# DA – Machine Learning – huiswerk week 6

In sklearn.cluster vind je een implementatie van het K-Means algoritme, dat als volgt werkt:

1. gegeven een dataset (features) en , het aantal gewenste cluster;
2. kies willekeurig centroids (clustercentra);
3. bepaal voor ieder datapunt bij welke centroid hij hoort (kleinste Euclidische afstand);
4. als de clustering van de datapunten niet is veranderd ten opzichte van de laatste stap,  
   stop dan het algoritme;
5. bepaal voor ieder cluster het gemiddelde van alle datapunten in het cluster en benoem dit tot het nieuwe clustercentrum;
6. ga verder bij stap 3.

Het lijkt me dat dit algoritme voldoende eenvoudig is om zelf te implementeren. Implementeer het algoritme als een functie in Python (die uit meerdere deelfuncties mag bestaan). De functie heeft twee argumenten: de dataset die wordt geclusterd en het gewenste aantal clusters.

De functie heeft twee outputs: voor ieder cluster wordt de centroid teruggegeven en voor ieder punt in de dataset wordt teruggegeven bij welk cluster het hoort.

Probeer de functie correct, efficiënt en zo charmant mogelijk te implementeren. Ik geef de voorkeur aan numpy-arrays als datastructuur voor alle meerdimensionale data, maar dat is geen verplichting.

Gebruik je zelfgebrouwen algoritme om de data in ***test\_clustering.txt*** te clusteren. Gaat dat goed?