

# Intelligent Systems

## Excercise 4- Representation / Feature Selection

---

Simon Reichhuber

December 09, 2020

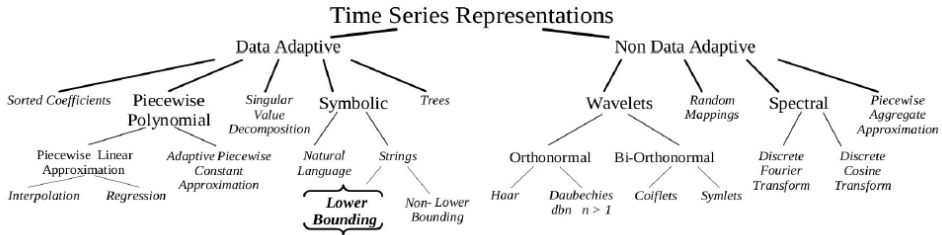
University of Kiel, Winter Term 2019

1. Data representation
2. Principal component analysis
3. Python PCA
4. Feature Selection Basics
5. Inconsistency Rate
6. Feature Selection Techniques
7. SAX Algorithm with Python

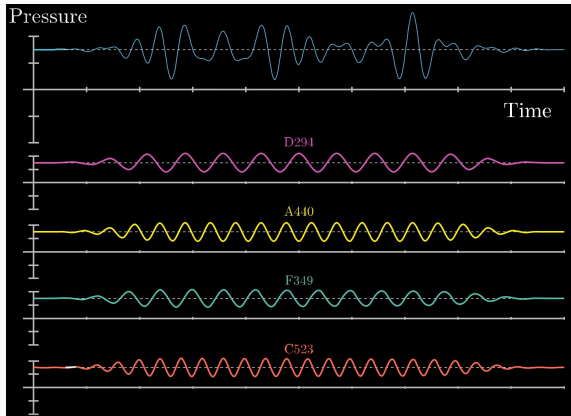
# Data representation

---

# NON DATA ADAPTIVE VS. DATA ADAPTIVE APPROXIMATION

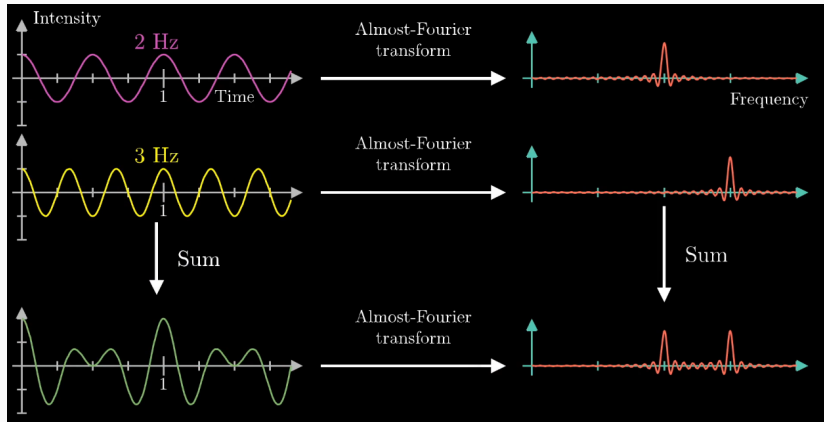


The FT represents the time series in the frequency domain. The signal is constructed as a sequence of sine and cosine terms.



**Figure 1:** Fourier Transform

The FT represents the time series in the frequency domain.  
The signal is constructed as a sequence of sine and cosine terms.



**Figure 2:** Fourier Transform

A. Explain the idea of the *Shape Definition Language* and its application?

B. Approximate the time series with the following approximations:

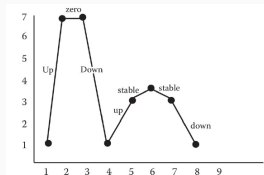
- *Piecewise Aggregate Approximation (PAA)* with 4 segments.
- *Clipping* to binary values (→ search the procedure on the internet).
- *Picewise Linear Approximation* with 4 segments.
- *Run-Length Encoding (RLE)*.

C. Aggregate the timeseries to the following statistical measures:

- *Mean*
- *Standard deviation*
- *Mode*

D. What are the advantages and disadvantages of the *clipping* procedure?

Given the case that the most important information of a time series can be extracted from the rough shape, limited terms of the *Shape Definition Language (SDL)* are enough to model it, i.e. *Up, up, stable, zero, down, Down*. These represent the different slopes of the time series. Advantageously, the representation can easily be processed by algorithms for symbolic sequences, like *Longest Common Subsequence (LCSS)*.



