Clusteringsalgoritmen

Bachelorproef

Victor Miclotte

Universiteit Gent

26 mei 2016

Inhoud

Motivatie

Veel toepassingen in verschillende takken van de wetenschap.

- Biologie: genexpressies
- Economie: prijsfluctuaties
- Taalkunde: teksten vergelijken
- ...

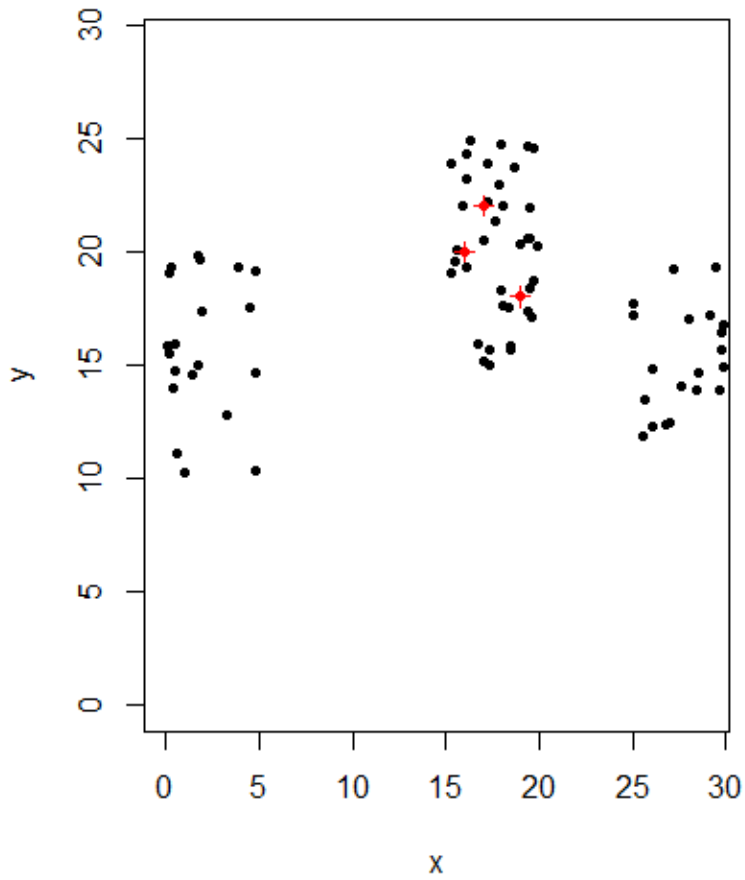
k-means

- Partitioneel clusteringsalgoritme
- Elke cluster wordt geassocieerd met een center
- k op voorhand vastgelegd

