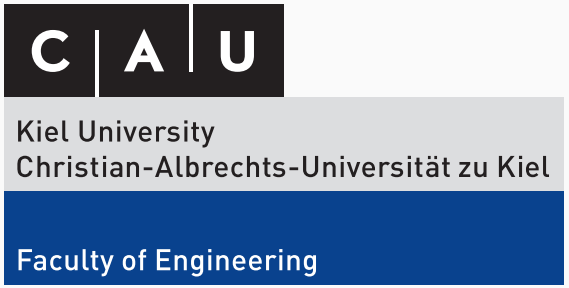


Intelligent Systems

Excersice 5- Feature Selection

Simon Reichhuber

November 25, 2019

University of Kiel, Winter Term 2019

Table of Content

Feature Selection Basics

C|A|U
Kiel University
Christian-Albrechts-Universität zu Kiel
Faculty of Engineering

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.

Kiel University
Christian-Albrechts-Universität zu Kiel
Faculty of Engineering

What are the tasks and goals of feature selection?

C|A|U
Kiel University
Christian-Albrechts-Universität zu Kiel
Faculty of Engineering

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, regresseion).
  - Minimal number of features simultanously guaranteeing at least a required minimum accuracy. (no redundant features)
  - Minimum costs (in computing time and monetary).
  - Improved comprehensability.

Goal:

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

C|A|U
Kiel University
Christian-Albrechts-Universität zu Kiel
Faculty of Engineering

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.

C|A|U
Kiel University
Christian-Albrechts-Universität zu Kiel
Faculty of Engineering

Benefits:

- Saving of time through less amount of data (fewer feautures, possibly fewer patterns)
- Better data quality by elimination of unimportant (e.g. unecessary or high-noisy) information.
- Better comprehensability of the results of the pattern recognition process in smaller dimensional spaces; facilitated evaluation; interpretability can be impossibly 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)

C|A|U
Kiel University
Christian-Albrechts-Universität zu Kiel
Faculty of Engineering

A. What are the tasks and goals of feature selection?

B. What are benefits of feature selection?

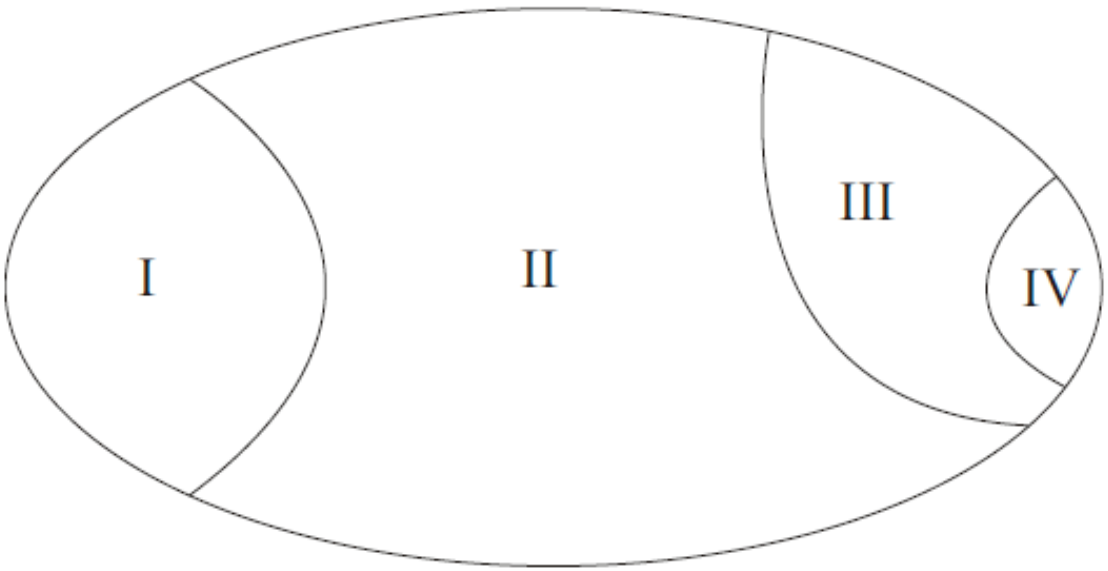
C. Describe the term “weakly relevant but non-redundant features” ?