**Figure 3:** Shape Definition Language<sup>1</sup>



A. Explain the idea of the *Shape Definition Language* and its application?

**B. Approximate the time series with the following approximations:**

- *Piecewise Aggregate Approximation (PAA)* with 4 segments.
- *Clipping* to binary values ( $\rightarrow$  search the procedure on the internet).
- *Picewise Linear Approximation* with 4 segments.
- *Run-Length Encoding (RLE)*.

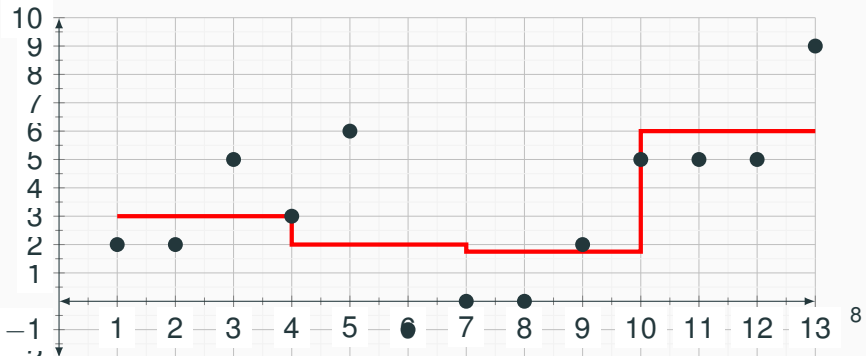
C. Aggregate the timeseries to the following statistical measures:

- *Mean*
- *Standard deviation*
- *Mode*

D. What are the advantages and disadvantages of the *clipping* procedure?

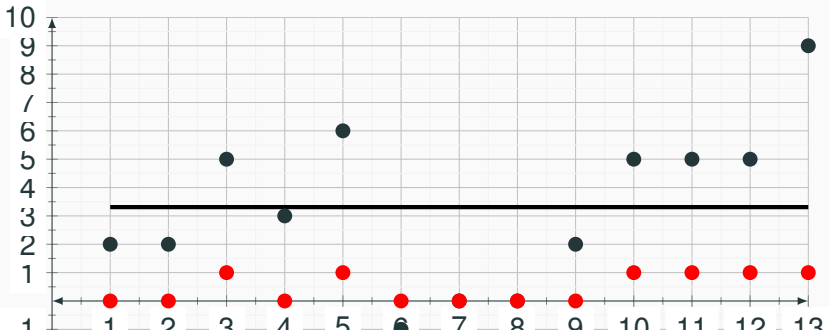
# 1. B PIECEWISE AGGREGATE APPROXIMATION

- PAA Segment Length =  $\frac{t_{\text{end}} - t_{\text{start}}}{\# \text{Segments}}$
- 4 segments  $\rightarrow$  PAA Segment Length =  $\frac{13-1}{4} = 3$
- Segment 1:  $\frac{2+2+5+3}{4} = 3$
- Segment 2:  $\frac{3+6+(-1)+0}{4} = 2$
- ...

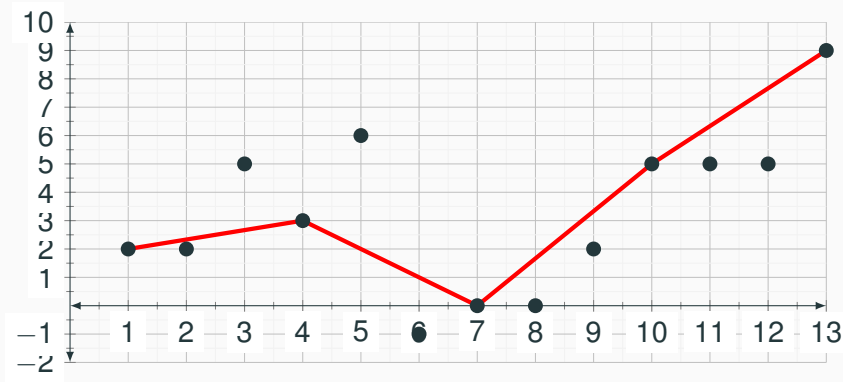


# 1. B CLIPPING

- Calculate time series' mean:  $\mu = \frac{1}{N} \sum_{i=1}^N x_i$
- The clipped values are given by:  $y_i^* = \begin{cases} 1, & \text{for } x_i \geq \mu \\ 0, & \text{otherwise} \end{cases}$
- $\mu = \frac{2+2+5+3+6+(-1)+0+0+2+5+5+5+9}{13} \approx 3.31$
- Clipping: 0, 0, 1, 0, 1, 0, 0, 0, 0, 1, 1, 1, 1



# 1. B PIECEWISE LINEAR APPROXIMATION



- Counter  $\{n\}$  represents the number of repetitions of the following symbol
- $2, 2, 5, 3, 6, -1, 0, 0, 2, 5, 5, 5, 9 \rightarrow$   
 $\{2\}2, 5, 3, 6, -1, \{2\}0, 2, \{3\}5, 9$

- A. Explain the idea of the *Shape Definition Language* and its application?
- B. Approximate the time series with the following approximations:
- *Piecewise Aggregate Approximation (PAA)* with 4 segments.
  - *Clipping* to binary values ( $\rightarrow$  search the procedure on the internet).
  - *Picewise Linear Approximation* with 4 segments.
  - *Run-Length Encoding (RLE)*.
- C. Aggregate the timeseries to the following statistical measures:**
- *Mean*
  - *Standard deviation*
  - *Mode*
- D. What are the advantages and disadvantages of the *clipping* procedure?