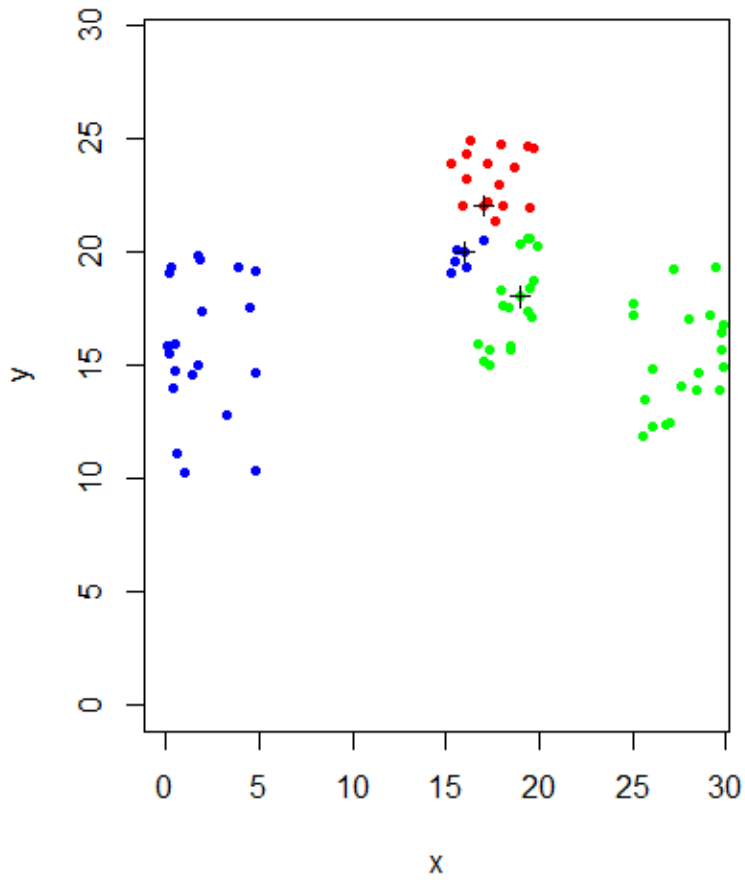
Algoritme 1 *k*-means

1: Selecteer k punten als initiële centers.
2: **while** centers zijn veranderd **do**
3: Voeg elk punt toe aan cluster met dichtste center.
4: Herbereken centers van elke cluster.
5: **end while**

