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 $\omega_1$ and $\omega_2$ 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.

# Fisher Linear Discriminant

In this exercise, we apply Fisher Linear Discriminant as described in Chapter 3.8.2 of Duda et al. on the UCI Abalone dataset. A description of the dataset is given at the page https://archive.ics.uci.edu/ml/datasets/Abalone (https://archive.ics.uci.edu/ml/datasets/Abalone). The following two methods are provided for your convenience:

- **`utils.Abalone.__init__(self)`** reads the Abalone data and instantiates two data matrices corresponding to: *infant (I)*, *non-infant (N)*.
- **`utils.Abalone.plot(self,w)`** produces a histogram of the data when projected onto a vector `w`, and where each class is shown in a different color.

Sample code that makes use of these two methods is given below. It loads the data, looks at the shape of instantiated matrices, and plots the projection on the first dimension of the data representing the length of the abalone.

In [1]:

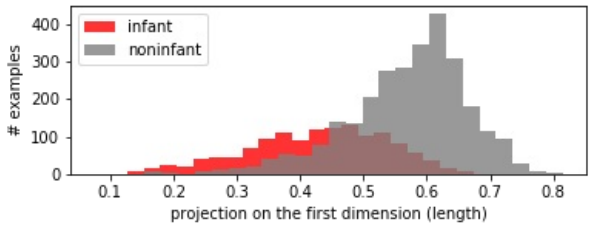
```
%matplotlib inline
import utils,numpy

# Load the data
abalone = utils.Abalone()

# Print dataset size for each class
print(abalone.I.shape, abalone.N.shape)

# Project data on the first dimension
w1 = numpy.array([1,0,0,0,0,0,0])
abalone.plot(w1,'projection on the first dimension (length)')
```

```
(1342, 7) (2835, 7)
```




## Implementation (10 + 5 + 5 = 20 P)

- **Create a function `w = fisher(X1,X2)` that takes as input the data for two classes and returns the Fisher linear discriminant.**
- **Create a function `objective(X1,X2,w)` that evaluates the objective defined in Equation 96 of Duda et al. for an arbitrary projection vector `w`.**
- **Create a function `z = phi(X)` that returns a quadratic expansion for each data point `x` in the dataset. Such expansion consists of the vector `x` itself, to which we concatenate the vector of all pairwise products between elements of `x`.** In other words, letting $x = (x_1, \ldots, x_d)$ denote the $d$-dimensional data point, the quadratic expansion for this data point is a $d \cdot (d+3)/2$ dimensional vector given by $\phi(x) = (x_i)_{1 \le i \le d} \cup (x_i x_j)_{1 \le i \le j \le d}$. For example, the quadratic expansion for $d = 2$ is $(x_1, x_2, x_1^2, x_2^2, x_1 x_2)$.

In [2]:

```
def fisher(X1,X2):
    ##### Replace by your code
    import solutions
    return solutions.fisher(X1,X2)
    #####

def objective(X1,X2,w):
    ##### Replace by your code
    import solutions
    return solutions.objective(X1,X2,w)
    #####

def expand(X):
    ##### Replace by your code
    import solutions
    return solutions.expand(X)
    #####
```

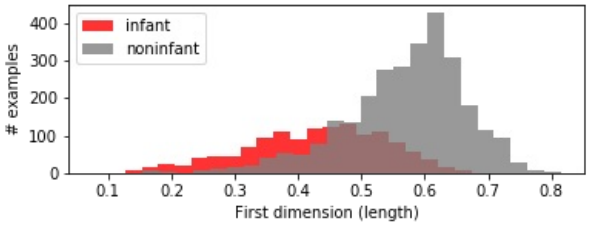
## Analysis (5 + 5 = 10 P)

- **Print value of the objective function and the histogram for several values of `w` :**
  - `w` is a canonical coordinate vector for the first feature (length).
  - `w` is the difference between the mean vectors of the two classes.
  - `w` is the Fisher linear discriminant.
  - `w` is the Fisher linear discriminant (after quadratic expansion of the data).

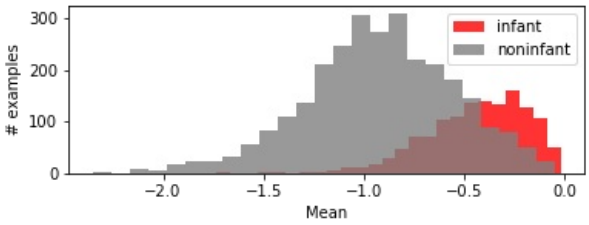
In [3]:

```
##### REPLACE BY YOUR CODE
%matplotlib inline
import solutions
solutions.analysis()
#####
```