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

Kiel University
Christian-Albrechts-Universität zu Kiel
Faculty of Engineering



I: irrelevant features
II: weakly relevant features
III: weakly relevant but non-redundant
IV: strongly relevant features
III+IV: optimal subset

C|A|U
Kiel University
Christian-Albrechts-Universität zu Kiel
Faculty of Engineering

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 $\rightarrow$ Reduced runtime
  - \- Not guaranteed to finde the optimum
- Non-deterministic Search:
  - Random selection of the next feature subset
  - \+ Fixed number of iterations $\rightarrow$ Reduced runtime
  - \- Global optimum not very likely

Inconsistency Rate

A. **Look at the Table. This data set represents relevant data for the decision wheter 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 plays 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

2. A Binning II

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 wheter 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 plays 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.

C|A|U
Kiel University
Christian-Albrechts-Universität zu Kiel
Faculty of Engineering

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 $|D_i| <= |D_j|$

Feature Selection with IR II

	Hair color	Size	Weight	Sun cream	Sun burned?
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

Kiel University
Christian-Albrechts-Universität zu Kiel
Faculty of Engineering

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.

Kiel University
Christian-Albrechts-Universität zu Kiel
Faculty of Engineering

Filter analyse the quality of features in the feature space.

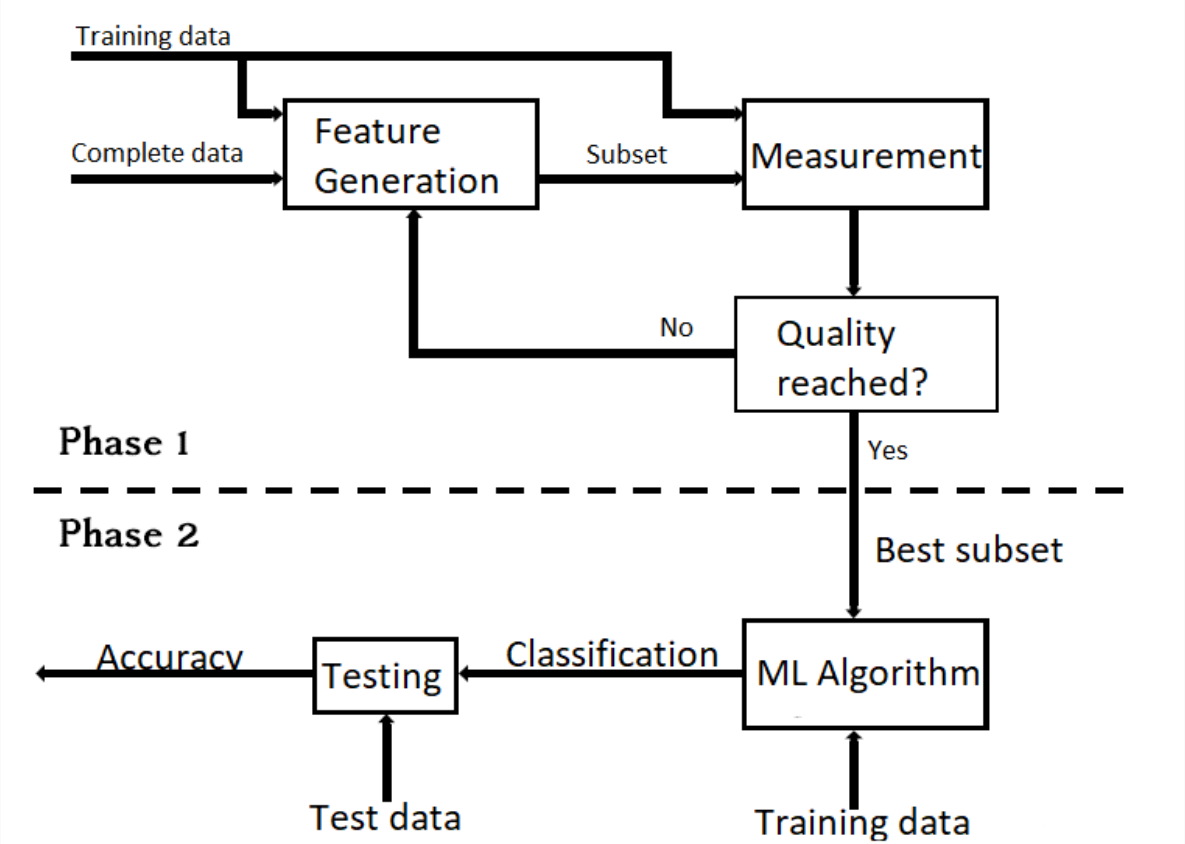


Figure 1: Filter Pipeline

Kiel University
Christian-Albrechts-Universität zu Kiel
Faculty of Engineering

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

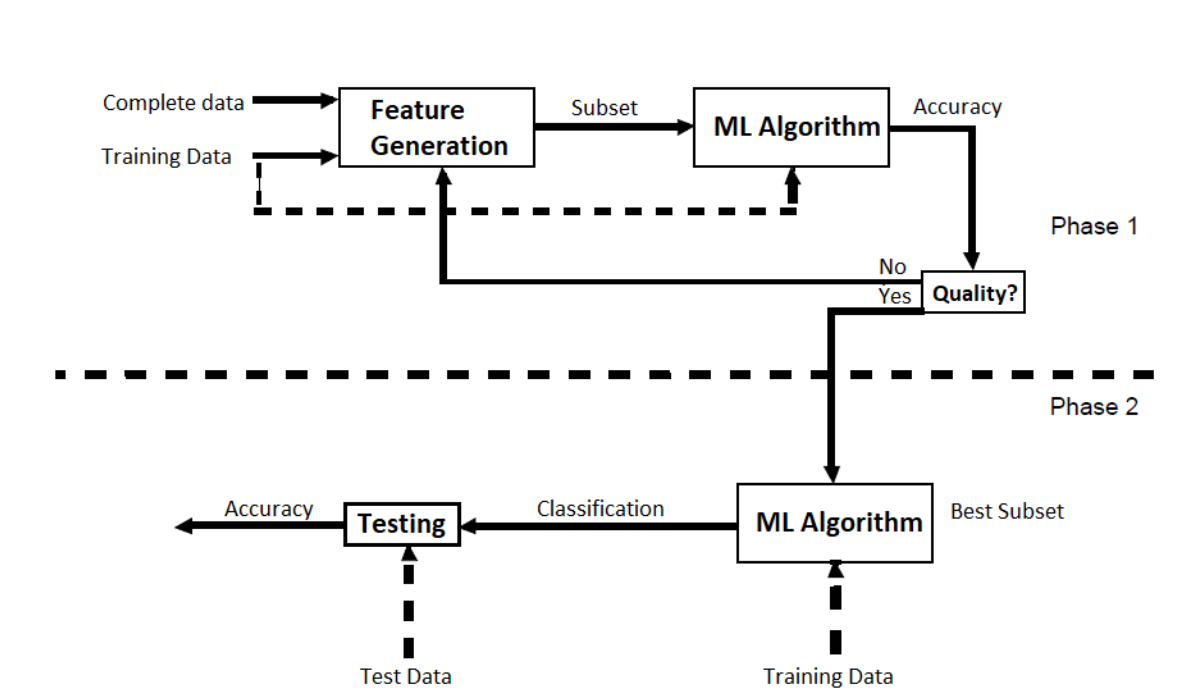


Figure 2: Wrapper Pipeline

Kiel University
Christian-Albrechts-Universität zu Kiel
Faculty of Engineering

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.

