Knowledge Discovery in Databases

SS 2016

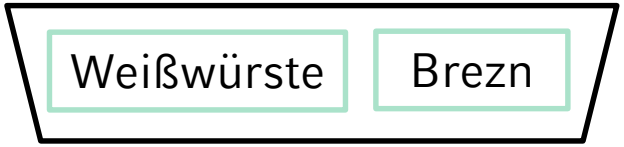
Exercise 1

Lecture: Prof. Dr. Thomas Seidl

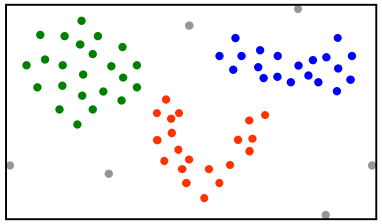
Tutorials: Julian Busch, Evgeniy Faerman,
Florian Richter, Klaus Schmid

Repetition

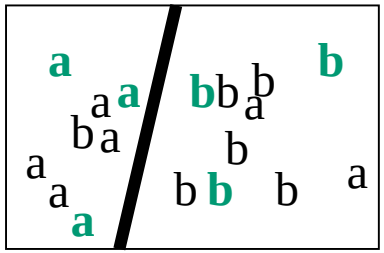
- Market-basket analysis
  - Find dependencies between items



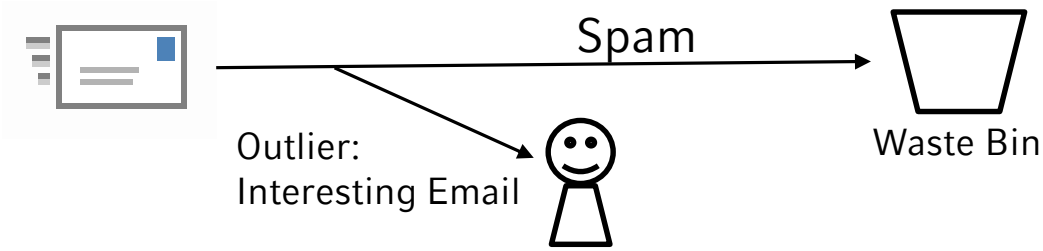
- Clustering
  - Aggregation of „similar“ objects