Given  $N$   $D$ -dimensional points  $x_i \in \mathbb{R}^d, i = 1, \dots, N$ , the following statistical measures can be defined:

- *Mean*:  $\mu = \frac{1}{N} \sum_{i=1}^N x_i \rightarrow 3.31$
- *Standard deviation*:  $\sigma = \sqrt{\frac{1}{N-1} \sum_{i=1}^N (x_i - \mu)^2} \rightarrow 2.84$
- *Mode*:  $m = \operatorname{argmax}_x |\{y | y \in \{x_1, \dots, x_N\}, x = y\}| \rightarrow 5$

- A. Explain the idea of the *Shape Definition Language* and its application?
- B. Approximate the time series with the following approximations:
- *Piecewise Aggregate Approximation (PAA)* with 4 segments.
  - *Clipping* to binary values ( $\rightarrow$  search the procedure on the internet).
  - *Picewise Linear Approximation* with 4 segments.
  - *Run-Length Encoding (RLE)*.
- C. Aggregate the timeseries to the following statistical measures:
- *Mean*
  - *Standard deviation*
  - *Mode*
- D. What are the advantages and disadvantages of the *clipping procedure*?



What are the advantages and disadvantages of the *clipping* procedure?

## **Advantages:**

- Extreme compression rate (*float*  $\rightarrow$  *bool*)
- Simple representation (only binary values)
- Rough patterns can be found easily

## **Disadvantage:**

- Very unprecise representation of the signal

- A. Explain the idea of the *Shape Definition Language* and its application?
- B. Approximate the time series with the following approximations:
- *Piecewise Aggregate Approximation (PAA)* with 4 segments.
  - *Clipping* to binary values ( $\rightarrow$  search the procedure on the internet).
  - *Picewise Linear Approximation* with 4 segments.
  - *Run-Length Encoding (RLE)*.
- C. Aggregate the timeseries to the following statistical measures:
- *Mean*
  - *Standard deviation*
  - *Mode*
- D. What are the advantages and disadvantages of the *clipping* procedure?

What is the main difference between the *Adaptive Picewise Aggregate Approximation (APAA)* and the *PAA*?

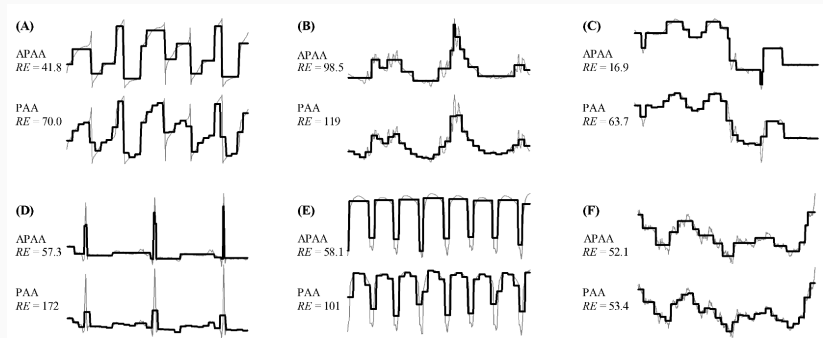
**Main difference:**

- Variable length of the (temporal) sections
- Adaptive according to local details of a time series
  - Sections of frequent movement will be parted in smaller intervals
  - A Section without significant information will be represented as a large interval

What is the main difference between the *Adaptive Picewise Aggregate Approximation (APAA)* and the *PAA*?

**Advantage:**

- The error (e.g. Least-Squares) between the raw data and the representation is reduced.<sup>2</sup>



# Principal component analysis

---

**A. What is the goal of the *Principal Component Analysis (PCA)* and what is its basic assumption.**

**B. What is the benefit of the *PCA*?**

**C. Describe the following items:**

- Zero-mean feature
- Variance
- Standard deviation
- Covariance matrix
- Arithmetic mean
- Eigenvector
- Eigenvalue
- Projection onto new feature space

**D. How can we get a dimensionality reduction with the means of Eigenvalues?**

What is the goal of the *Principal Component Analysis* (PCA) and what is its basic assumption. **Basic assumption:**

- The larger the variance, the higher is the level of information.

### Goal:

- Dimensionality reduction:
  - Remove dimensions with poor information gain
  - Generate new dimensions that better fit to the structure of the data, i.e. at which the data has the largest variance.

