SoSe 2020
Dr. Joscha Reimer

Kiel University
Department of Computer Science

Optimization and Data Science
9. Homework exercises

Programming exercise 1:
The "D'Agostino and Pearson's normaltest" is a hypothesis test which tests the hypothesis that values are the realization of normal distributed random samples. It is also available in SciPy at scipy.stats.normaltest. Apply this test to the data

a) "data1.txt"

b) "data2.txt"

from programming exercise 1 of homework exercises 8 using a significance level of 0.05. Does the test say that the data originates from a normal distribution or not?

Programming exercise 2:
Write a function $normal(n, eta, sigma)$ which returns an array with n values which are realizations of normal distributed random samples with expectation value eta and standard deviation sigma.

You are allowed to use an already implemented function to generate uniform distributed values, but you are not allowed to use an already implemented function to generate normal distributed values.

Use the hypothesis test mentioned in the previous exercise to test your implementation.

Hint: Use the Box-Muller algorithm to generate standard-normal-distributed values from uniform distributed values.

Programming exercise 3:
The file "Iris_data.csv" contains data from three different species of iris flowers. The data collected are sepal length, sepal width, petal length and petal width. You should investigate whether the data can be used to reliably distinguish between these species. Proceed as follows:

a) Calculate the principal components of the data.

b) Plot the first and second principal component in a 2D scatter plot. Mark which values belong to which species. Interpret you plot. Do you think that the first two principal components can be used to distinguish all species?

c) Plot the first, second and third principal component in a 3D scatter plot. Mark which values belong to which species. Interpret you plot. Do you think that the first three principal components can be used to distinguish all species?

d) *Calculate what proportion of the variance is described by the first two and the first three principal components, respectively. Interpret you results. Do you think that the iris flowers data (i. e. all four principal components) can be used to distinguish all species?*

The iris flower data set is a very famous and often for test purposes used application data set in data science

The solutions of the theoretical exercises will be discussed on 08. June 2020.