- Classification
  - Mapping of new objects to known groups

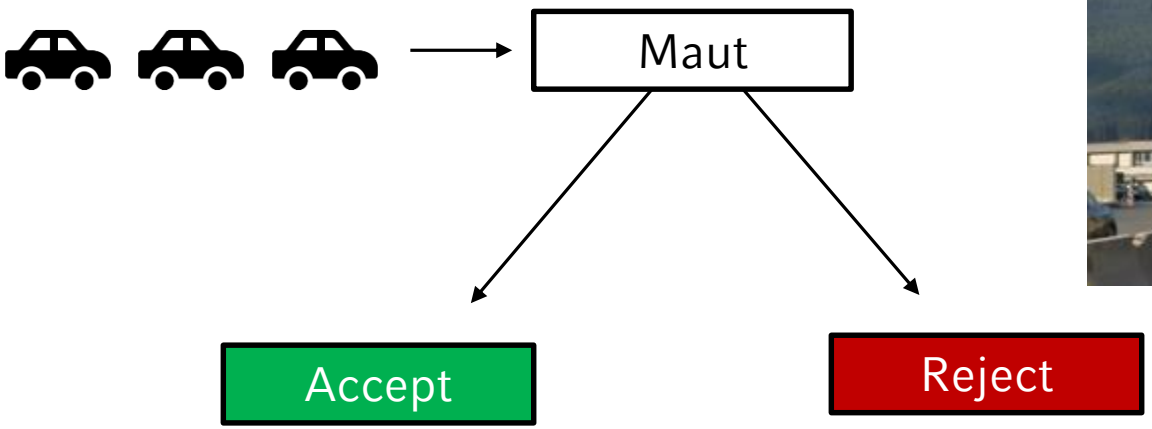


- Outlier Detection
  - Identify objects with significant dissimilarity



Exercise 1

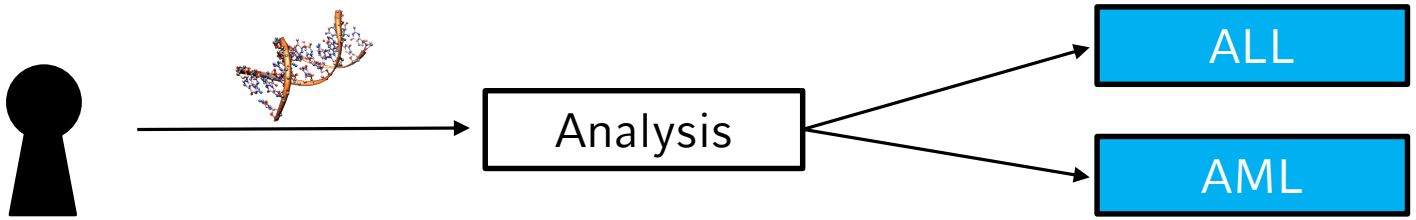
(a) **Optical character recognition/OCR**:
When crossing the alps using the Brenner Autobahn, there is the option to pay electronically in advance. When approaching the toll station, the barrier automatically opens when the number plate was recognized. The reconition happens fully automatically by a digital camera system.



Classification

Exercise 1

(b) **Computer Aided Diagnosis:**
Patients that suffer from blood cancer can be characterized in two categories (ALL and AML). The therapies for these two types partially differ, and the therapy for AML can sometimes be detrimental to patients suffering from ALL and the other way around. To avoid these complications, special gene expression data is used to differentiate between these two types by comparing them to the data from patients where the cancer type is already known.

