Time	Group	Submission in Moodle; Mails with subject: [SMD2023]
Th. 12:00–13:00	A	lukas.beiske@udo.edu and tristan.gradetzke@udo.edu
Fr. 08:45–09:45	B	jonas.hackfeld@ruhr-uni-bochum.de and ludwig.neste@udo.edu
Fr. 10:00–11:00	C	stefan.froese@udo.edu and vincent.latko@udo.edu

Exercise 12 *Fisher-Discriminant: By Hand* **3 p.**

Perform a Linear Fisher Discriminant analysis by hand.
Population 0: (1; 1) (2; 1) (1.5; 2) (2; 2) (2; 3) (3; 3)
Population 1: (1.5; 1) (2.5; 1) (3.5; 1) (2.5; 2) (3.5; 2) (4.5; 2)

(a) Calculate the mean $\vec{\mu}$ and scatter matrices S_i, as well as the combined scatter matrix S_{ij}.

(b) What is $\vec{\lambda}$?

(c) Draw the points of the two populations on a graph along with the projection line $\vec{\lambda} = \lambda \cdot \vec{e}_{\vec{\lambda}}$. Make sure the aspect ratio of your plot is equal (`aspect('equal')`).

(d) Project the individual points onto the projection line.

(e) Choose a suitable parameter λ_{cut} and calculate the corresponding precision and recall with respect to population 1. Why did you choose this parameter?

Exercise 13 *Fisher-Discriminant: Implementation* **7 p.**

In the file `two_populations.h5` the two populations `P_0_10000` and `P_1` are given. (You can find this file in Moodle.) Read the file using the command:

Listing 1: Example to read the populations:

```
import pandas as pd

p0 = pd.read_hdf('two_populations.h5', key='P_0_10000')
p1 = pd.read_hdf('two_populations.h5', key='P_1')

p0_1000 = pd.read_hdf('two_populations.h5', key='P_0_1000')
```

Note: It is allowed to use packages for linear algebra, however, no functions that perform discriminance analysis may be used.

(a) Calculate the mean μ_{P0} and μ_{P1} of both populations.

(b) Calculate the covariance matrices V_{P0} and V_{P1} of both populations and combine these to obtain the matrix $V_{P0,P1}$.

(c) Construct a Linear Fisher Discriminant $\vec{\lambda} = \lambda \cdot \vec{e}_{\vec{\lambda}}$. State this linear equation.

(d) Plot the populations as a projection onto the line from **(c)** in a one-dimensional histogram.

(e) Consider $P0$ as the signal and $P1$ as the background. Calculate the precision and the recall of the signal as functions of a cut λ_{cut} in λ and plot the results.

(f) At what value of λ_{cut} does the signal to background ratio S/B have its maximum after separation? Create a plot for this as well.

(g) At what value of λ_{cut} is the significance $S/\sqrt{S+B}$ maximal? Create a plot for this as well.

(h) Repeat the steps **(a)** to **(g)** for the case that $P0$ denotes the population `P_0_1000`. What do you notice? Interpret the results.