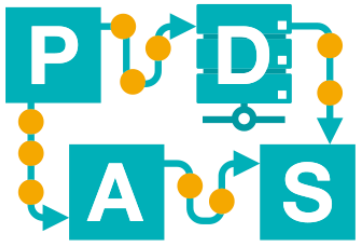
# Data Quality & Preprocessing

*(Tsung-Hao Huang)*

IDS-I-L17




Chair of Process and Data Science

RWTH AACHEN UNIVERSITY

# Outliers & Missing Values

## Outlier detection techniques based on previous lectures:

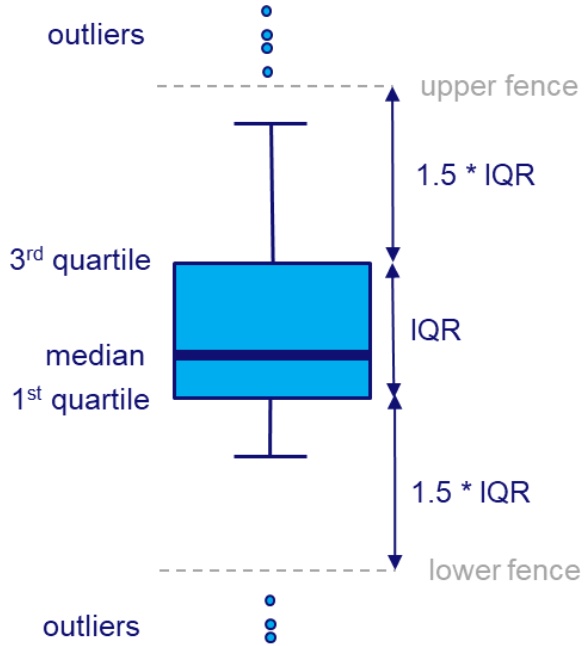
- Boxplots (Lecture 2)
- Decision Trees (Lecture 3)
- Regression (Lecture 4)
- SVMs (Lecture 5)
- Clustering (Lecture 9)

# Outliers & Missing Values

**We can handle outliers/missing values in several ways, depending on our needs, e.g.:**

- **Ignore the feature or instance**
- **Fill in the correct value (domain knowledge)**
- **Fill in a value based on other data**
  - **Mean, median, min, max, most frequent... (possibly focus on similar data)**
  - **Prediction model (regression, SVM, decision tree, NN, …)**

