Ludwig-Maximilians-Universität München
Institut für Informatik
Dr. Eirini Ntoutsi
Erich Schubert

Knowledge Discovery in Databases
SS 2012

Übungsblatt 1: Data mining tasks, feature scales

Aufgabe 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.

Aufgabe 1-2   *Feature scales*

For the features of the following data set, decide the feature types.

Obs.	Geschlecht	Grösse (cm)	Gewicht (kg)	Haarfarbe	Blutgruppe	Brille	Rauchen	Wohnlage
67	Frau	175	60	dunkelbl./braun	A	nein	gelegentlich	ruhig
68	Frau	176	52	hellblond	AB	ja	gelegentlich	ruhig
69	Frau	176	63	schwarz	A	ja	selten	sehr ruhig
70	Frau	179	65	dunkelbl./braun	0	ja	nie	ruhig
71	Frau	180	62	dunkelbl./braun	B	ja	nie	ruhig
72	Frau	180	70	dunkelbl./braun	A	ja	nie	ruhig
73	Frau	185	72	dunkelbl./braun	B	nein	nie	sehr ruhig
74	Frau	195	62	rot	0	ja	sehr viel	sehr ruhig
75	Frau	203	62	rot	AB	ja	sehr viel	sehr lärmig
76	Mann	165	53	dunkelbl./braun	A	nein	selten	ruhig
77	Mann	169	63	dunkelbl./braun	B	ja	selten	ruhig
78	Mann	169	72	dunkelbl./braun	A	nein	nie	ruhig
79	Mann	170	61	dunkelbl./braun	A	nein	nie	sehr ruhig
80	Mann	171	71	dunkelbl./braun	A	nein	viel	lärmig
81	Mann	173	61	schwarz	A	ja	nie	sehr ruhig
82	Mann	173	63	rot	A	nein	selten	lärmig
83	Mann	173	67	dunkelbl./braun	B	ja	nie	ruhig
84	Mann	175	68	dunkelbl./braun	.	nein	nie	ruhig
85	Mann	175	71	dunkelbl./braun	AB	nein	viel	ruhig
86	Mann	176	60	dunkelbl./braun	A	nein	selten	ruhig
87	Mann	177	64	dunkelbl./braun	AB	nein	nie	sehr lärmig