3. B Information gain

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

Kiel University
Christian-Albrechts-Universität zu Kiel
Faculty of Engineering

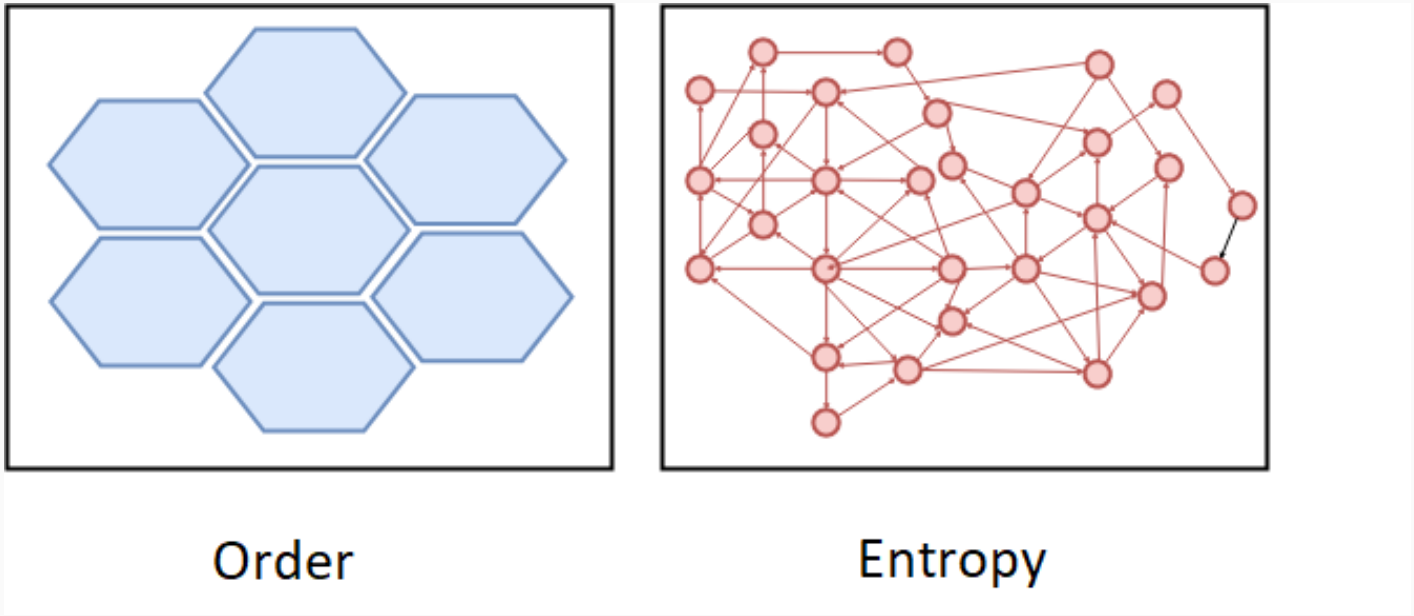


Figure 3: Order vs. Entropy

- Entropy is a measure for for information based on the probabilty 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

Kiel University
Christian-Albrechts-Universität zu Kiel
Faculty of Engineering

Shannon's entropy of events

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

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

The unit of information is bit.[1]

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: $\log_2 \frac{6}{3} = 1$

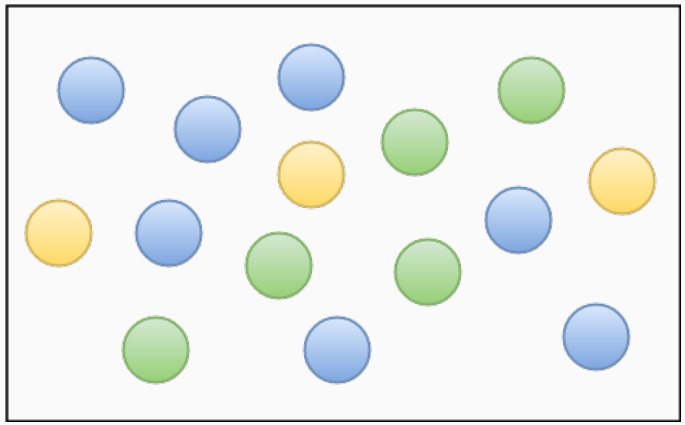
[1] If the logarithm of base 2 is used.

C|A|U
Kiel University
Christian-Albrechts-Universität zu Kiel
Faculty of Engineering

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:
  $I = -\log_2 \frac{3}{15} = 2.32$



C|A|U
Kiel University
Christian-Albrechts-Universität zu Kiel
Faculty of Engineering

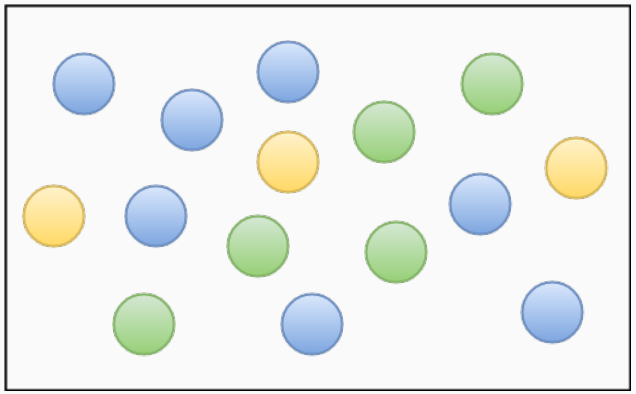
Entropie

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.

