Heinrich Heine Universität Düsseldorf

HEIDERLAB

Exercise 2 on Machine Learning WS 2023/24
Prof. Dr. Dominik Heider
M. Sc. Arsam (Mohammad) Tajabadi
Submission: Until 08.11.2023, 23:59 on Ilias

Information

Familiarize yourself with the Python libraries Pandas, Numpy, SciPy and Matplotlib.

Task 1. Key figures (2 points)

Calculate the relevant statistics using Python for the values 4, 2, 5, 6, 1, 6, 8, 3, 4, 9 and answer the questions: (0.5 P. each)

a) What is the standard deviation?

b) Are the data skewed to the right or left?

c) What are the values of the Quartiles?

d) Are there any outliers? If yes, which ones?

Task 2. Correlation (6 points)

a) Write a function that calculates the Pearson correlation coefficient between two variables. In this subtask, you are not allowed to use any packages. (2.5 P.)

b) Import the file *icecream.csv* in the folder *data* as a table in a Python editor. Examine the data and perform Data Cleaning if necessary.
What needs to be considered before the correlation coefficient can be calculated? (1.5 P.)

c) Calculate the correlation coefficient between *Temperature* and *Sales*. Interpret the result. (1 P.)

d) Plot *Temperature* and *Sales* in an appropriate way to see the correlation. (1 P.)

Task 3. Variance (2 points)

Show that the variance of a data set with n entries can be calculated using the following formula:

$$Var(X) = E(X^2) - E(X)^2 \tag{1}$$