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

Exercise Sheet 4
Intelligent Systems

Preprocessing / Feature Selection

This exercise sheet will be discussed on December 09, 2020

Exercise 1 - Representation

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

B. Approximate the time series in Figure 1 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 time series to the following statistical measures:

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

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

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

Exercise 2 - Data Adaptive Representations

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:

- Zero-mean feature
- Variance
- Standard deviation
- Covariance matrix

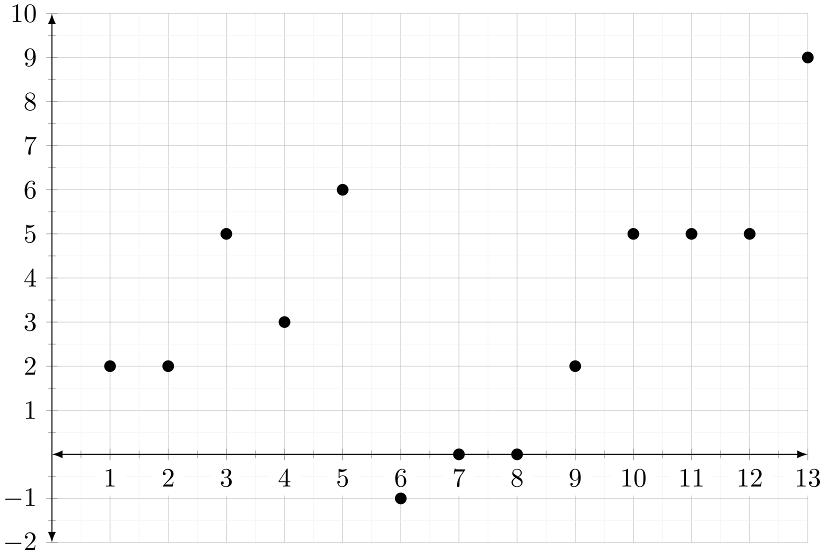

Abbildung 1: Point sequence.

- Arithmetic mean
- Eigenvector
- Eigenvalue
- Projection onto new feature space

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

Exercise 3 - Data Adaptive Representations

A. Download the file *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*.

Exercise 4 - Feature Selection I

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

B. What are benefits of feature selection?

C. Describe the term “wakly 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.

Exercise 5 - Feature Selection II

A. Look at the Table 1. 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?

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

Tabelle 1: Tennis data set.

Exercise 6 - Feature Selection III

A. What is the difference between the wrapper and the filter?

B. Calculate the Information Gain of every feature in Table 2. 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 2.

Exercise 7 - SAX Algorithm with Python

A.

Download the jupyter notebook ***SAX.ipynb*** from Open Olat. First, calcluate the Euclidean di-

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

Tabelle 2: Tennis data set.

stance 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?