### Benefits:

- **Reduce computing time:** Through the usage of data mining algorithms applied on dimensionality-reduced data sets.
- **Feature selection:** Very easy; just select the PCA meta features.

- A. What is the goal of the *Principal Component Analysis* (*PCA*) and what is its basic assumption.
- B. What is the benefit of the *PCA*?
- C. **Describe the following items:**
- Zero-mean feature
  - Variance
  - Standard deviation
  - Covariance matrix
  - Arithmetic mean
  - Eigenvector
  - Eigenvalue
  - Projection onto new feature space
- D. **How can we get a dimensionality reduction with the means of Eigenvalues?**



**Given:** Data set with  $N$  samples and  $D$  dimensions (features).

### 1. Standardisation:

- Calculate the arithmetic mean for each feature  $j$ :

$$\mu_j = \frac{1}{N} \sum_{i=1}^N x_{j,i}$$

- Zero-mean sample:

$$x'_{j,i} = x_{j,i} - \mu_j$$

- Calculate the standard deviation for each feature  $j$ :

$$\sigma_j = \sigma_j = \sqrt{\frac{1}{N-1} \sum_{i=1}^N x'^2_{j,i}}$$

- Standardisation for every feature  $j$ :

$$x_{j,i} = \frac{x'_{j,i}}{\sigma_j}$$

### 2. Calculate covariance matrix:

- $s_{j,j'} = \frac{1}{N-1} \sum_{i=1}^N x'_{j,i} \cdot x'_{j',i}$
- $S = \begin{pmatrix} s_{1,1} & \cdots & s_{1,D} \\ \vdots & \ddots & \vdots \\ s_{D,1} & \cdots & s_{D,D} \end{pmatrix}$

### 3. Calculate eigenvectors and eigenvalues:

- Eigenvalues represent the ratio of the variance along their corresponding eigenvectors
- The eigenvector with the largest eigenvalue represents the direction, in which the data has the largest variance
- The eigenvector with the second largest eigenvalue represents a orthogonal direction w.r.t. the first eigenvector, in which the data has the second largest variance, etc.
- Eigenvectors are the *principal components (PC)*, which are more suitable to model the structure of the data than the original features (assuming that high information gain corresponds to high variance).

**4. Dimensionality reduction:** Choose the most important eigenvalues according to their eigenvalues:

- Method 1: The sum of the remaining eigenvalues shall be larger than a predefined ratio (e.g. 0.75) of the sum of all eigenvalues.
- Method 2: Dimensions shall be removed if the eigenvalue of the corresponding eigenvector is lower than the mean of all eigenvalues.

**4. Projection onto the new feature space:**

- Spanned by the new selected eigenvectors (*PCs*).

# Python PCA

---

- A. Download the file *04\_Representation.ipyn* from *OpenOlat*.
- B. In order to solve the tasks, you can use the library *numpy*.
- C. Compare your results afterwards with the help of *sklearn*.

# Feature Selection Basics

---

- A. What are the tasks and goals of feature selection?**
- B. What are benefits of feature selection?
- C. Describe the term “weakly relevant but non-redundant features” ?
- D. Creating feature subsets, what is the advantage of *Random Generation (RG)* over *Sequential Forward Generation (SFG)*, *Sequential Backward Generation (SBG)*, and *Bidirectional Generation (BG)*?
- E. Enumerate and describe the three different search strategies for finding an adequate subset of features. Additionally, mention their advantages and disadvantages.



What are the tasks and goals of feature selection?

## Task:

- Find and select the optimal subset of the feature set.
- The subset should meet the following conditions:
  - Provide the highest accuracy for the chosen task (classification, regression).
  - Minimal number of features simultaneously guaranteeing at least a required minimum accuracy. (no redundant features)
  - Minimum costs (in computing time and monetary).
  - Improved comprehensibility.

## Goal:

- Distinguish important (good) features from unimportant (bad) features.

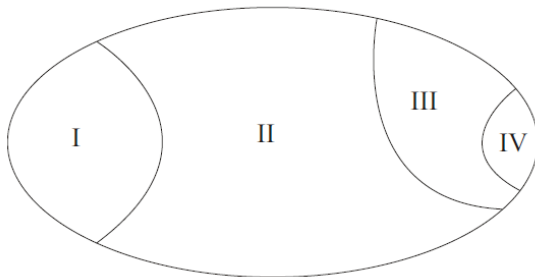
- A. What are the tasks and goals of feature selection?
- B. What are benefits of feature selection?**
- C. Describe the term “weakly relevant but non-redundant features” ?
- D. Creating feature subsets, what is the advantage of *Random Generation (RG)* over *Sequential Forward Generation (SFG)*, *Sequential Backward Generation (SBG)*, and *Bidirectional Generation (BG)*?
- E. Enumerate and describe the three different search strategies for finding an adequate subset of features. Additionally, mention their advantages and disadvantages.