# Outliers & Missing Values - Boxplots




**Instances outside the fences can be considered outliers.**

**See Instruction 3 for exercises.**

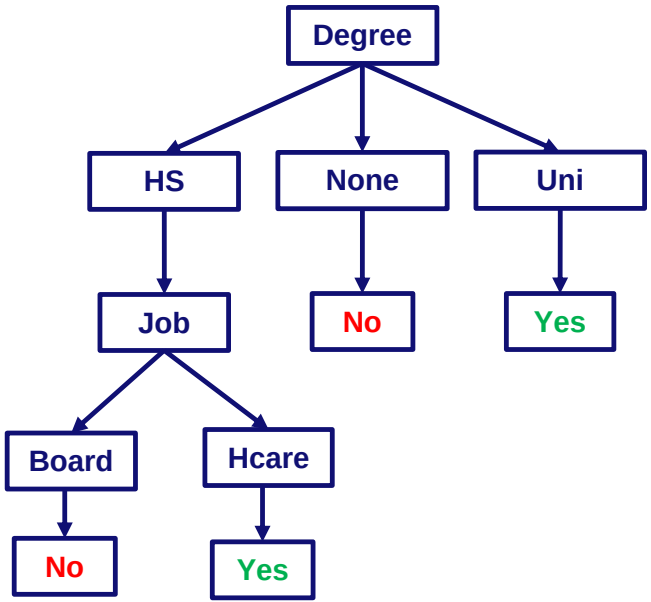



Chair of Process and Data Science

# Outliers & Missing Values – Decision trees

| | Experience | Degree | Job | Class |
|---|---|---|---|---|
| 1 | Exp >10 | HS | Board | No |
| 2 | 5< Exp <10 | Uni | Board | Yes |
| 3 | Exp >10 | HS | Board | No |
| 4 | 5< Exp <10 | HS | Hcare | Yes |
| 5 | Exp < 5 | HS | Hcare | Yes |
| 6 | Exp < 5 | HS | Board | No |
| 7 | Exp < 5 | None | Edu | No |
| 8 | Exp >10 | None | Hcare | No |
| 9 | Exp < 5 | Uni | Edu | Yes |
| 10 | Exp >10 | Uni | Board | Yes |
| 11 | Exp >10 | HS | Board | Yes |

Consider the following data about accepting or rejecting job applications based on “Experience”, “Degree”, and “Job”, as well as the corresponding decision tree (extension of Instruction 3).



# Outliers & Missing Values – Decision trees

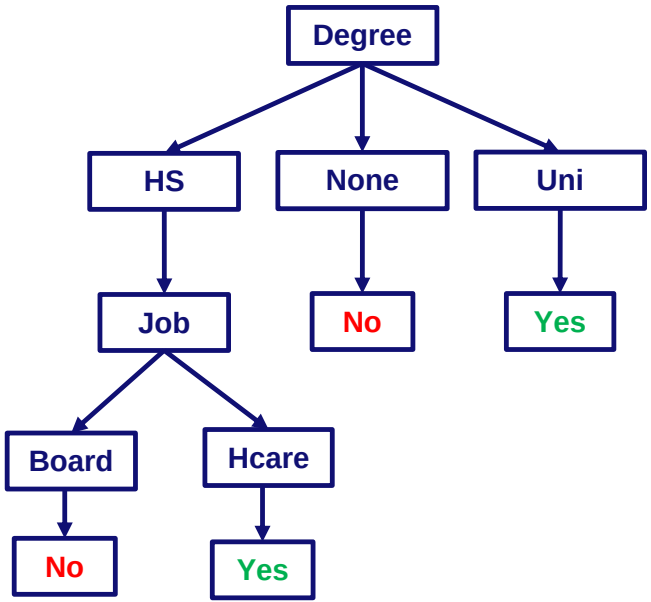
| | Experience | Degree | Job | Class |
|---|---|---|---|---|
| 1 | Exp >10 | HS | Board | No |
| 2 | 5< Exp <10 | Uni | Board | Yes |
| 3 | Exp >10 | HS | Board | No |
| 4 | 5< Exp <10 | HS | Hcare | Yes |
| 5 | Exp < 5 | HS | Hcare | Yes |
| 6 | Exp < 5 | HS | Board | No |
| 7 | Exp < 5 | None | Edu | No |
| 8 | Exp >10 | None | Hcare | No |
| 9 | Exp < 5 | Uni | Edu | Yes |
| 10 | Exp >10 | Uni | Board | Yes |
| 11 | Exp >10 | HS | Board | Yes |

**Exercise 1**

Consider the given the data table and corresponding decision tree.

a) Which instances can be identified as outliers based on the given decision tree?

b) Using the decision tree as predictive model to replace the outliers target features values, what would be the resulting data table?



# Outliers & Missing Values – Regression

| Age [months] | Weight [kg] | Size [cm] |
|---|---|---|
| 0 | 2.5 | 46.4 |
| 0 | 3.8 | 52.9 |
| 1 | 3.2 | 50.4 |
| 3 | 6.7 | 63.4 |
| 9 | 7.5 | 68 |
| 9 | 10.2 | 75 |
| 12 | 8.4 | 71.7 |
| 12 | 11.2 | 79.1 |
| 18 | 9.6 | 77.5 |
| 18 | 12.8 | 86.1 |
| 24 | 10.5 | 82.3 |
| 24 | 14.1 | 92 |

$$\mathbb{M}_{\mathbf{w}}(d) = -8.4 + 1 \times \text{WEIGHT} + 0.12 \times \text{SIZE}$$

**Exercise 2**

Consider the given baby data and corresponding regression function $\mathbb{M}_{\mathbf{w}}(d)$, predicting *Age* based on *Weight* and *Size*.

a) For an error threshold of 3, (using the absolute error), which data points will be identified as outliers?

b) Use the regression function to replace the outliers target feature values by the predicted values.