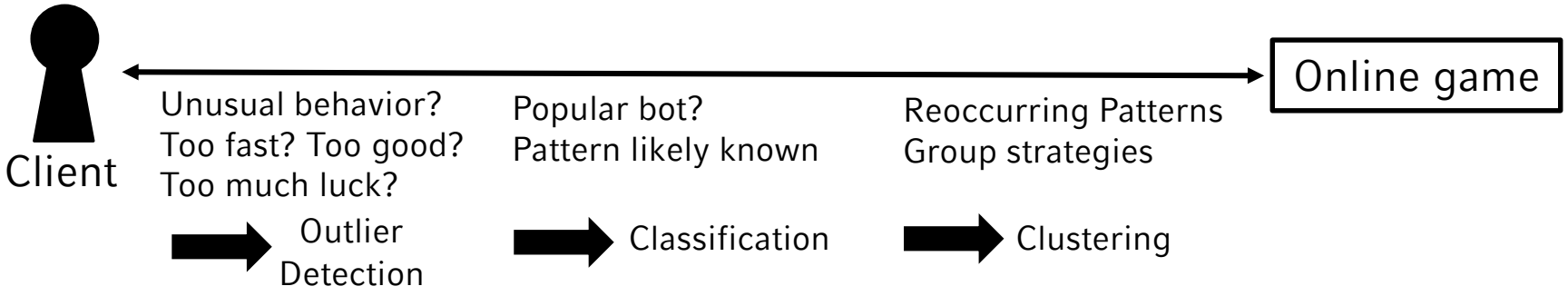


Classification

Exercise 1

(c) **Cheat Detection**

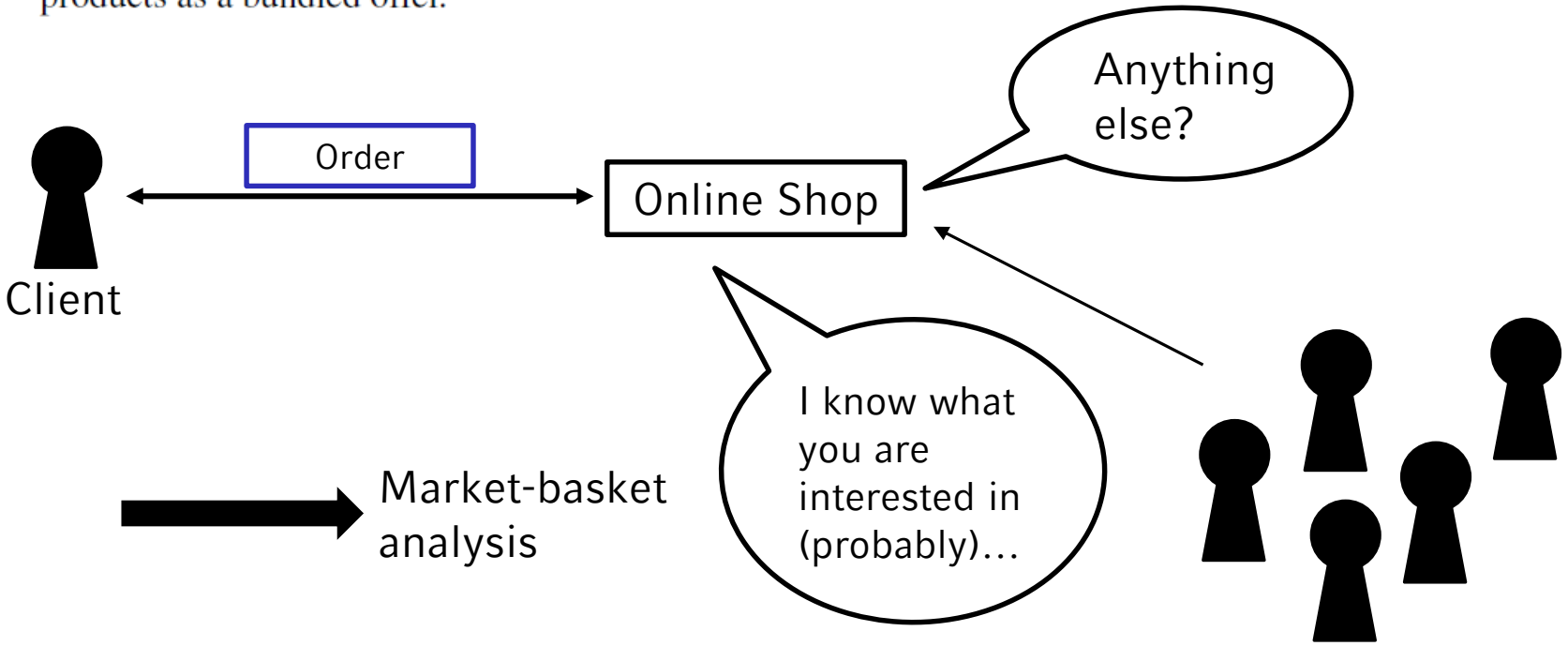
The operator of a multi player online game wants to protect his system against various violations of the terms of service. Particular problems are the use of game bot programs, the manipulation of timestamps in the communication protocol and attempts to predict random numbers used. To prevent this misuse, data mining is used on the available user data.



Exercise 1

(d) **Recommendation Systems**

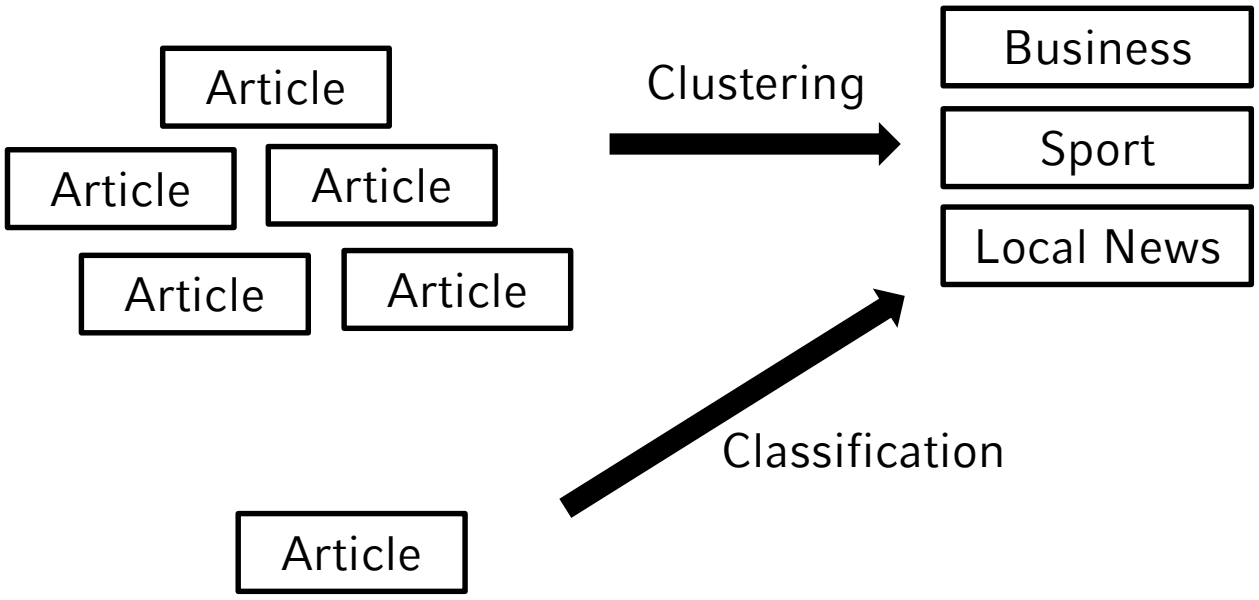
An online shopping portal wants to determine products that are automatically offered to registered customers upon login. The available data in particular includes products previously bought by the customer to predict his interests. For example a user that bought the book "Lord of the rings" might be offered the DVDs of the movie triology. A related task might be suggesting additional products for already chosen products as a bundled offer.



