In [ ]:
```
## settings ↔
```

\usepackage*amssymb*

4. Clustering Methods

4.1. Relevance for Data Mining

- Clusters are the simplest form of a structure
- prevalence of clusters suggests relatedness / affinity and presence of correlations.
- multidimensional clusters can be interpreted as a simple form of rule
- cluster centers offer an economic description of the data (data reduction)

Note that the definition of an objective criterion for the discrimination of clusters is difficult.

4.2. Distance Measures

Starting point: Distance matrix

$$D = \begin{bmatrix} d_{11} & d_{12} & \dots & d_{1N} \\ \vdots & & & \vdots \\ d_{N1} & \dots & \dots & d_{NN} \end{bmatrix}$$

- Note that numbers mean the unsimiliarity (distance) between the corresponding data points
- Example: distance between the tastes of different pudding flavours

Requirements for a distance measure ($\forall i, j, k$)

$$\begin{aligned} d_{ij} &= d_{ji} && \text{symmetry} \\ d_{ij} &\geq 0 && \text{positive definite} \\ d_{ij} + d_{jk} &\geq d_{ik} && \text{triangle inequality} \end{aligned}$$

The definition / derivation of meaningful distances d_{ij} from the data depends on the fundamental question of the meaning (semantics) of the data:

- stating two data points x_i and x_j as similar (i.e. a small value of d_{ij}) requires a decision about what features seem to be meaningful.
- for that reason, there is no general procedure.

4.2.1 Distance measures for data points

Distance measure for real valued data vectors are

1. Euclidean Distance

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

- most simple, straightforward, and frequently used distance meassure
- but often insufficient as the following example illustrates
- Example: data set of broomstick features:
  - $\vec{x} = (\text{length in cm}, \text{diameter of the stick in cm})$
  - typical broomsticks {(150, 2.0), (158, 2.1), (165, 2.5), (180, 2.4), (170, 2.2)}
  - a difference of 10 cm in diameter is semantically much more 'different' than the equal difference in length
  - however, the euclidean distance of (160, 2.0) and (150, 12.0) to (150, 2.0) is the same dissimiliarity!

2. Pearson- or χ^2 distance

$$d(\vec{x},\vec{y}) = \sqrt{\sum_i \frac{(x_i-y_i)^2}{\sigma_i^2}}$$

with $\sigma_i^2 = \langle (x_i - \langle x\rangle)^2\rangle_x$ as scaling factor

- (+) leads to a more balanced weighting of different dimensions for the result
- (-) Pearson-distance assumes uncorrelated vector components x_i. This presumption is often not valid
  - Example: repeated features within the vector, e.g. with different units, or basically a 1:1-dependency

$$\vec{x} = \left.\begin{pmatrix} u \\ \vdots \\ u \\ v \end{pmatrix}\right\} (n-1)-\text{times}$$

- Iso-Distance-surface to a vector $\vec{x}$ is an ellipsoid whose principal axes are aligned with the coordinates.
  - With reference to this, an improved distance measure is

3. Mahalanobis-Distance:

$$d(\vec{x},\vec{y}) = \sqrt{(\vec{x}-\vec{y})^\tau \Sigma^{-1} (\vec{x}-\vec{y})}$$

with

$$\Sigma = \langle (\vec{x}-\langle\vec{x}\rangle)(\vec{x}-\langle\vec{x}\rangle)^\tau\rangle$$

- This basically scales as pearson, but with prior change into the PCA basis
- Iso-distance surface around $\vec{x}$ are now rotated ellipsoids (according to the variance ellipsoid of the whole data set)
- Example:
  - w.l.o.g. let $\vec{y} = 0$.
  - Let's look at the vector of the ith principal component of length $\sqrt{\lambda_i}$, so $\vec{x} = \sqrt{\lambda_i}\hat{u}_i$.
  - Then the distance d

$$d = \sqrt{\lambda_i}\hat{u}_i^\tau [UD^{-1}U^\tau]\sqrt{\lambda_i}\hat{u}_i = \sqrt{\lambda_i}\lambda_i^{-1}\sqrt{\lambda_i} = 1$$

  - so the iso-distance surface for the distance 1 has an extension of λ_i along the eigenvector $\hat{u}_i$
- Attention: Scaling can sometimes even destroy a prevalent clustering structure

4. `City Block'-Distance

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

- 'distance to be traveled if streets and avenues are perpendicular, as in Manhattan

5. Supremum Distance

$$d(\vec{x}, \vec{y}) = \max_{i=1\ldots d} |y_i - x_i|$$

- that is the largest component of the city-block sum terms.

6. Minkowski-Distance

$$d(x, y) = \left\{ \sum_{i=1}^{d} |y_i - x_i|^p \right\}^{1/p}$$

Some of the above distance measures result as special case, namely:

- $p = 1$: City-Block distance
- $p = 2$: Euklidean distance
- $p \to \infty$: Supremum distance

Remarks:

- All above distance measures assume implicityl a topology of $\mathbb{R}^n$
- They are not suitable to represent angles (which have a topology of a circle)
- Different topologies can be tackled by embedding into a suitable $\mathbb{R}^m$
- Example:
  - $\phi_1 = 0$ and $\phi_2 = 2\pi$ represent the same angle, but have a numeric distance of 2π.
  - Embedding the angle variable into a 2D-space by $(\cos(\phi), \sin(\phi))$ gives a representation where this problem does not occur anymore.
- Dealing with nominal attributes
  - if a fixed number (e.g. K) of alternative values are given, the variable can be embedded into a K-dimensional vector space, e.g.

$$\{\text{vanille, chocolade, strawberry}\} \Rightarrow \left\{ \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \right\}$$

  - Using a numbering $\{V, C, S\} \Rightarrow \{1, 2, 3\}$ would instead induce an ordering (e.g. $d(V, C) < d(V, S)$)
  - Trick: if many value alternatives are given, the embedding space would become inadequately high-dimensional. A sometimes acceptable compromise is then to use random projections: select K random vectors of length 1 in a vector space $\mathbb{R}^L,\ L < K$. If L is large enough, the vectors are approximately orthogonal on eachother and therefore more or less uncorrelated.

4.2.2. Distance measures between clusters

Many clustering methods require a distance between clusters. Let X, Y be clusters, we can define the following frequently used distance measures

$$d_1(X,Y) = \min_{\substack{\vec{x}\in X \\ \vec{y}\in Y}} d(\vec{x},\vec{y}) \text{ minimal distance}$$

$$d_2(X,Y) = \max_{\substack{\vec{x}\in X \\ \vec{y}\in Y}} d(\vec{x},\vec{y}) \text{ maximal distance}$$

$$d_3(X,Y) = \frac{1}{N_X N_Y} \sum_{\substack{\vec{x}\in X \\ \vec{y}\in Y}} d(\vec{x},\vec{y}) \text{ average distance}$$

$$d_4(X,Y) = d\left(\frac{1}{N_X}\sum_{\vec{x}\in X}\vec{x}, \frac{1}{N_Y}\sum_{\vec{y}\in Y}\vec{y}\right) \text{ centroid distance}$$