Ludwig-Maximilians-Universität München
Institut für Informatik
Prof. Dr. Thomas Seidl
Julian Busch, Evgeniy Faerman,
Florian Richter, Klaus Schmid

Knowledge Discovery in Databases

SS 2016

Exercise 1: Data Mining Tasks, Distance Functions

Regarding tutorials on 20.04.-22.04.2016.

Exercise 1-1 Data mining tasks

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

(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.

(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.

(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.

(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.

(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 1-2 Distance functions

Distance functions can be classified into the following categories:

$d : O \times O \to \mathbb{R}_0^+$ $x, y, z \in O$:	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	×	×	×	×

So if a distance measure satisfies $d : O \times O \to \mathbb{R}_0^+$ and for any vector $x, y, z \in O$: 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! Both directions yield the identity of indiscernibles as defined in the lecture slides.

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!

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

- Euclidean distance $dist(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$
- Weighted Euclidean distance $dist(x, y) = \sqrt{\sum_{i=1}^{n} w_i (x_i - y_i)^2}$, where $w = (0.6, 0.1)$
- Quadratic form distance $dist(x, y) = \sqrt{\sum_{i=1}^{n} \sum_{j=1}^{n} (x_i - y_i) w_i (x_j - y_j)}$,
  where $w = \begin{pmatrix} 1.25 & -0.75 \\ -0.75 & 1.25 \end{pmatrix}$

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