Exercise 1

(e) **News Aggregation**

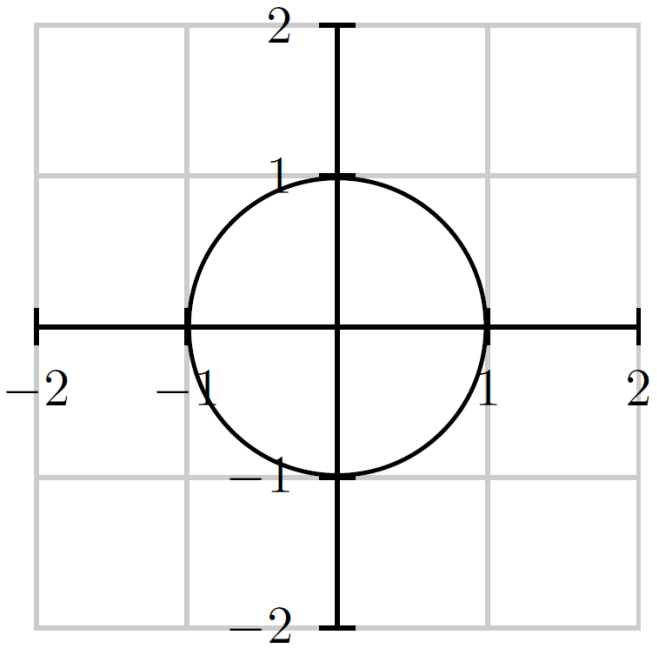
A news summary web site automatically collects current news from various sites to keep the visitor informed. However, news reports about the same subject are common and should be grouped by subject. This happens at multiple levels: there are obviously broad categories like politics and sports, and subcategories such as soccer. But even on a single soccer game, there will likely be different news sites reporting. Some articles will be identical to the report of a major agency, some will only be slightly modified, others will be original works.



Exercise 2

(a) Draw the „circle“ of distance 1 around a query object $q = (0, 0)$ for the following distance functions:

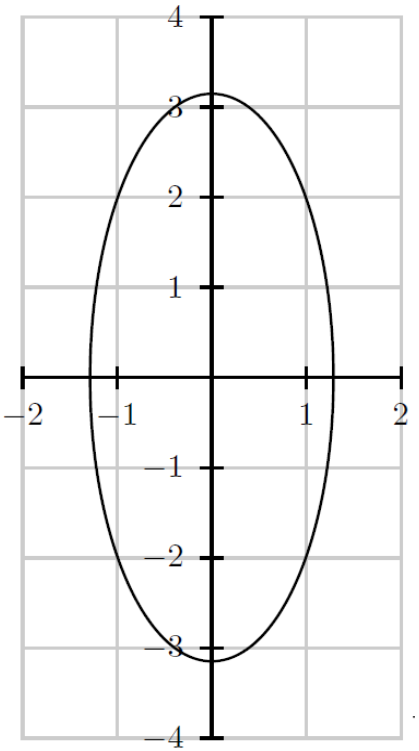
- Euclidean distance $dist(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$



Exercise 2

(a) Draw the „circle“ of distance 1 around a query object $q = (0, 0)$ for the following distance functions:

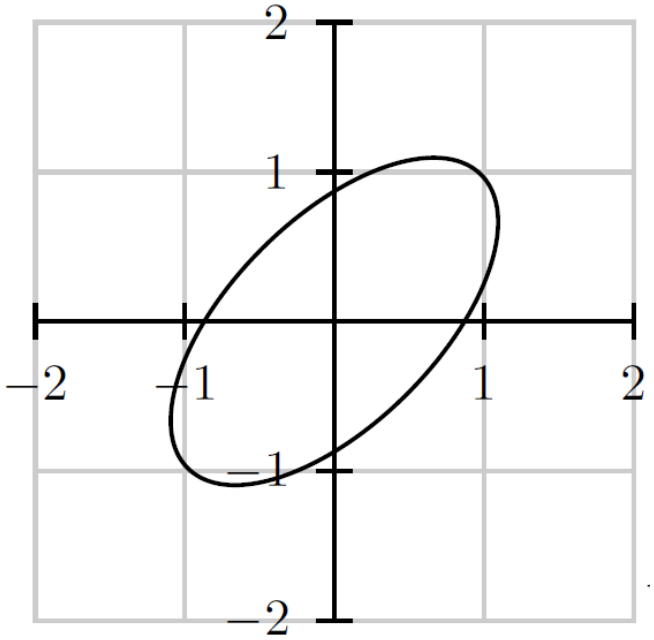
- Weighted Euclidean distance $dist(x, y) = \sqrt{\sum_{i=1}^{n} w_i(x_i - y_i)^2}$, where $w = (0.6, 0.1)$



Exercise 2

(a) Draw the „circle“ of distance 1 around a query object $q = (0, 0)$ for the following distance functions:

- Quadratic form distance $dist(x, y) = \sqrt{\sum_{i=1}^{n} \sum_{j=1}^{n} (x_i - y_i) w_{ij} (x_j - y_j)}$,
  where $w = \begin{pmatrix} 1.25 & -0.75 \\ -0.75 & 1.25 \end{pmatrix}$



Note: if you use as similarity matrix A the identity matrix => the Quadratic Form corresponds to the Euclidean distance

Zusatz: Distanzfunktionen

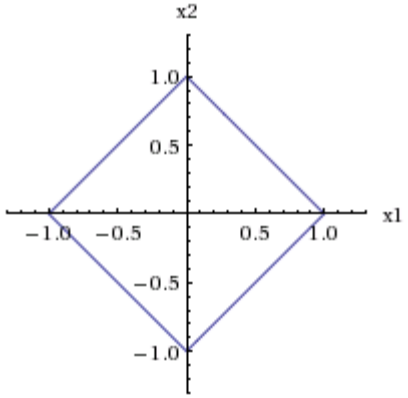
- Distanzfunktionen bilden ein Paar von Objekten (x,y) auf einen positiven reellen Wert ab: d: $X \times X \rightarrow \mathbb{R}_0^+$
- Eine Funktion, die negative Werte annehmen kann, ist als Distanzmaß ungeeignet: Was bedeutet d(x,y) < 0? Gleicher als gleiche Objekte?
- Erfüllt eine Distanzfunktion
  - Symmetry: $\forall p, q \in O: d(p,q) = d(q,p)$
  - Identity of indiscernibles $\forall p, q \in O: d(p,q) = 0 \Leftrightarrow p = q$ (Striktheit+Reflexivität)
  - Triangle inequality $\forall p, q, o \in O: d(p,q) \leq d(p,o) + d(o,q)$

  so ist d eine Metrik. Eine Menge X in Verbindung mit einer Metrik d nennt man metrischer Raum.
- Es gibt viele praktisch nützliche Distanzfunktionen, die keine Metriken sind.
- Vorsicht: Begriffe werden gerne und oft in der Literatur äquivalent genutzt.

Zusatz: Distanzfunktionen

- Isodistanz-Linien sind Mengen von Punkten, die den gleichen Distanzwert zu einem gemeinsamen Referenzpunkt besitzen: $isodist(x, range) = \{y \mid dist(x, y) = range\}$
- Einheitskreise sind Spezialfälle mit Distanz 1 um den Referenzpunkt 0 isodist(0,1)
- Für viele ungewichteten klassischen Distanzfunktionen führt die Isodistanzlinie für Distanz 1 (Einheitskreis) auf jeder Achse durch 1
- Manhattan-Distanz: $d_{MAN}(x, y) = \sum_{i=1}^{n} |x_i - y_i|$

