Ludwig-Maximilians-Universität München
Institut für Informatik
Prof. Dr. Thomas Seidl
Janina Sontheim, Maximilian Hünemörder

Knowledge Discovery in Databases

WS 2019/20

Exercise 1: Data Mining Tasks, Data Types, Distance Functions

Exercise 1-1 Data mining tasks

Which data mining tasks (association rule mining, clustering, outlier detection, classification, etc.) are at the heart of the following use cases? Are the tasks supervised or unsupervised?

(a) **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

(b) **Identification of most important supplier**:
A large online vendor would like to know which of its suppliers are the most important ones, i.e. contribute the most to revenue. Connections to these suppliers could be intensified, they could be taken over or a new logistics center could be established in the vicinity to reduce delivery times.

Data selection and simple aggregation. Could be answered using a common SQL Query:
`SELECT SUM(Revenue) FROM data GROUP BY supplier`
Not "Knowledge Discovery-according to our definition, but trivial knowledge".

(c) **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 as belonging to a paying customer. The characters on the plates are automatically recognized by a digital camera system.

Classification

(d) **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.

Outlier detection, sometimes classification (for known bots), clustering (to recognize strategies)

(e) **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.

Market basket analysis, association rules, frequent itemset mining

(f) **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.

Clustering / Classification (as categories)

(g) **Extraction of Data / Web Scraping**:
From a well-known movie database a list of movies and a list of actors shall be extracted. (Let licensing issues be ignored at this point.)

Data selection, by our definition, is not "Data Mining", rather it is an earlier step in the KDD-process. Therefore it is not in the scope of this lecture.

(h) **Image segmentation in medical image data**:
Segmentation is the process of segmenting an image in different parts. In medical imaging, segments usually correspond to different cell types, organs or pathologies, or other biologically relevant structures. Medical imaging is complicated by poor image quality, low contrast and noise or other types of uncertainty. Even though there exists a range of methods for image data, these usually need to be adapted to applications in the medical field. Terms in this subject area include:

(i) Atlas-Based Segmentation:
An expert produces evaluations for some example images which are then applied to to new images by extrapolation. Thereby, the training data is abstracted and a model is developed.
- Classification

(ii) Shape-Based Segmentation:
This method utilizes parameterized models of shapes describing certain structures. These shapes are then modified to fit to unknown images.
- Classification, maybe Outlier Detection

(iii) Interactive Segmentation:
A medical doctor supplies information during a surgery, such as a region or boarder of a segment. An algorithm then refines these preliminary results and is able to provide a more precise segmentation of a cell type.
- Clustering (semi-supervised)

Exercise 1-2 Attribute data types

For the attributes of the following data set, decide whether the attributes are numerical, nominal or ordinal.

Obs.	Sex	Height (cm)	Weight (kg)	Hair color	Blood type	Glasses	Smoker	Residential area
67	female	175	60	dark blonde / brown	A	no	occasionally	quiet
68	female	176	52	light blonde	AB	yes	occasionally	quiet
69	female	176	63	black	A	yes	rarely	very quiet
70	female	179	65	dark blonde / brown	0	yes	never	quiet
71	female	180	65	dark blonde / brown	B	yes	never	quiet
72	female	180	70	dark blonde / brown	A	yes	never	quiet
73	female	185	72	dark blonde / brown	B	no	never	very quiet
74	female	195	62	red	0	yes	very often	very quiet
75	female	203	62	red	AB	yes	very often	quiet
76	male	165	53	dark blonde / brown	A	no	rarely	quiet
77	male	169	63	dark blonde / brown	B	yes	rarely	quiet
78	male	169	72	dark blonde / brown	A	no	never	quiet
79	male	170	61	dark blonde / brown	A	no	never	very quiet
80	male	171	71	dark blonde / brown	A	no	often	noisy
81	male	173	61	black	A	yes	never	very quiet
82	male	173	63	red	A	no	rarely	noisy
83	male	173	67	dark blonde / brown	B	yes	never	quiet
84	male	175	68	dark blonde / brown	0	no	never	quiet
85	male	175	71	dark blonde / brown	AB	no	often	quiet
86	male	176	60	dark blonde / brown	A	no	rarely	quiet
87	male	177	64	dark blonde / brown	AB	no	never	very noisy

Attribute	Data type
Obs	nominal, if the primary key, otherwise numerical discrete
Sex	nominal, dichotom
Height (cm)	numerical, discrete
Weight (kg)	numerical, discrete
Hair color	nominal, possibly ordinal (when considering hair color in a color spectrum)
Blood type	nominal
Glasses	nominal, dichotom
Smoker	ordinal
Residential Area	ordinal

Exercise 1-3 Distance functions

Distance functions can be classified into the following categories:

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

So if a distance measure satisfies $d : D \times D \to \mathbb{R}_0^+$ and for any vector $o, p, q \in D$: is reflexive, symmetric and strict and also satisfies the triangle inequality, then it is a metric.