## Benefits:

- Saving of time through less amount of data (fewer features, possibly fewer patterns)
- Better data quality by elimination of unimportant (e.g. unnecessary or high-noisy) information.
- Better comprehensability of the results of the pattern recognition process in smaller dimensional spaces; facilitated evaluation; interpretability can be impossible by given a high number of features
- Reduced costs, for example by only buying the relevant features (e.g. direct marketing) or by reduced computing resources (e.g. in technical applications like signal processing)

- A. What are the tasks and goals of feature selection?
- B. What are benefits of feature selection?
- C. Describe the term “weakly relevant but non-redundant features” ?**
- D. Creating feature subsets, what is the advantage of *Random Generation (RG)* over *Sequential Forward Generation (SFG)*, *Sequential Backward Generation (SBG)*, and *Bidirectional Generation (BG)*?
- E. Enumerate and describe the three different search strategies for finding an adequate subset of features. Additionally, mention their advantages and disadvantages.

# 1. C WAEKLY RELEVANT, BUT NON-REDUNDANT FEATURES



I: irrelevante Merkmale

II: schwach relevante Merkmale

III: schwach relevante, aber nicht-redundante Merkmale

IV: stark relevante Merkmale

III+IV: optimale Teilmenge

- A. What are the tasks and goals of feature selection?
- B. What are benefits of feature selection?
- C. Describe the term “weakly relevant but non-redundant features” ?
- D. Creating feature subsets, what is the advantage of *Random Generation (RG)* over *Sequential Forward Generation (SFG)*, *Sequential Backward Generation (SBG)*, and *Bidirectional Generation (BG)*?**
- E. Enumerate and describe the three different search strategies for finding an adequate subset of features. Additionally, mention their advantages and disadvantages.

Creating feature subsets, what is the advantage of *Random Generation (RG)* over *Sequential Forward Generation (SFG)*, *Sequential Backward Generation (SBG)*, and *Bidirectional Generation (BG)*?

**Problem with SFG, SBG, and BG:**

Incrementally removing the most unimportant or adding the most important feature may not end in an optimal solution!

**Goal of RG:**

Prevent local optima



- A. What are the tasks and goals of feature selection?
- B. What are benefits of feature selection?
- C. Describe the term “weakly relevant but non-redundant features” ?
- D. Creating feature subsets, what is the advantage of *Random Generation (RG)* over *Sequential Forward Generation (SFG)*, *Sequential Backward Generation (SBG)*, and *Bidirectional Generation (BG)*?
- E. Enumerate and describe the three different search strategies for finding an adequate subset of features. Additionally, mention their advantages and disadvantages.**

- Exhaustive Search:
  - + Searches through all combinations of features, finds optimum
  - Runtime
- Heuristic Search:
  - Heuristics are applied (e.g. based on expert knowledge)
  - + Certain paths of the search space → Reduced runtime
  - Not guaranteed to find the optimum
- Non-deterministic Search:
  - Random selection of the next feature subset
  - + Fixed number of iterations → Reduced runtime
  - Global optimum not very likely

# Inconsistency Rate

---

- A. Look at the Table. This data set represents relevant data for the decision whether to play tennis or not. The column “Play” represents the class of the sample. First apply binning of the temperature values, in order to reduce the continuous temperature value range to three ordinal values. Also, take care of an equal interval size.**
- B. Calculate the inconsistency rate (IR) of the new data set. Why does the IR play a role for the feature selection?**

## 2. A BINNING I

Outlook	Temperature	Humidity	Windy	Play
overcast	24	high	false	no
rainy	12	normal	false	no
sunny	18	low	true	yes
overcast	13	low	true	no
sunny	23	high	true	yes
rainy	24	normal	false	yes
rainy	19	high	true	no
overcast	17	normal	false	yes
sunny	14	high	false	yes
overcast	21	high	false	no
sunny	17	low	true	yes
rainy	18	high	true	no
rainy	22	normal	false	yes
sunny	12	high	false	yes
overcast	10	low	true	no
sunny	11	high	false	no
overcast	12	low	true	yes
overcast	20	high	false	yes
sunny	16	low	true	no
rainy	15	high	true	yes
rainy	21	normal	false	no

### Values:

10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24

### Bins:

- 10, 11, 12, 13, 14 **Cold**
- 15, 16, 17, 18, 19 **Mild**
- 20, 21, 22, 23 ,24 **Warm**