# Outlier Detection - SVMs

| Weight [kg] | Age [months] | Class |
|---|---|---|
| 2.5 | 0 | underweight |
| 4.2 | 0 | overweight |
| 2.4 | 0 | underweight |
| 3.8 | 0 | overweight |
| 3.2 | 1 | underweight |
| 5.4 | 1 | overweight |
| 3 | 1 | underweight |
| 4.9 | 1 | overweight |
| 4.4 | 3 | underweight |
| 7.4 | 3 | overweight |
| 4.2 | 3 | underweight |
| 6.7 | 3 | overweight |
| 6.2 | 6 | underweight |
| 9.5 | 6 | overweight |
| 5.8 | 6 | underweight |
| 8.7 | 6 | overweight |

**Exercise 3**

Consider the given data table and corresponding separating hyperplane.

a) Based on this hyperplane, which instances can be considered outliers?

b) Replace the outliers target feature values by the values predicted by the SVM.

$$0 = 0.4 \cdot \text{AGE} - \text{WEIGHT} + 4.0$$
$$\Leftrightarrow (w_1, w_2) = (-1.0, 0.4), b = 4.0$$

# Outlier Detection - Clustering

| Weight [kg] | Age [months] | Class |
|---|---|---|
| 2.5 | 0 | underweight |
| 4.2 | 0 | overweight |
| 2.4 | 0 | underweight |
| 2.7 | 0 | overweight |
| 3.2 | 1 | underweight |
| 5.4 | 1 | overweight |
| 3 | 1 | underweight |
| 4.9 | 1 | overweight |
| 4.4 | 3 | underweight |
| 7.4 | 3 | overweight |
| 4.2 | 3 | underweight |
| 6.7 | 3 | overweight |
| 6.2 | 6 | underweight |
| 9.5 | 6 | overweight |
| 5.8 | 6 | underweight |
| 8.7 | 6 | overweight |

## Exercise 4

Consider the given data clustered into the blue, red, yellow and purple clusters. Which instances can be considered outliers?

# Transformation

**Getting the right data type, e.g.**

- **One-hot encoding: categorical to numerical**
- **Binning: numerical to categorical (ordinal)**

**Known from earlier lectures!**

# Transformation - Binning

## Introduced in Lecture 2!

## Many details can vary in real applications:

- What to do, if there is not enough data to ‘fill the last bin’?
- What to do, if the data is not dividable into bins according to requested parameters?
- Do we define interval limits closed or open ?
- …

# Transformation - Binning

| Weight [kg] |
|---|
| 2.5 |
| 3.8 |
| 3.2 |
| 4.9 |
| 4.4 |
| 6.7 |
| 6.2 |
| 7.5 |
| 8.4 |
| 9.6 |
| 12.8 |
| 10.5 |
| 14.1 |

## Exercise 5

Apply equal-width binning to the given data. Your lowest bin interval should have 0 as open lower limit, and your highest interval should have 16 as closed upper limit.

1) Give a mapping of the data to two bins. Indicate the interval limits as well as the bins boundary values.
2) Give a mapping of the data to four bins. Indicate the interval limits as well as the bins boundary values.

# Transformation - Binning

| Weight [kg] |
|---|
| 2.5 |
| 3.2 |
| 3.8 |
| 4.4 |
| 4.9 |
| 6.2 |
| 6.7 |
| 7.5 |
| 8.4 |
| 10.5 |
| 12.8 |
| 14.1 |

**(sorted)**

## Exercise 6

Apply equal-frequency binning to the given data.

1) Give a mapping of the data to two bins. Indicate the interval limits as well as the bins boundary values.
2) Give a mapping of the data such that the bin size is four. Indicate the interval limits as well as the bins boundary values.

