Kiel University
Christian-Albrechts-Universität zu Kiel

Exercise Sheet 3
Intelligent Systems
November 11, 2019

Preprocessing

Exercise 1 - Preprocessing: Characteristica

A. Describe the process of preprocessing.

B. Why do we need the preprocessing step?

C. How to deal with *missing values* and *outliers*?

D. Give a short description of how to detect them automatically. Apply your methodology to the points given in Figure 1

E. What are reasons for applying standardisation and normalisation to raw data?

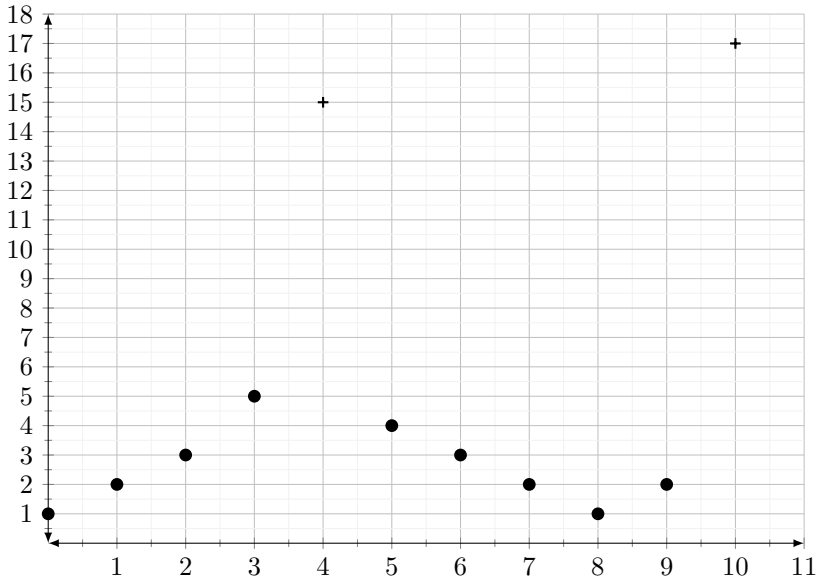

Figure 1: Point sequence with outliers at $x = 4$ and $x = 10$

Exercise 2 - Preprocessing with Python I

Download the file *usStatesData.csv* and the Python Jupyter Notebook *03_Preprocessing.ipynb* from OpenOlat. The *csv*-file contains characteristics about the *states* and *territories* of the US. Your task is to analyse these states regarding their political direction. You can find a column *Vote*, which represents the results of the elections in the year 2016.
Start by loading the *csv*-file into a Pandas DataFrame and detect possible missing values. Treat the missing values by using one of the following methods:

- Delete all rows with missing values
- Replace missing values with a constant (for example "?")
- Replace missing values on your own with the usage of domain knowledge
- Replace missing values with the class mean.

After having treated missing values, plot the data with the given plot function. What attracts your attention? Consider a suitable transformation for the data and replot the data again.