## 2. A BINNING III

Outlook	Temperature	Humidity	Windy	Play
overcast	warm	high	false	no
rainy	cold	normal	false	no
sunny	mild	low	true	yes
overcast	cold	low	true	no
sunny	warm	high	true	yes
rainy	warm	normal	false	yes
rainy	mild	high	true	no
overcast	mild	normal	false	yes
sunny	cold	high	false	yes
overcast	warm	high	false	no
sunny	mild	low	true	yes
rainy	mild	high	true	no
rainy	warm	normal	false	yes
sunny	cold	high	false	yes
overcast	cold	low	true	no
sunny	cold	high	false	no
overcast	cold	low	true	yes
overcast	warm	high	false	yes
sunny	mild	low	true	no
rainy	mild	high	true	yes
rainy	warm	normal	false	no

- A. Look at the Table. This data set represents relevant data for the decision whether to play tennis or not. The column “Play” represents the class of the sample. First apply binning of the temperature values, in order to reduce the continuous temperature value range to three ordinal values. Also, take care of an equal interval size.
- B. Calculate the inconsistency rate (IR) of the new data set. Why does the IR play a role for the feature selection?**



## 2. B INCONSISTENCY RATE

Muster	Klasse: Yes	Klasse: No	IZ
overcast 1	1	2	$IZ = 3-2 = 1$
overcast 2	1	2	$IZ = 3-2 = 1$
overcast 3	1	0	$IZ = 1-1 = 0$
rainy 1	0	1	$IZ = 1-1 = 0$
rainy 2	2	1	$IZ = 3-2 = 1$
rainy 3	1	2	$IZ = 3-2 = 1$
sunny 1	2	1	$IZ = 3-2 = 1$
sunny 2	1	0	$IZ = 1-1 = 0$
sunny 3	2	1	$IZ = 3-2 = 1$

- $$IR = \frac{1+1+0+0+1+1+1+0+1}{21} = \frac{6}{21} = \frac{2}{7}$$

**Why should you check the dataset of inconsistencies after the feature selection?**

- Feature selection can produce inconsistencies by removing features.

Let  $D_i$  and  $D_j$  be two (sub)sets of features with inconsistency rates  $IR(D_i)$  and  $IR(D_j)$ , then:

- $D_i$  and  $D_j$  are indistinguishable, if  $IR(D_i) = IR(D_j)$  and  $|D_i| = |D_j|$
- $D_i$  is preferred to  $D_j$ , if:
  - $IR(D_i) = IR(D_j)$  and  $|D_i| < |D_j|$
  - $IR(D_i) < IR(D_j)$  and  $|D_i| \leq |D_j|$

	Haar- farbe	Größe	Gewicht	Sonnen- creme	Ergebnis
$i_1$	1	2	1	0	1
$i_2$	1	3	2	1	0
$i_3$	2	1	2	1	0
$i_4$	1	1	3	0	1
$i_5$	3	2	3	0	1
$i_6$	2	3	3	0	0
$i_7$	2	2	3	0	0
$i_8$	1	1	1	1	0

Which features would you select in this data set?

**Minimal Inconsistency Rate 0 for:**

- Hair color and sun creame
- Hair color, size, weight
- Any super set of the two sets above

**Result:** Hair color and sun creame

# Feature Selection Techniques

---

- A. What is the difference between the wrapper and the filter?**
- B. Calculate the Information Gain of every feature in Table 1. Sort your results and begin with the most important one.
- C. What is known by “Automated Branch and Bound Algorithmus” and what are its properties? Create an ABB search tree from the data of Table 1.

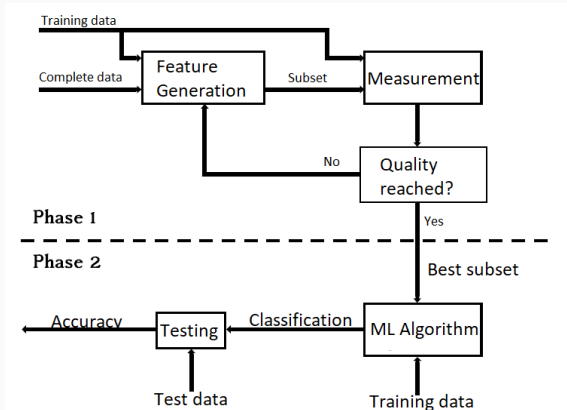
There are two basic methodologies for the feature selection:

- **Filter** analyse the quality of features in the feature space.
- **Wrapper** analyse the quality of features according to a suitable Machine Learning algorithm (e.g. Classification)

In principle, Wrapper achieve better results, as the actual Machine Learning task is taken into account. Though, the computing time is much higher, sometimes not computable in a reasonable amount of time. Hybrid approaches are possible.