# Normalization

**Adjusting the influence of features.**

**Introduced in the lecture:**

- **Min-Max Normalization**
- **Standard Score Normalization**
- **Decimal Scaling**

# Normalization – Min-Max

| Age [months] | Weight [kg] | Size [cm] |
|---|---|---|
| 0 | 2.5 | 46.4 |
| 0 | 3.8 | 52.9 |
| 1 | 3.2 | 50.4 |
| 1 | 4.9 | 56.9 |
| 3 | 4.4 | 56.7 |
| 3 | 6.7 | 63.4 |
| 6 | 6.2 | 63.4 |
| 6 | 8.7 | 70.2 |
| 9 | 7.5 | 68 |
| 9 | 10.2 | 75 |
| 12 | 8.4 | 71.7 |
| 12 | 11.2 | 79.1 |
| 18 | 9.6 | 77.5 |
| 18 | 12.8 | 86.1 |
| 24 | 10.5 | 82.3 |
| 24 | 14.1 | 92 |

## Exercise 7

Apply Min-Max Normalization for each feature of the given data with a target interval of [1,10].

# Normalization – Standard Score

| Age [months] | Weight [kg] | Size [cm] |
|---|---|---|
| 0 | 2.5 | 46.4 |
| 0 | 3.8 | 52.9 |
| 1 | 3.2 | 50.4 |
| 1 | 4.9 | 56.9 |
| 3 | 4.4 | 56.7 |
| 3 | 6.7 | 63.4 |
| 6 | 6.2 | 63.4 |
| 6 | 8.7 | 70.2 |
| 9 | 7.5 | 68 |
| 9 | 10.2 | 75 |
| 12 | 8.4 | 71.7 |
| 12 | 11.2 | 79.1 |
| 18 | 9.6 | 77.5 |
| 18 | 12.8 | 86.1 |
| 24 | 10.5 | 82.3 |
| 24 | 14.1 | 92 |

## Exercise 8

Scale each feature of the given data points using Standard Score Normalization.

# Normalization – Decimal Scaling

| Age [months] | Weight [kg] | Size [cm] |
|---|---|---|
| 0 | 2.5 | 46.4 |
| 0 | 3.8 | 52.9 |
| 1 | 3.2 | 50.4 |
| 1 | 4.9 | 56.9 |
| 3 | 4.4 | 56.7 |
| 3 | 6.7 | 63.4 |
| 6 | 6.2 | 63.4 |
| 6 | 8.7 | 70.2 |
| 9 | 7.5 | 68 |
| 9 | 10.2 | 75 |
| 12 | 8.4 | 71.7 |
| 12 | 11.2 | 79.1 |
| 18 | 9.6 | 77.5 |
| 18 | 12.8 | 86.1 |
| 24 | 10.5 | 82.3 |
| 30 | 17.1 | 102 |

## Exercise 9

Scale each feature of the given data points using Decimal Scaling.

# Reduction

**Making the data smaller for analysis, but produce the same or similar analysis results.**

- **Two types of reduction:**
  - **Feature reduction (keep all the rows, but reduce the number of columns in dataset)**
  - **Instance reduction (store summary, remove or reduce the number of rows)**

# Reduction - Aggregation

| Weight [kg] | Age [months] | Class |
|---|---|---|
| 2.5 | 0 | underweight |
| 4.2 | 0 | overweight |
| 2.4 | 0 | underweight |
| 2.7 | 0 | overweight |
| 3.2 | 1 | underweight |
| 5.4 | 1 | overweight |
| 3 | 1 | underweight |
| 4.9 | 1 | overweight |
| 4.4 | 3 | underweight |
| 7.4 | 3 | overweight |
| 4.2 | 3 | underweight |
| 6.7 | 3 | overweight |
| 6.2 | 6 | underweight |
| 9.5 | 6 | overweight |
| 5.8 | 6 | underweight |
| 8.7 | 6 | overweight |

## Exercise 10

Consider the data table on the left. Reduce the data by dropping features and applying aggregation to

a) Show the minimum weight of overweight babies, for each age class.

b) For each age class, show the average weight.

c) For age class, show the summarized weight of all underweight babies and of all overweight babies.