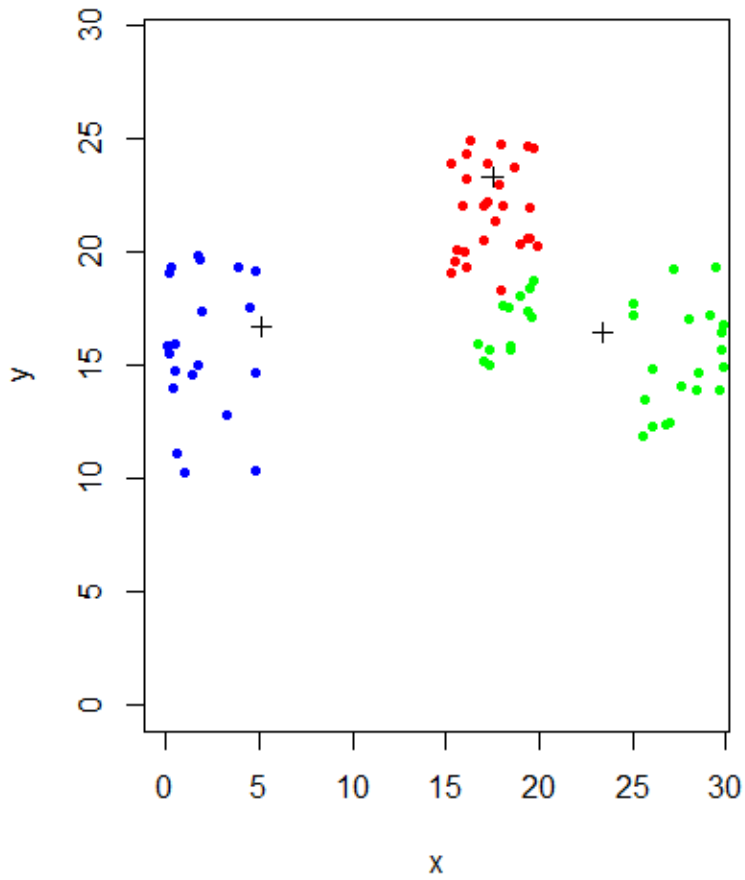
k-means: voorbeeld



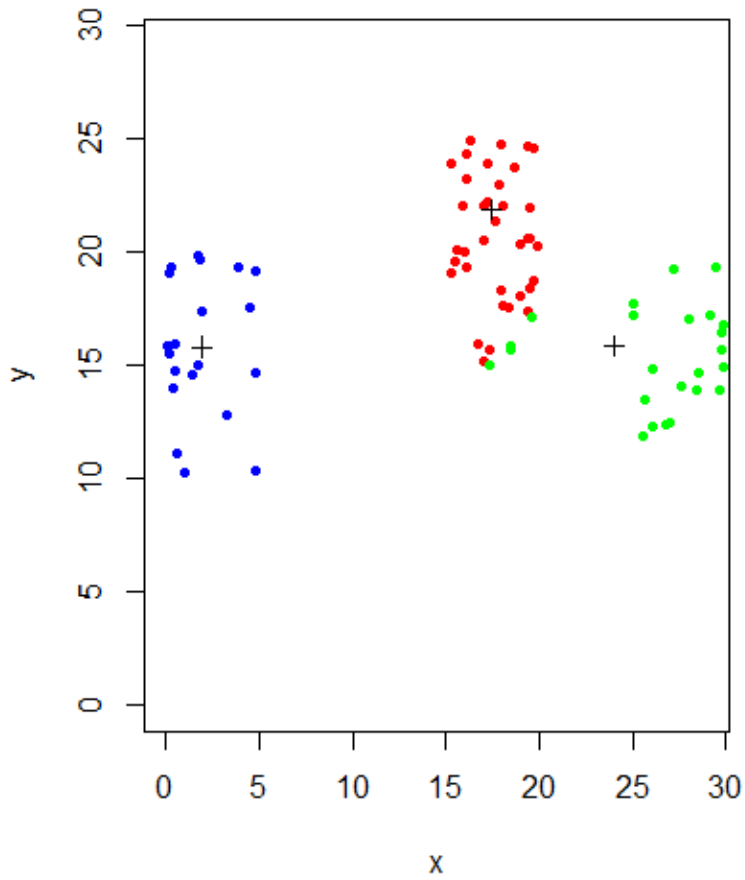
k-means: voorbeeld



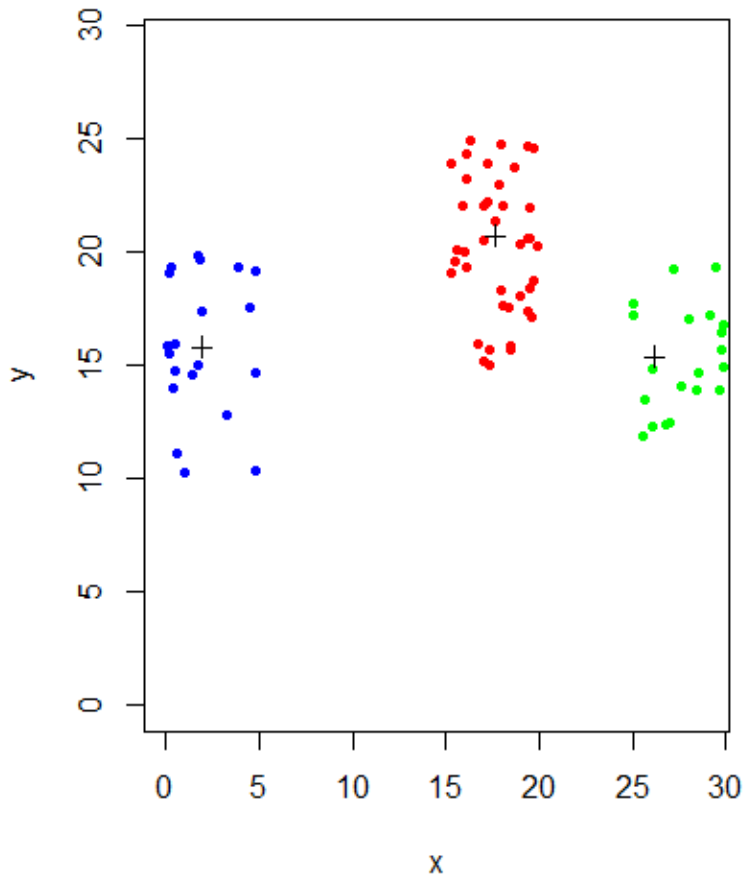
k-means: voorbeeld



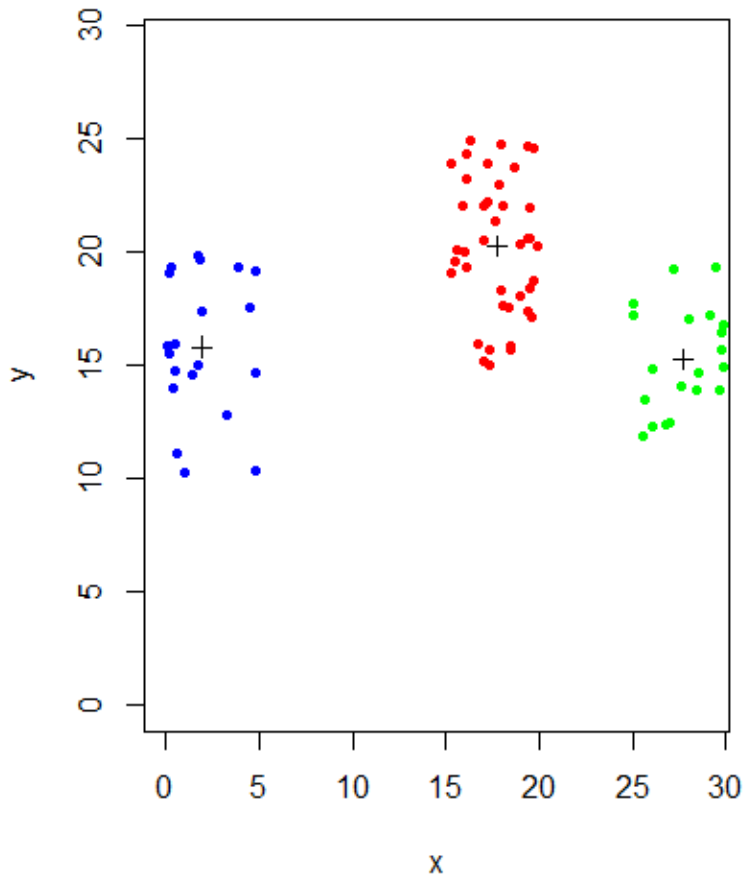
k-means: voorbeeld