### 3. A FILTER

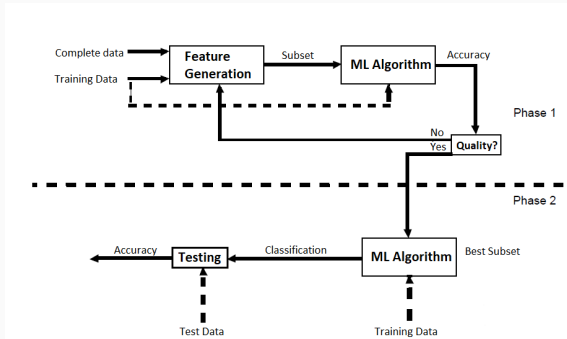
**Filter** analyse the quality of features in the feature space.



**Figure 6:** Filter Pipeline

### 3. A WRAPPER

**Wrapper** analyse the quality of features according to a suitable Machine Learning algorithm (e.g. Classification)



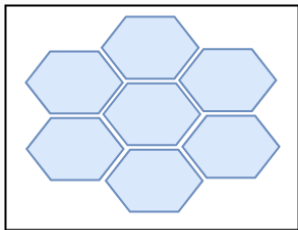
**Figure 7: Wrapper Pipeline**



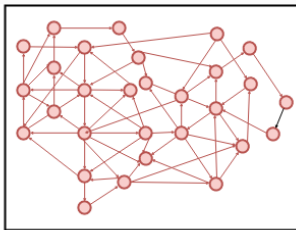
- A. What is the difference between the wrapper and the filter?
- B. Calculate the Information Gain of every feature in Table 1. Sort your results and begin with the most important one.**
- C. What is known by “Automated Branch and Bound Algorithmus” and what are its properties? Create an ABB search tree from the data of Table 1.

Calculate the Information Gain of every feature in Table 1. Sort your results and begin with the most important one.

- $IG(d) = I(X) - \sum_{l=1}^L \frac{|X_{d_l}|}{|X|} I(X_{d_l})$
- $I(X_{d_l}) = - \sum_{c=1}^C p_{X_{d_l}}(c) * \log_2 p_{X_{d_l}}(c)$
- $I(X) = - \sum_{c=1}^C p_X(c) * \log_2 p_X(c)$



Order



Entropy

**Figure 8:** Order vs. Entropy

- Entropy is a measure for information based on the probability distribution of events
- Hence, it is also a quantification for disorder or chaos
- The more chaotic a system is, the more information is necessary to describe it.
- General idea:
  - Given an unlikely event, then the information we gain, if it occurs is high
  - Given a likely event, then the information we gain, if it occurs is low

#### Shannon's entropy of events

Shannon's entropy (information content)  $I$  of an event with a probability of occurrence  $p$  is defined as:

$$I = \log_2 \frac{1}{p} = -\log_2 p$$

The unit of information is bit.<sup>3</sup>

Examples:

- $I$  of the event *Head* in a "Head or Tails" game:  $\log_2 \frac{2}{1} = 1$
- $I$  of the event 5 rolling a dice :  $\log_2 \frac{6}{1} \approx 2,59$
- $I$  of the event *odd number* rolling a dice:  $\log_2 \frac{6}{3} = 1$

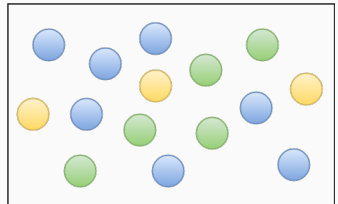
---

<sup>3</sup>If the logarithm of base 2 is used.

#### Information of events

$$I = \log_2 \frac{1}{p} = -\log_2 p$$

- blue ball:  $I = -\log_2 \frac{7}{15} = 1.09$
- green ball:  
 $I = -\log_2 \frac{5}{15} = 1.58$
- yellow ball:  
 $I = -\log_2 \frac{3}{15} = 2.32$



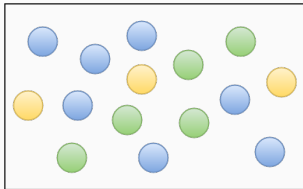
## Entropie

The entropy of a series of events is defined as:

$$H = - \sum_{i=1}^n p_i \log_2 p_i,$$

where  $n$  is the number of possible events.

- $H = -\left[\frac{7}{15} * \log_2 \frac{7}{15}\right] - \left[\frac{5}{15} * \log_2 \frac{5}{15}\right] - \left[\frac{3}{15} * \log_2 \frac{3}{15}\right] = 1.51$





- $IG(d) = I(X) - \sum_{l=1}^L \frac{|X_{d_l}|}{|X|} I(X_{d_l})$
- $I(X_{d_l}) = - \sum_{c=1}^C p_{X_{d_l}}(c) * \log_2 p_{X_{d_l}}(c)$
- $I(X) = - \sum_{c=1}^C p_X(c) * \log_2 p_X(c)$

### 3. B INFORMATION GAIN VI

Outlook	Temperature	Humidity	Windy	Play
Sunny	Hot	High	False	No
Sunny	Hot	High	True	No
Overcast	Hot	High	False	Yes
Rainy	Mild	High	False	Yes
Rainy	Cool	Normal	False	Yes
Rainy	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Sunny	Mild	High	False	No
Sunny	Cool	Normal	False	Yes
Rainy	Mild	Normal	False	Yes
Sunny	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Rainy	Mild	High	True	No