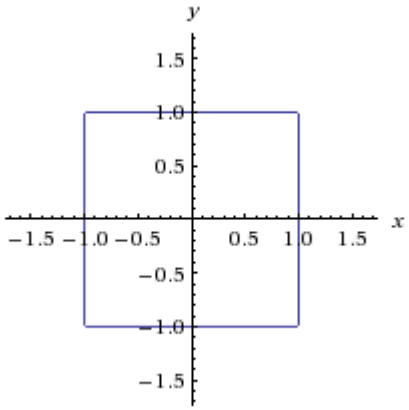
  $\text{isodist}_{MAN}(\binom{0}{0}, 1) = \{y \mid |y_1| + |y_2| = 1\}$



Zusatz: Distanzfunktionen

- Maximums-Distanz: $d_{MAX}(x,y) = \max_{i=1..n} |x_i - y_i|$

  $\text{isodist}_{MAX}(\binom{0}{0},1) = \{y \mid |y_1| = 1 \vee |y_2| = 1\}$



- Gewichtet man Distanzmaße, so verzerrt man die Isodistanzlinien in den entsprechenden Dimensionen
- Gewichte kleiner als 1 reduzieren die Wichtigkeit der Dimension. Eine Distanz ist weniger signifikant. Dadurch vergrößern sich die Abstände der Isodistanzen in dieser Dimension. Der Achsenabschnitt des Einheitskreises entfernt sich von 0.
- Gewichte größer 1 legen mehr Signifikanz auf diese Dimension. Selbst eine kleine Distanz hat mehr Einfluss auf den Gesamtabstand und die Isodistanzen rücken näher zusammen

Zusatz: Distanzfunktionen

- Dazu betrachte Aufgabe 2.a.ii)
- Für w = (0.6,0.1) erhält man den Einheitskreis
  - $isodist_{w-euclid}\left(\binom{0}{0}, 1\right) = \left\{\binom{y_1^2}{y_2^2} \middle| \sqrt{w_1 y_1^2 + w_2 y_2^2} = 1\right\} =$
    $\left\{\binom{y_1^2}{y_2^2} \middle| w_1 y_1^2 + w_2 y_2^2 = 1\right\}$
  - Achsenabschnitt $y_1 = 0: w_2 y_2^2 = 1 = 0.1 y_2^2 \Longrightarrow y_2 = \sqrt{\frac{1}{0.1}} = 3.16$
  - Achsenabschnitt $y_2 = 0: w_1 y_1^2 = 1 = 0.6 y_1^2 \Longrightarrow y_1 = \sqrt{\frac{1}{0.6}} = 1.29$

Zusatz: Distanzfunktionen

- Möglichkeit der Normierung
  - Um verschiedene Distanzen als Ensemble zu verwenden oder zu vergleichen, wird „Gleichwertigkeit“ benötigt
  - Meist: Mapping (Normierung) auf [0,1]. 1 entspricht maximal möglicher Unterschied (z.B. alle Vektoreneinträge zweier Vektoren ungleich)
  - Gewichte bei gewichteten Distanzen müssen angepasst werden

- Bei Hammingdistanz und gegebener Dimension sehr leicht:
  - $klassisch: dist(x,y) = \sum_{i=1}^{n} \begin{cases} 0 & , x_i = y_i \\ 1 & , x_i \neq y_i \end{cases}$
  - $normiert: dist'(x,y) = \frac{1}{n} \sum_{i=1}^{n} \begin{cases} 0 & , x_i = y_i \\ 1 & , x_i \neq y_i \end{cases}$

Exercise 2

(b) Decide for each of the following functions $d : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}_0^+$, whether they are a distance, and if so of which type.

(i) $d(x, y) = \sum_{i=1}^{n} (x_i - y_i)$

(ii) $d(x, y) = \sum_{i=1}^{n} (x_i - y_i)^2$

(iii) $d(x, y) = \sqrt{\sum_{i=1}^{n-1} (x_i - y_i)^2}$

(iv) $d(x, y) = \sum_{i=1}^{n} \begin{cases} 1 & \text{iff} \quad x_i = y_i \\ 0 & \text{iff} \quad x_i \neq y_i \end{cases}$

(v) $d(x, y) = \sum_{i=1}^{n} \begin{cases} 1 & \text{iff} \quad x_i \neq y_i \\ 0 & \text{iff} \quad x_i = y_i \end{cases}$