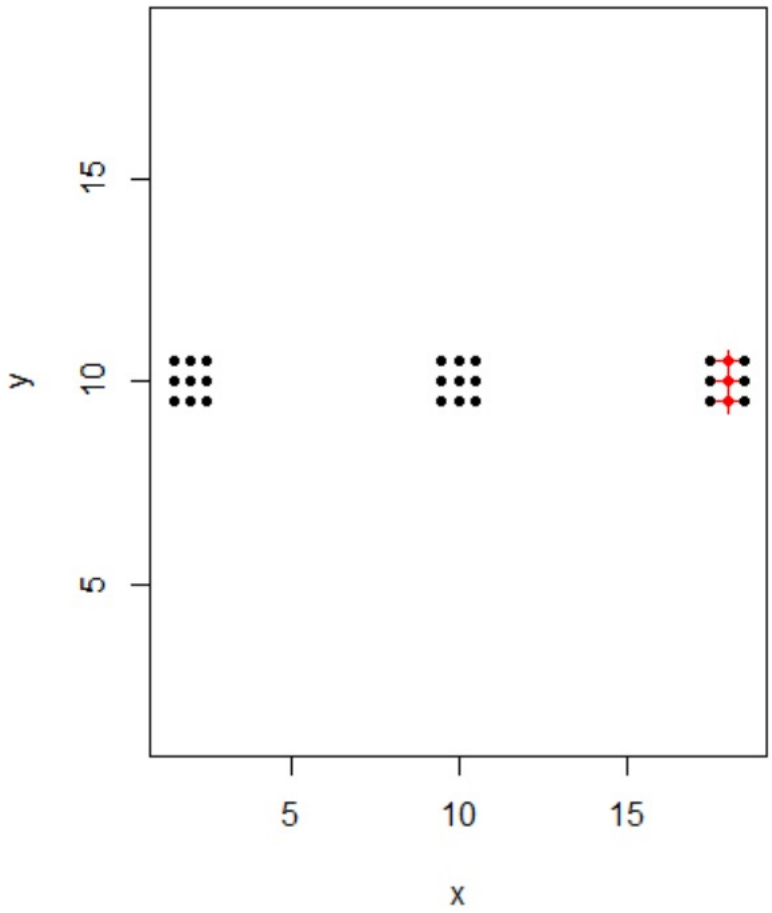
k-means: voorbeeld

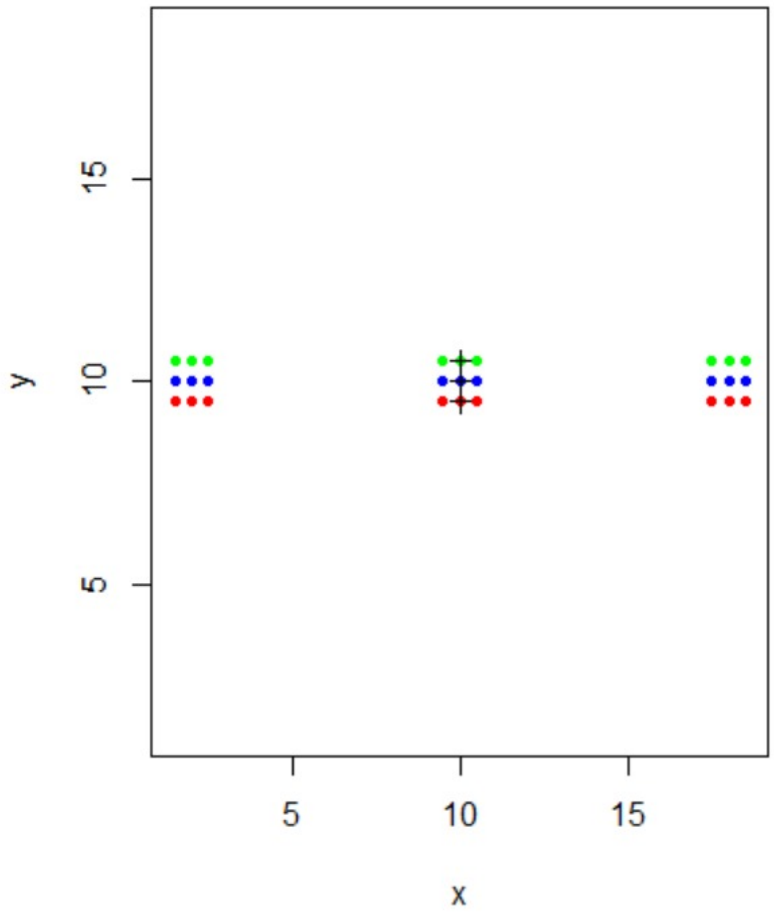


k-means: voorbeeld



Initiële centers





Andere problemen

- Lege clusters
- Niet-sferische clusters
- Clusters zijn in elkaar verweven

Uitbreidingen

- Pre-processing
- Post-processing: verlagen van de SSE
- Bisecting *k*-means

Post-processing: verlagen van de SSE

Twee fasen:

- Aantal clusters verhogen en SSE verlagen
- Clusters samenvoegen waarbij SSE zo weinig mogelijk stijgt

