Types clusteren

Hard of soft

Hard, een datapunt hoort of wel bij een cluster of niet. Het is binair

Soft, een datapunt krijgt een kans dat het bij een cluster hoort, dus 0.8 bij punt 1 0.15 bij 2 enz enz

Hierbij moet je zelf bepalen waar je de grens legt waar je het indeeld als je dat uiteindelijk wel hard moet gaan doen.

Er zijn mega veel clustering algoritmes, onhandig dus

Bepaalde categorieën zijn:

Connectivity modellen:  
als punten in de dataspace dicht bij elkaar liggen is de kans dat deze bij het zelfde cluster hoort groter dan dat ze verder uit elkaar liggen. Komt een beetje neer op nearest neightbour methode.

Of je defeineerd de clusters vooraf al en kijkt waar de data in valt

Andere optie is alle punten zijn een cluster en als er veel bij elkaar liggen dan partition je het zodat je uit eindelijk een paar clusters over houd.

Centroid modellen.

Iteratieve methode K-means is hier een voorbeeld van, je geeft hier al in aan hoeveel clusters je wil hebben op het einde en het algoritme deelt dan alles in hierbij is het dus wel belangrijk dat je al weet hoe je dataset er uitziet en een beetje weet wat voor verschillende dingen je hebt in je data. Stel bij dna heb je 4 clusters van ATCG base en dan kan je ze indelen op waterstof bruggen en gewicht oid waardoor je 4 clusters krijgt en waarvan je dan kunt bepalen welk datapunt welke base is mocht je dit nog niet weten(dit zou handig zijn als dataset om te leren hoe je clustert en dan is het BMT gerelateerd.)

Distribution modellen:

Dit is een model dat kansen gebruikt om te bepalen welk datapunt bij welk cluster hoort. Hierbij kun je of zelf al cluster punten kiezen of ze iteratief laten bepalen maar dit is wel een hele lastige en het zorgt niet voor een harde scheiding en vind ik echt super vaag

Density modellen

Dit is super handig als je echt veel groepen hebt in je data waar veel punten bij elkaar liggen. Dus elke regio met heel veel punten is een cluster en zo trek je de data uit elkaar.