$d : O \times O \to \mathbb{R}_0^+$ $x, y, z \in O$:	reflexive reflexiv $x = y \Rightarrow d(x, y) = 0$	symmetric symmetrisch $d(x, y) = d(y, x)$	strict strikt $d(x, y) = 0 \Rightarrow x = y$	Triangle inequality Dreiecksungleichung $d(x, z) \leq d(x, y) + d(y, z)$
Dissimilarity function Unähnlichkeitsfunktion	×			
(Symmetric) Pre-metric (Symmetrische) Prämetrik	×	×		
Semi-metric, Ultra-metric Semimetrik, Ultrametrik	×	×	×	
Pseudo-metric Pseudometrik	×	×		×
Metric Metrik	×	×	×	×

Exercise 2

- i)

$$dist(x,y) = \sum_{i=1}^{n} (x_i - y_i)$$

Sei $x, y \in \mathbb{R}, x = 0, y = 1$

$$dist(x,y) = dist(0,1) = \sum_{i=1}^{n} (0 - 1) = -1 < 0$$

aber: Distanzfunktionen bilden auf $\mathbb{R}_0^+$ab! Damit ist diese Funktion keine Distanzfunktion.

- ii)

$$dist(x,y) = \sum_{i=1}^{n} (x_i - y_i)^2$$

Funktion ist offensichtlich reflexiv, strikt und symmetrisch. Aber Dreiecksungleichung wird verletzt:

Sei $x, y, z \in \mathbb{R}, x = 0, y = 1, z = 2$

$$dist(x,z) = \sum_{i=1}^{n} (x_i - z_i)^2 = (-2)^2 = 4 > 1 + 1 = dist(x,y) + dist(y,z)$$

Exercise 2

- iii)

$$dist(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

Reflexivität und Symmetrie sind gegeben. Striktheit aber nicht:

Sei $x, y \in \mathbb{R}^2, x = \binom{0}{0}, y = \binom{0}{1}$

$$dist\left(\binom{0}{0}, \binom{0}{1}\right) = 0$$

- iv)

$$dist(x,y) = \sum_{i=1}^{n} \begin{cases} 1 & , x_i = y_i \\ 0 & , x_i \neq y_i \end{cases}$$

Funktion ist symmetrisch, verletzt aber die Reflexivität:

Sei $x, y \in \mathbb{R}, x = y = 0$

$$dist(x,y) = dist(x,x) = \sum_{i=1}^{n} \begin{cases} 1 & , x_i = y_i \\ 0 & , x_i \neq y_i \end{cases} = 1 \neq 0$$

Exercise 2

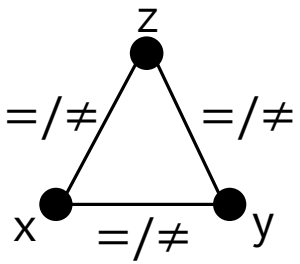
- v)

$$dist(x,y) = \sum_{i=1}^{n} \begin{cases} 0 & , x_i = y_i \\ 1 & , x_i \neq y_i \end{cases}$$

Reflexivität, Symmetrie und Striktheit lassen sich einfach zeigen.

Dreiecksungleichung ist ebenfalls gegeben, benötigt aber eine Fallunterscheidung:

Sei $x, y, z \in \mathbb{R}$.



1) $x = y = z: dist(x,z) = 0 \leq 0 + 0 = dist(x,y) + dist(y,z)$
2) $x = y \neq z: dist(x,z) = 1 \leq 0 + 1 = dist(x,y) + dist(y,z)$
3) $x \neq y = z: dist(x,z) = 1 \leq 1 + 0 = dist(x,y) + dist(y,z)$
4) $x \neq y \neq z, x = z: dist(x,z) = 0 \leq 1 + 1 = dist(x,y) + dist(y,z)$
5) $x \neq y \neq z, x \neq z: dist(x,z) = 1 \leq 1 + 1 = dist(x,y) + dist(y,z)$

Damit können wir für $x', y', z' \in \mathbb{R}^n$ zeigen:

$$dist(x,z) = \sum_{i=1}^{n} dist(x_i, z_i) \leq \sum_{i=1}^{n} \big(dist(x_i, y_i) + dist(y_i, z_i)\big)$$

$$= \sum_{i=1}^{n} dist(x_i, y_i) + \sum_{i=1}^{n} dist(y_i, z_i) = dist(x,y) + dist(y,z)$$