Hiërarchisch clusteren

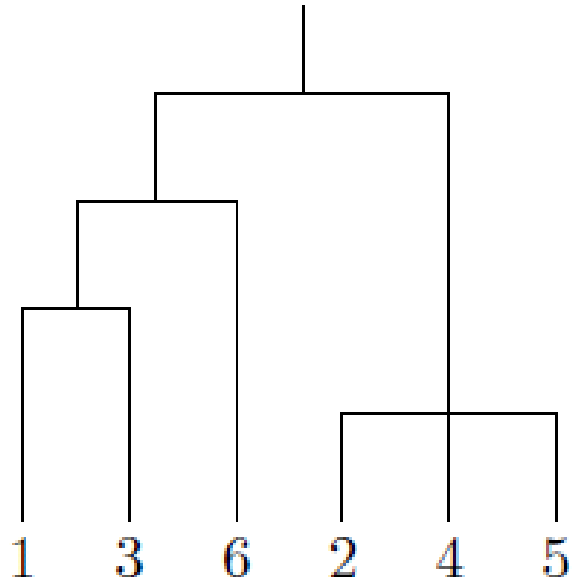
- Sequentie van partitionele clusteringen
- Afstandsmatrix
- Single link / Complete link / Group average

Algoritme 2 Hiërarchisch clusteren

Input: een $n \times n$-afstandsmatrix.

1: Maak n clusters met elk één punt.
2: **while** Er is meer dan één cluster **do**
3: Voeg de twee dichtste clusters samen.
4: Pas de afstandsmatrix aan.
5: **end while**

