Cluster Analysis

Contents

- Required packages

```
knitr::opts_chunk$set(echo = TRUE)
#install.packages("dplyr","ade4","magrittr","cluster","factoextra","cluster.datasets")

knitr::opts_chunk$set(echo = TRUE)
```

Definition of a distance

- A distance function or a metric on $\mathbb{R}^n$, $n \geq 1$, is a function $d : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$.
- A distance function must satisfy some required properties or axioms.
- There are three main axioms.
- A1. $d(\mathbf{x}, \mathbf{y}) = 0 \iff \mathbf{x} = \mathbf{y}$ (identity of indiscernibles);
- A2. $d(\mathbf{x}, \mathbf{y}) = d(\mathbf{y}, \mathbf{x})$ (symmetry);
- A3. $d(\mathbf{x}, \mathbf{z}) \leq d(\mathbf{x}, \mathbf{y}) + d(\mathbf{y}, \mathbf{z})$ (triangle inequality), where $\mathbf{x} = (x_1, \cdots, x_n)$, $\mathbf{y} = (y_1, \cdots, y_n)$ and $\mathbf{z} = (z_1, \cdots, z_n)$ are all vectors of $\mathbb{R}^n$.
- We should use the term *dissimilarity* rather than *distance* when not all the three axioms A1-A3 are valid.
- Most of the time, we shall use, with some abuse of vocabulary, the term distance.