1. Calculate the information Berechne den Informationsgehalt der Klassenverteilung

- $I(X) = - \sum_{c=1}^C p_{X(c)} * \log_2 p_{X(c)}$
- $I(X) = -\frac{5}{14} \log_2(\frac{5}{14}) - \frac{9}{14} \log_2(\frac{9}{14}) = 0.94$

2. Calculate the information content of every feature value (= manifestation or possible instance):

- $I(X_{d_l}) = - \sum_{c=1}^C p_{X_{d_l}(c)} * \log_2 p_{X_{d_l}(c)}$
- $I(X_{sunny}) = -\frac{3}{5} \log_2(\frac{3}{5}) - \frac{2}{5} \log_2(\frac{2}{5}) = 0.97$
- $I(X_{rainy}) = -\frac{2}{5} \log_2(\frac{2}{5}) - \frac{3}{5} \log_2(\frac{3}{5}) = 0.97$
- $I(X_{overcast}) = 0 - \log_2(\frac{4}{4}) = 0$

3. Calculate the Information Gain of every feature:

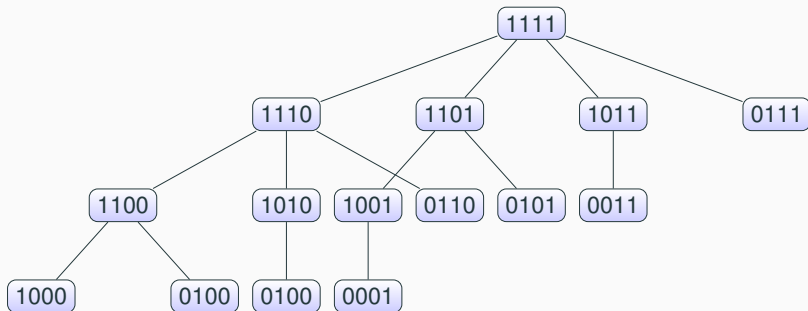
- $IG(Outlook) = I(X) - \sum_{l=1}^L \frac{|X_{d_l}|}{|X|} I(X_{d_l})$
- $IG(Outlook) = 0.94 - (\frac{5}{14} * 0.97 + \frac{5}{14} * 0.97 + \frac{4}{14} * 0) = 0.247$

4. Compare the features by sorting them by their Information Gain values:

- $IG(Outlook) = 0.247$ ,  $IG(Humidity) = 0.15$ ,  $IG(Windy) = 0.05$ ,  $IG(Temp) = 0.03$

- A. What is the difference between the wrapper and the filter?
- B. Calculate the Information Gain of every feature in Table 1.  
Sort your results and begin with the most important one.
- C. What is known by “Automated Branch and Bound Algorithmus” and what are its properties? Create an ABB search tree from the data of Table.**

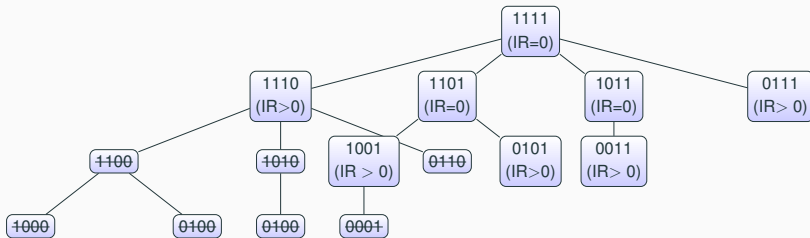
### 3. C ABB SEARCH TREE



**Figure 9:** ABB Search Tree

**Remark:** (1, 1, 1, 1) corresponds to  
(Outlook, Temperature, Humidity, Windy)

### 3. C ABB SEARCH TREE



**Figure 10: ABB Suchbaum Pruning**

⇒ One out of the feature selections

(Outlook, Temperature, Windy) and

(Outlook, Humidity, Windy) will be chosen.

# **SAX Algorithm with Python**

---



- A.** Download the jupyter notebook **4\_SAX.ipynb** from Open Olat. First, calculate the Euclidean distance of the two time series. Afterwards, apply the steps of the SAX algorithm and compare the distance of the two strings. What attracts your attention? Which parameters can be adapted to achieve better results?

Download the jupyter notebook **4\_SAX.ipynb** from Open Olat.  
First, calculate the Euclidean distance of the two time series.  
Afterwards, apply the steps of the SAX algorithm and compare  
the distance of the two strings. What attracts your attention?  
Which parameters can be adapted to achieve better results?

# Signature Task

---

- A.** Construct different features for the signatures and calculate the IG and IR values.