Dendrogram

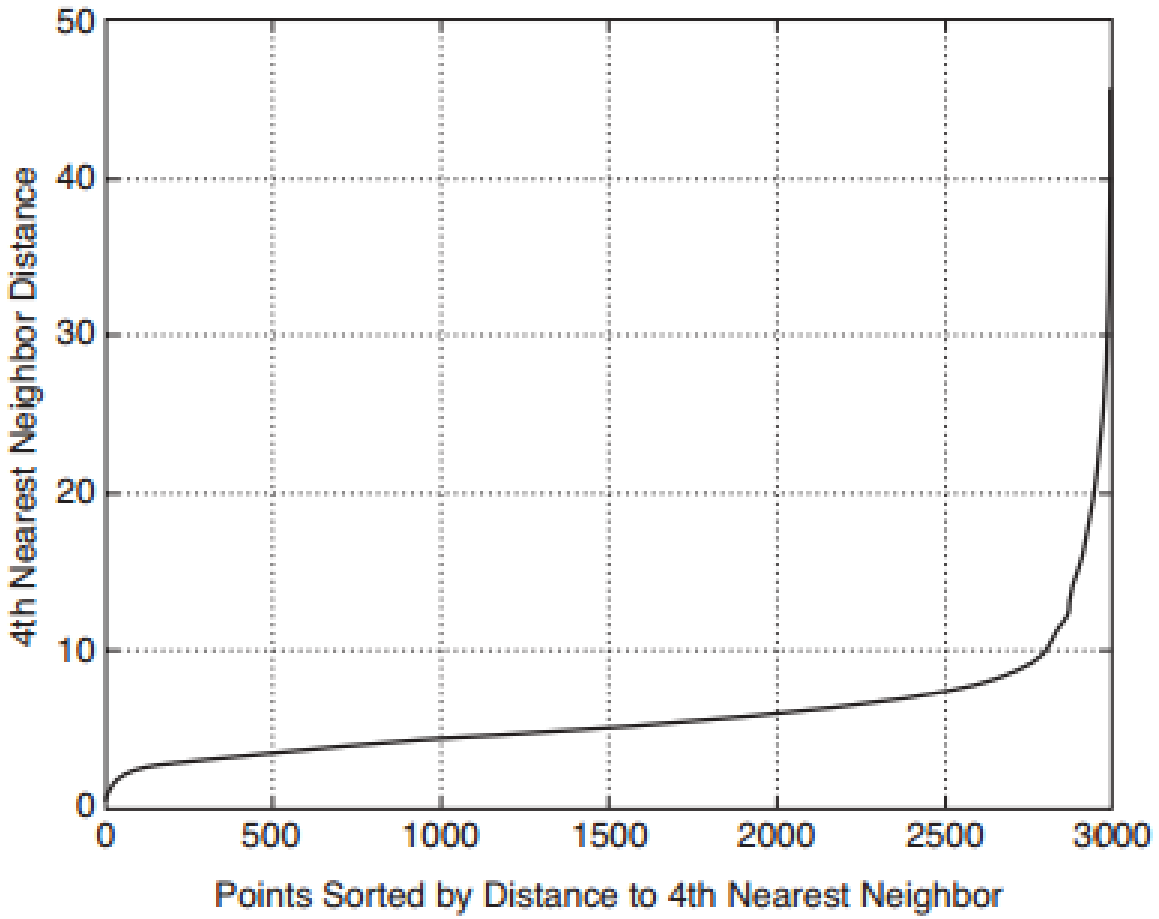


DBScan

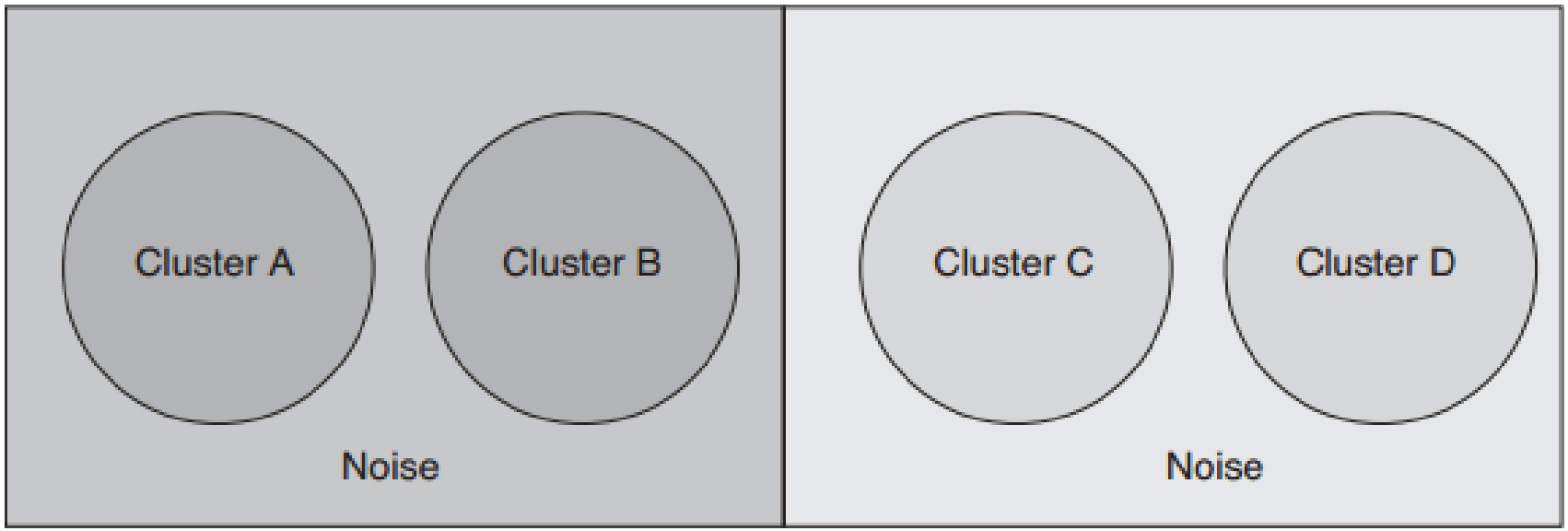
Densiteit (= aantal punten binnen bepaalde straal ϵ) i.p.v. afstanden gebruiken om clusters te vormen.

- *MinPts*: density-treshold
- **Kernpunt:** punt met dichtheid $\geq$ *MinPts*
- **Randpunt:** punt in omgeving van kernpunt
- **Ruispunt:** elk ander punt

Bepalen van *MinPts* en ϵ



Clusters met verschillende densiteit



Clusterevaluatie

- Kwaliteit nagaan: cohesie en separatie
- Aantal clusters bepalen
- Geschikt voor clustering