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



Kiel University
Christian-Albrechts-Universität zu Kiel
Faculty of Engineering

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

C|A|U
Kiel University
Christian-Albrechts-Universität zu Kiel
Faculty of Engineering

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

Kiel University
Christian-Albrechts-Universität zu Kiel
Faculty of Engineering

1. Calculate the information Berechne den Informationsgehalt der Klassenverteilung
   - $I(X) = -\sum_{c=1}^{C} px(c) * log_2 px(c)$
   - $I(X) = -\frac{5}{14} log_2(\frac{5}{14}) - \frac{9}{14} log_2(\frac{9}{14}) = 0.94$
2. Calculate the information content of every feature value (= manifestation or possible instance):
   - $I(X_{d_l}) = -\sum_{c=1}^{C} px_{d_l}(c) * log_2 px_{d_l}(c)$
   - $I(X_{sunny}) = -\frac{3}{5} log_2(\frac{3}{5}) - \frac{2}{5} log_2(\frac{2}{5}) = 0.97$
   - $I(X_{rainy}) = -\frac{2}{5} log_2(\frac{2}{5}) - \frac{3}{5} log_2(\frac{3}{5}) = 0.97$
   - $I(X_{overcast}) = 0 - log_2(\frac{4}{4}) = 0$
3. Calculate the Information Gain of every feature:
   - $IG(Outlook) = I(X) - \sum_{l=1}^{L} \frac{|X_{d_l}|}{|X|} I(X_{d_l})$
   - $IG(Outlook) = 0.94 - (\frac{5}{14} * 0.97 + \frac{5}{14} * 0.97 + \frac{4}{14} * 0) = 0.247$