As you can see, a pre-metric does not necessarily need to be *strictly* reflexive. Make sure you understand the difference between reflexivity and strictness!

Note: these terms as well as "distance function" are used inconsistently in literature. In mathematics, "distance function" is commonly used synonymous with "metric". In a database (and thus data mining) context, strictness

is often not relevant at all, and a "distance function" usually refers to a pseudo-metric, pre-metric or even dissimilarity function. Do not rely on Wikipedia, it uses multiple definitions within itself!

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

(a) $d(o,p) = \sum_{i=1}^{n}(o_i - p_i)$

d is not even positive definite: it can become negative.

(b) $d(o,p) = \sum_{i=1}^{n}(o_i - p_i)^2$

d is reflexive, symmetric, strict, but the triangle inequality is not satisfied.
Counter example: Consider $o = (0,0)$, $p = (1,0)$, $q = (2,0)$:

$$d(o,q) = 4 \geq 1 + 1 = d(o,p) + d(p,q)$$

(c) $d(o,p) = \sqrt{\sum_{i=1}^{n-1}(o_i - p_i)^2}$

d is reflexive, symmetric, satisfies the triangle inequality, but is not strict.

(d) $d(o,p) = \sum_{i=1}^{n} \begin{cases} 1 & \text{iff} \quad o_i = p_i \\ 0 & \text{iff} \quad o_i \neq p_i \end{cases}$

d is not reflexive – the other properties are irrelevant to us.

(e) $d(o,p) = \sum_{i=1}^{n} \begin{cases} 1 & \text{iff} \quad o_i \neq p_i \\ 0 & \text{iff} \quad o_i = p_i \end{cases}$

d defines the so-called Hamming distance, a metric which plays an important role in information theory. On binary representations, it corresponds to the number of ones after an XOR operation of two binary vectors.

Reflexivity, symmetry and strictness should be obvious.

Proof of triangle inequality by case distinction on the individual positions:

(i) $o_i = p_i \wedge o_i = q_i$:

$$\begin{aligned} d(o_i,p_i) + d(p_i,q_i) &\geq d(o_i,q_i) \\ d(o_i,o_i) + d(p_i,o_i) &\geq d(o_i,o_i) \\ 0 + 0 &\geq 0 \end{aligned}$$

(ii) $o_i = p_i \wedge o_i \neq q_i$:

$$\begin{aligned} d(o_i,p_i) + d(p_i,q_i) &\geq d(o_i,q_i) \\ d(o_i,o_i) + d(o_i,q_i) &\geq d(o_i,q_i) \\ 0 + 1 &\geq 1 \end{aligned}$$

(iii) $o_i = q_i \wedge o_i \neq p_i$:

$$\begin{aligned} d(o_i,p_i) + d(p_i,q_i) &\geq d(o_i,q_i) \\ d(o_i,p_i) + d(p_i,o_i) &\geq d(o_i,o_i) \\ 1 + 1 &\geq 0 \end{aligned}$$

(iv) $o_i \neq p_i \wedge p_i = q_i$:

$$\begin{aligned} d(o_i,p_i) + d(p_i,q_i) &\geq d(o_i,q_i) \\ d(o_i,p_i) + d(p_i,p_i) &\geq d(o_i,p_i) \\ 1 + 0 &\geq 1 \end{aligned}$$

(v) $o_i \neq p_i \wedge p_i \neq q_i \wedge o_i \neq q_i$:

$$\begin{aligned} d(o_i, p_i) + d(p_i, q_i) &\geq d(o_i, q_i) \\ 1 + 1 &\geq 1 \end{aligned}$$

Which implies:

$$\begin{aligned} d(o, p) + d(p, q) &= \sum_{i}^{n} d(o_i, p_i) + \sum_{i}^{n} d(p_i, q_i) \\ &= \sum_{i}^{n} (d(o_i, p_i) + d(p_i, q_i)) \\ &\geq \sum_{i}^{n} d(o_i, q_i) \\ &= d(o, q) \end{aligned}$$