Kiel University
Christian-Albrechts-Universität zu Kiel

Faculty of Engineering

Intelligent Systems

Excersice 3- Preprocessing

Simon Reichhuber

November 11, 2019

University of Kiel, Winter Term 2019

Table of Content

1. Preprocessing: Characteristica
2. Preprocessing with Python

Preprocessing: Characteristica

Question 1A

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 the Figure below.

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

1A Preprocessing Definition

Definition: Preprocessing

Preprocessing is an important step before Machine Learning can be applied. In detail, it is about concerning the existance of erroneous or missing data (i.e. identifying and elimination of outliers or noise) and finding an adequate data representation (for example: the choice of variable types or the modelling of unknown or missing data).

Question 1B

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?

1B Why preprocessing? Error types

Preprocessing is very important for producing reasonable results in Machine Learning!

Data can be uncomplete: Missing values, missing feautres of various data sources

Data can contain noise: Measurement errors, unimportant outliers

Data can be inconsistent: Contradictory measurements of different data sources, occasionally “only” different scaling.

1B Why preprocessing? Error reasons

Reasons:

- Breakdown of sensors whilst measuring physical units
- Receiving or transmission errors (e.g. GPS in the underground)
- Irrelevant feauture for the objective
- Change of the experimental setting
- Merging different data sets

1B Why preprocessing? Preprocessing Tasks

Adjustment

- Treatment of missing values (e.g. replace them)
- Detection and treatment of outliers
- Remove inconsistencies

Integration: Merge information from different data sources

Transformation: Normalise, aggregate, or change the mathematical basis

Reduction: Preferably without (or nearly without) loss of information

Question 1C

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?

1C Adjustment

Treatment of missing values or outliers:

1. Sample (data points) with missing values will not be taken into account (deleted)
   → Only applicable if only a few data points are affected, not applicable for time series
2. Missing values will be propagated to the machine learning model
3. Missing values will be replaced by:
   - the mean.
   - the most frequent value.
   - an estimation inferred from values of other features.
   - repeating the last valid value.
   - replacing the missing values by a global constant, e.g. “?”.
   - replacing the missing values with the help of domain knowledge.
   - interpolation of timeseries data.

Question 1D

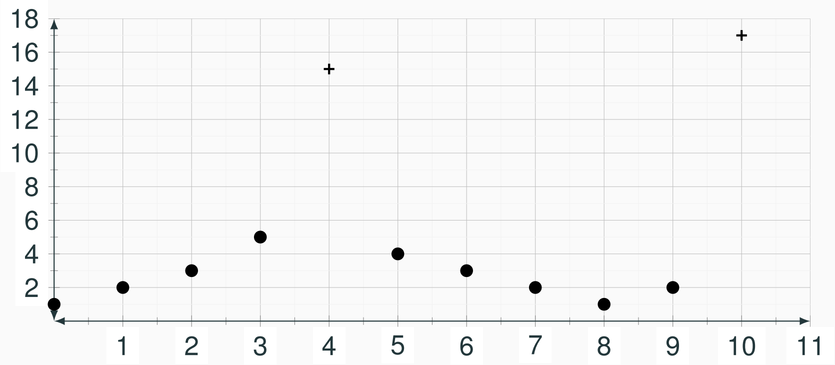
A. Describe the process of preprocessing.

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?



1. D Outlier detection - Problem

Problem: How to distinguish between outliers and exotic samples?

Definition: exotic sample

Correct meassurement of unusual or exceptional phenomena. The so called *exotic sample* can carry important information.

1. D Outlier detection - Solution

Detection of outliers:

1. Value of at least one feature exceeds the valid range.
2. Distance to the mean of the data of a certain feauture is larger than two or three times the standard deviation.
   $\longrightarrow$ 03_Preprocessing.ipynb
3. Distance to a model-based value is larger than a predefined threshold.

1. D Outlier detection - Mean & Std

Given:

Points $x_1, \ldots, x_N$

Searched:

$f : \mathbb{R} \to \{0, 1\}$

$$f(x) = \begin{cases} 1, & \text{for x is outlier} \\ 0, & \text{otherwise} \end{cases}$$

1. D Outlier detection - Mean & Std

Given:

Points $x_1, \dots, x_N$

Searched:

$f : \mathbb{R} \to \{0, 1\}$

$$f(x) = \begin{cases} 1, & \text{for x is outlier} \\ 0, & \text{otherwise} \end{cases}$$

Tools:

- **Mean:** $\bar{x} = \frac{1}{n} \sum_{i=1}^{N} x_i$
- **Standard Deviation:** $\sigma = \sqrt{\frac{1}{N-1} \sum_{i=1}^{N} (x_i - \bar{x})^2}$
- **signum:** $\text{sign}(x) = \begin{cases} 1, & \text{for } x > 0 \\ 0, & \text{for } x = 0 \\ -1, & \text{for } x < 0 \end{cases}$

1. D Outlier detection - Mean & Std

Given:

Points $x_1, \ldots, x_N$

Searched:

$f : \mathbb{R} \rightarrow \{0, 1\}$

$$f(x) = \begin{cases} 1, & \text{for x is outlier} \\ 0, & \text{otherwise} \end{cases}$$

Tools:

- **Mean:** $\bar{x} = \frac{1}{n} \sum_{i=1}^{N} x_i$
- **Standard Deviation:** $\sigma = \sqrt{\frac{1}{N-1} \sum_{i=1}^{N} (x_i - \bar{x})^2}$
- **signum:** $\text{sign}(x) = \begin{cases} 1, & \text{for } x > 0 \\ 0, & \text{for } x = 0 \\ -1, & \text{for } x < 0 \end{cases}$

Solution:

$\rightarrow f(x) = \max(\text{sign}(|x - \bar{x}| - 2 \times \sigma), 0)$

QUESTION 1E

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?

1E Reasons for Normalisation & Standardisation

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

Problem: Different domains of feautures

For example:

1. Measurement of temperature values:
   - Interpretable physical units, like *Celsius*, *Fahrenheit*, or *Rankine*
   - Comparison fails since domains differ (different reference system or basis, etc.)
2. Smartphone user's behavior:
   - User 1: 25000 KB, 12 minutes, 19.99$, 18 messages
   - User 2: 17000 KB, 10 minutes, 19.99$, 20 messages
   - Calculation of point sample distance beteen User 1 and User 2 makes no sence, since the storage size dominates other features

Solution: Normalisation & Standardisation

Preprocessing with Python