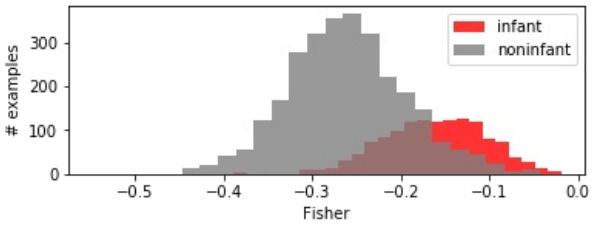
First dimension (length):  0.00048



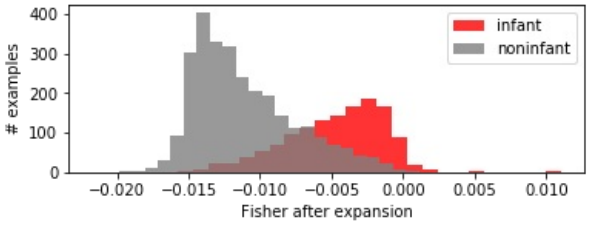

Means Linear:  0.00050




Fisher:  0.00057




Fisher after expansion:  0.00077



# Sheet 4

$\bar{x}$ : vector

$\underline{X}$ : matrix

## 1

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

$$\overset{1)}{=} \max_{\bar{w}} \frac{\bar{w}^T \underline{S}_B \bar{w}}{1} = \max_{\bar{w}} \bar{w}^T \underline{S}_B \bar{w}$$

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

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

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

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

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

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

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

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

2]

a) According to Sheet 1 exercise 1, the Bayes error $P(\text{error}|\vec{x})$ is $\min[P(\omega_1|\vec{x}), P(\omega_2|\vec{x})]$ for two classes $\omega_1$ and $\omega_2$:

$$P(\text{error}|\vec{x}) = \min\left[P(\omega_1|\vec{x}), P(\omega_2|\vec{x})\right]$$

$$= M_{-\infty}\left[P(\omega_1|\vec{x}), P(\omega_2|\vec{x})\right]$$

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

$$\leq M_0\left[P(\omega_1|\vec{x}), P(\omega_2|\vec{x})\right]$$

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

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

b) According to Eq. 1 in sheet 1:

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

according to 2 d):

$$\leq \int \sqrt{P(\omega_1|\vec{x})\, P(\omega_2|\vec{x})}\; p(\vec{x})\, d\vec{x}$$

according to Bayes' rule:

$$= \int \sqrt{\frac{P(\vec{x}|\omega_1)P(\omega_1)}{P(\vec{x})} \cdot \frac{P(\vec{x}|\omega_2)P(\omega_2)}{P(\vec{x})}}\; p(\vec{x})\, d\vec{x}$$

$$= \sqrt{P(\omega_1)P(\omega_2)} \int \sqrt{P(\vec{x}|\omega_1)\, P(\vec{x}|\omega_2)}\, d\vec{x}$$

$$= \sqrt{P(\omega_1)P(\omega_2)} \int \sqrt{\mathcal{N}(\vec{\mu}, \underline{\Sigma}) \cdot \mathcal{N}(-\vec{\mu}, \underline{\Sigma})}\, d\vec{x}$$

$$= \sqrt{P(\omega_1)P(\omega_2)} \int \sqrt{\frac{e^{-\frac{1}{2}(\vec{x}-\vec{\mu})^T \underline{\Sigma}^{-1}(\vec{x}-\vec{\mu})}}{\sqrt{(2\pi)^d |\underline{\Sigma}|}} \cdot \frac{e^{-\frac{1}{2}(\vec{x}+\vec{\mu})^T \underline{\Sigma}^{-1}(\vec{x}+\vec{\mu})}}{\sqrt{(2\pi)^d |\underline{\Sigma}|}}}\, d\vec{x}$$

$$= \sqrt{P(\omega_1)P(\omega_2)} \int \frac{e^{-\vec{x}^T \underline{\Sigma}^{-1} \vec{x}}\; e^{-\vec{\mu}^T \underline{\Sigma}^{-1} \vec{\mu}}}{\sqrt{(2\pi)^d |\underline{\Sigma}|}}\, d\vec{x} = \sqrt{P(\omega_1)P(\omega_2)}\, e^{-\frac{1}{2}\vec{\mu}^T \underline{\Sigma}^{-1} \vec{\mu}} \int \underbrace{\frac{e^{-\frac{1}{2}\vec{x}^T \underline{\Sigma}^{-1} \vec{x}}}{\sqrt{(2\pi)^d |\underline{\Sigma}|}}}_{\mathcal{N}(\vec{0}, \underline{\Sigma})}\, d\vec{x}$$

$$= \sqrt{P(\omega_1)P(\omega_2)}\, e^{-\frac{1}{2}\vec{\mu}^T \underline{\Sigma}^{-1} \vec{\mu}}$$

3

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

$$\underline{S_w} = \begin{pmatrix} 4 & 0 \\ 0 & 1 \end{pmatrix}$$

The difference between the means is:

$$\mu_2 - \mu_1 = \begin{pmatrix} 1 \\ 1 \end{pmatrix} - \begin{pmatrix} -1 \\ -1 \end{pmatrix} = \begin{pmatrix} 2 \\ 2 \end{pmatrix}$$

According to 1c):

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

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