Clusteren — **introductie**

Stel we hebben een n-tal objecten (datapunten), en we willen die verdelen in een aantal *clusters*. Hoeveel? Dat weten we misschien nog niet.

Het betreft een vorm van *unsupervised* leren, in tegenstelling tot *classificatie*: daar weten we de indeling in groepen al, en willen we een nieuw object in de juiste groep krijgen (met bijvoorbeeld ID3).

De twee hoofdrichtingen binnen clustering zijn:

- *hierarchisch*: voeg kleine(re) clusters samen tot steeds grotere

- *niet-hierarchisch* of *partitioneren*: stap voor stap verbeteren van een bestaande clustering

Clusteren — **hierarchisch**

Het basisalgoritme is:

- stop ieder object in een eigen cluster
- herhaal (totdat je meent klaar te zijn): voeg de twee dichtstbijzijnde clusters samen

Nodig is een afstandsbegrip tussen objecten, oftewel: een *dissimilarity-matrix*. Twee voorbeelden voor de afstand d tussen objecten i en j, die gerepresenteerd worden door vectoren in een p-dimensionale ruimte, zijn de *Manhattan-afstand* en de *Euclidische afstand*:

$$d(i,j) = \sum_{\ell=1}^{p} |x_{i\ell} - x_{j\ell}|$$

$$d(i,j) = \sqrt{\sum_{\ell=1}^{p} (x_{i\ell} - x_{j\ell})^2}$$

Clusteren — **afstand tussen clusters**

Als we de afstand tussen elk tweetal objecten weten, wat is dan de afstand tussen twee clusters A en B? Er zijn verscheidene mogelijkheden, onder andere:

- *single linkage*:

$$d(A,B) = \min \{d(a,b) | a \in A, b \in B\}$$

- *complete linkage*:

$$d(A,B) = \max \{d(a,b) | a \in A, b \in B\}$$

- *average linkage*:

$$d(A,B) = \text{mean } \{d(a,b) | a \in A, b \in B\}$$

Clusteren — **niet-hierarchisch**

Het basisalgoritme is hier:

- begin met een willekeurige clustering in k clusters (nadeel: k ligt vooraf vast)

- herhaal (totdat je meent klaar te zijn): stop één object in een andere cluster, zodanig dat de totale kwaliteit verbetert

Nodig is een kwaliteitsmaat voor clusters. Voorbeelden zijn:

$$\sum_{i=1}^{n} d(i, c(i))^2 \qquad \text{of} \qquad \sum_{i=1}^{n} d(i, m(i))^2$$

Hierbij is $c(i)$ het gemiddelde (het zwaartepunt) van de objecten van de cluster waar object i in zit, en $m(i)$ is het "steunpunt" van de cluster waar object i in zit. Die steunpunten zijn "representatieve" objecten.
De tweede mogelijkheid leidt tot zogeheten k-*means clustering*.