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 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 recognition happens fully automatically by a digital camera system.

Classification

(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

(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.

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

(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 trilogy. A related task might be suggesting additional products for already chosen products as a bundled offer.

Market basket analysis, association rules

(e) **News Aggregation**
A news summary website 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 and Classification (for categories)

(f) 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", but a previous step in the KDD-process. It is not a focus of this lecture.

(g) 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 by a common DBMS:
SELECT SUM(Revenue) FROM data GROUP BY supplier
Not "Data Mining-according to our definition, but trivial knowledge”.

(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:
For this method, parametric models of shapes, which focus on particular structures or curves, are often employed. Thereby, the shape is exchanged to match a new image.
- 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.

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

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Data type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Obs</td>
<td>nominal, if the primary key, otherwise numerical discrete</td>
</tr>
<tr>
<td>Gender</td>
<td>nominal, dichotom</td>
</tr>
<tr>
<td>Height (cm)</td>
<td>numerical, discrete</td>
</tr>
<tr>
<td>Weight (kg)</td>
<td>numerical, discrete</td>
</tr>
<tr>
<td>Hair color</td>
<td>nominal, possibly ordinal</td>
</tr>
<tr>
<td>Blood type</td>
<td>nominal</td>
</tr>
<tr>
<td>Glasses</td>
<td>nominal, dichotom</td>
</tr>
<tr>
<td>Smoker</td>
<td>nominal, possibly ordinal</td>
</tr>
<tr>
<td>Residential Area</td>
<td>nominal, possibly ordinal</td>
</tr>
</tbody>
</table>

Exercise 1-3  Distance functions

Distance functions can be classified into the following categories:

<table>
<thead>
<tr>
<th>$d: D \times D \to \mathbb{R}_0^+$</th>
<th>reflexive reflexiv</th>
<th>symmetric symmetrisch</th>
<th>strict strikt</th>
<th>Triangle inequality Dreiecksungleichung</th>
</tr>
</thead>
<tbody>
<tr>
<td>$o, p, q \in D$</td>
<td>$o = p \Rightarrow d(o, p) = 0$</td>
<td>$d(o, p) = d(p, o)$</td>
<td>$d(o, p) = 0 \Rightarrow o = p$</td>
<td>$d(o, q) \leq d(o, p) + d(p, q)$</td>
</tr>
</tbody>
</table>

Dissimilarity function
Unähnlichkeitsfunktion ×

(Symmetric) Pre-metric
(Symmetrische) Prämertik × ×

Semi-metric, Ultra-metric
Semimertik, 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) \( d(o, p) = \sum_{i=1}^{n} \begin{cases} 1 & \text{iff } o_i = p_i \\ 0 & \text{iff } o_i \neq p_i \end{cases} \)
\( d \) is not reflexive – the other properties are irrelevant to us.

(e) \( d(o, p) = \sum_{i=1}^{n} \begin{cases} 1 & \text{iff } o_i \neq p_i \\ 0 & \text{iff } o_i = p_i \end{cases} \)
\( d \) 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) \( o_i = p_i \wedge o_i = q_i \):
\[
d(o, p) + d(p, q) \geq d(o, q) \]
\[
d(o, o) + d(p, p) \geq d(o, p) + d(p, p) \geq 0 + 0 \geq 0
\]

(ii) \( o_i = p_i \wedge o_i \neq q_i \):
\[
d(o, p) + d(p, q) \geq d(o, q) \]
\[
d(o, o) + d(p, q) \geq d(o, o) + d(p, q) \geq 0 + 1 \geq 1
\]

(iii) \( o_i = q_i \wedge o_i \neq p_i \):
\[
d(o, p) + d(p, q) \geq d(o, q) \]
\[
d(o, p) + d(p, p) \geq d(o, o) + d(p, p) \geq 1 + 1 \geq 0
\]

(iv) \( o_i \neq p_i \wedge p_i = q_i \):
\[
d(o, p) + d(p, q) \geq d(o, q) \]
\[
d(o, q) + d(q, q) \geq d(o, q) + d(q, q) \geq 1 + 0 \geq 1
\]
(v) \( o_i \neq p_i \land p_i \neq q_i \land o_i \neq q_i \):

\[
d(o_i, p_i) + d(p_i, q_i) \geq d(o_i, q_i)
\]

\[
1 + 1 \geq 1
\]

Which implies:

\[
d(o, p) + d(p, q) = \sum_{i=1}^{n} d(o_i, p_i) + \sum_{i=1}^{n} d(p_i, q_i)
\]

\[
= \sum_{i=1}^{n} (d(o_i, p_i) + d(p_i, q_i))
\]

\[
\geq \sum_{i=1}^{n} d(o_i, q_i)
\]

\[
= d(o, q)
\]