Kiel University
Christian-Albrechts-Universität zu Kiel
Faculty of Engineering

4. Compare the featues 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.

Kiel University
Christian-Albrechts-Universität zu Kiel
Faculty of Engineering

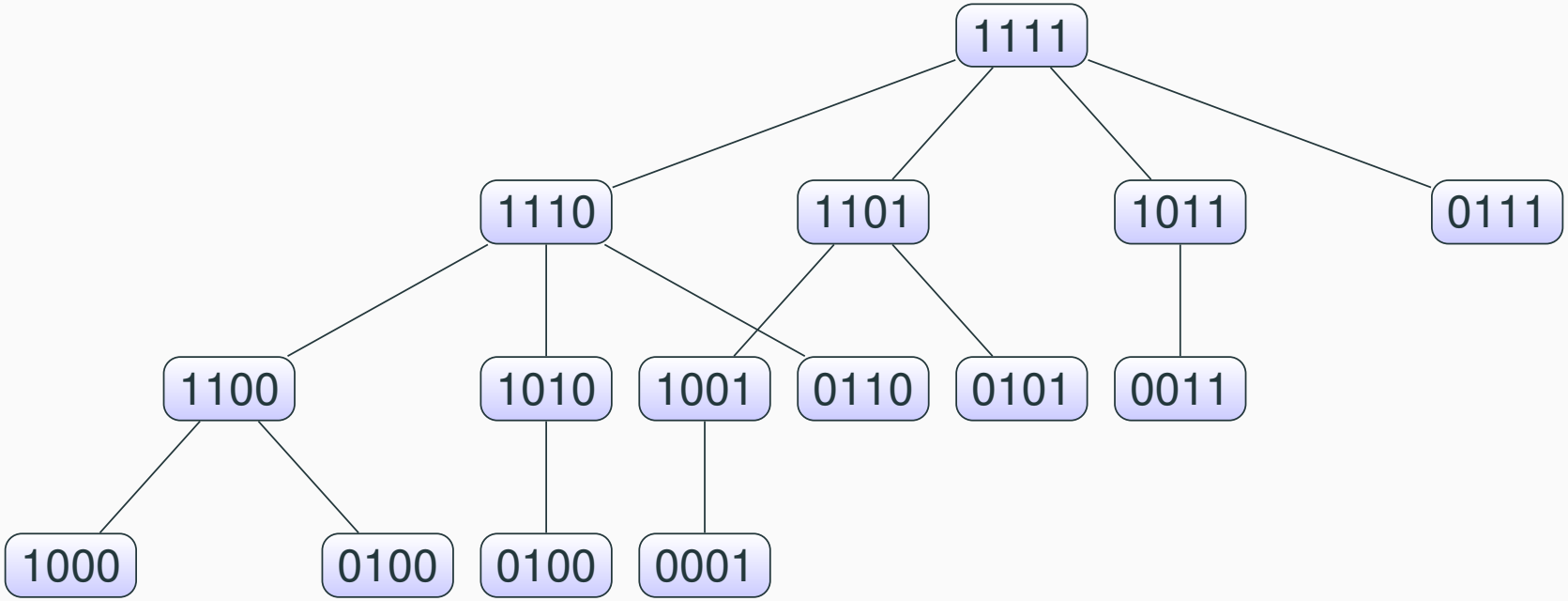


Figure 4: ABB Search Tree

Remark: $(1, 1, 1, 1)$ corresponds to
$(\texttt{Outlook}, \texttt{Temperature}, \texttt{Humidity}, \texttt{Windy})$

Kiel University
Christian-Albrechts-Universität zu Kiel
Faculty of Engineering

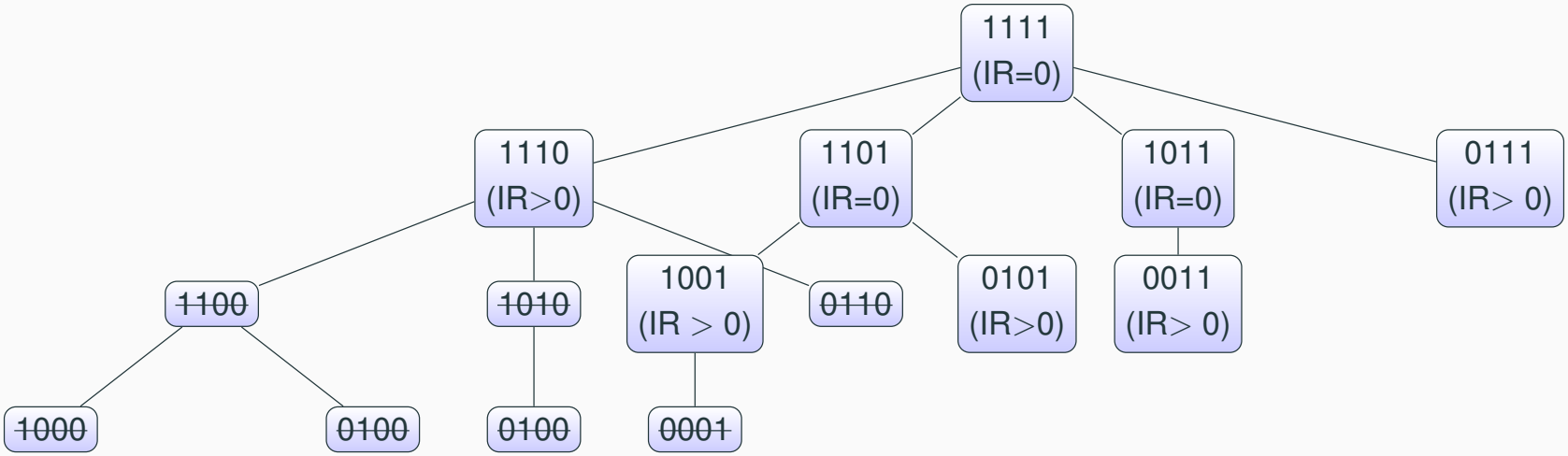


Figure 5: ABB Suchbaum Pruning

$\Rightarrow$ One out of the feature selections
(`Outlook`, `Temperature`, `Windy`) and
(`Outlook`, `Humidity`, `Windy`) will be chosen.

SAX Algorithm with Python

Kiel University
Christian-Albrechts-Universität zu Kiel
Faculty of Engineering

A. Download the jupyter notebook ***4_SAX.ipynb*** from Open Olat. First, calcluate 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 paramaters can be adapted two achieve better results?

Kiel University
Christian-Albrechts-Universität zu Kiel
Faculty of Engineering

Download the jupyter notebook ***4_SAX.ipynb*** from Open Olat.
First, calcluate 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 paramaters can be adapted two achieve better results?

Signature Task

Kiel University
Christian-Albrechts-Universität zu Kiel
Faculty of Engineering

A. Construct different feautures for the signatures and calculate the IG and IR values.