B = (\boldsymbol{m}_1 - \boldsymbol{m}_2)\,(\boldsymbol{m}_1 - \boldsymbol{m}_2)^\top$ is the between-class scatter matrix, and where $\boldsymbol{S}_W = \boldsymbol{S}_1 + \boldsymbol{S}_2$ is the within-class scatter matrix. We would like show how a solution $\boldsymbol{w}$ to this problem can be computed analytically. The derivation of an analytical solution is outlined at page 120 of Duda et al. However, some steps are left to the reader to demonstrate.

(a) *Show* that a vector $\boldsymbol{w}$ that maximizes the objective $J(\boldsymbol{w})$ is also a solution of the generalized eigenvalue problem

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

(b) *Show* that if $\boldsymbol{S}_W$ is invertible, then, a solution for $\boldsymbol{w}$ that maximizes $J(\boldsymbol{w})$ is given by

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

Exercise 2: LDA vs. Optimal Classification (5 + 5 + 5 + 5 + 5 + 5 P)

Consider the case of two classes ω_1 and ω_2 with associated data generating probabilities $p(\boldsymbol{x}|\omega_1)$ and $p(\boldsymbol{x}|\omega_2)$ with $\boldsymbol{x} \in \mathbb{R}^d$ the vector of measurements. Linear discriminants can be generally expressed as:

$$g(\boldsymbol{x}) = \boldsymbol{w}^\top \boldsymbol{x} + b$$

where we decide for class ω_1 if $g(\boldsymbol{x}) > 0$ and ω_2 otherwise.

Linear discriminant analysis (LDA) sets the parameters $\boldsymbol{w}$ and b in terms of the data means $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2$ and covariance Σ assumed to be equal between the two classes:

$$\begin{aligned} \boldsymbol{w} &= \Sigma^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \\ b &= \frac{1}{2}(T + \boldsymbol{\mu}_2 \Sigma^{-1} \boldsymbol{\mu}_2 - \boldsymbol{\mu}_1 \Sigma^{-1} \boldsymbol{\mu}_1) \end{aligned}$$

The parameter T is a threshold to be determined, and which is typically adjusted to the class prior probabilities. LDA can be shown to be optimal when data for each class is Gaussian distributed.

Assume that the data-generating distributions are two-dimensional and *non-Gaussian*, with probability densities:

$$\begin{aligned} p(\boldsymbol{x}|\omega_1) &= \frac{1}{16} \cdot 1_{\{0 \le x_1 \le 4\}} \cdot 1_{\{-1 \le x_2 \le 3\}} \\ p(\boldsymbol{x}|\omega_2) &= \frac{1}{16} \cdot 1_{\{-4 \le x_1 \le 0\}} \cdot 1_{\{-3 \le x_2 \le 1\}} \end{aligned}$$

where $1_{\{\cdot\}}$ is an indicator function. Assume prior probabilities $P(\omega_1) = P(\omega_2) = 0.5$.

(a) *Compute* the means $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2$ and covariances $\boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2$ associated to these two distributions.

(b) *Derive* the LDA discriminant for this dataset (i.e. express $\boldsymbol{w}$ and b).

(c) *Plot* the expected error of the LDA discriminant as a function of the threshold parameter T.

(d) *Determine* what is the Bayes error rate associated to this dataset.

(e) *Explain* whether the LDA discriminant is able to reach the Bayes error rate.

(f) *Find* an alternate linear discriminant (i.e. parameters $\boldsymbol{w}$ and b) that reaches the Bayes error rate on this data distribution.

Exercise 3: Programming (50 P)

Download the programming files on ISIS and follow the instructions.