Clusteranalyse

1 De Simpsons revisited

In deze oefening willen we nagaan welke soorten Simpsons er zijn.
Hiervoor gebruiken we de oorspronkelijke data (waarbij de variabelen continu zijn).
Je vindt deze in "simpsons_origineel.csv".

1. lees het bestand in. Bewaar de namen van de Simpsons in een aparte variabele. Wis daarna de kolommen "naam" en "geslacht" want deze zijn nominaal
2. bepaal de euclidische afstanden tussen de Simpsons (kijk eventueel ook eens naar deze functies: `scipy.spatial.distance.pdist en scipy.spatial.distance.squareform`)
3. welke Simpsons zijn het dichtst bij elkaar?
4. wat is de Manhattan afstand tussen Homer en Bart?
5. wat is de gestandaardiseerde euclidische afstand tussen Marge en Maggie?
6. maak nu een dendrogram (gebruik euclidische afstanden en complete linkage). Kan je de namen van de Simpsons onderaan in het diagram krijgen?
7. als je 2 clusters zou moeten selecteren uit het dendrogram, welke zijn dat dan? Hoe zou je die clusters benoemen?
8. maak een scatterplot van de Simpsons waarbij je in de X-as leeftijd zet en in de Y-as het gewicht. Gebruik het clusternummer voor de kleur.
9. bepaal nu 3 clusters uit het dendrogram en maak weer een plot. Welke Simpsons zitten nu in welke categorie?
10. probeer nu 2 clusters te maken met het k-means algoritme. Kijk naar de centroids. In welke coördinaten verschillen deze centroids het meest?
11. maak een beslissingsboom die bepaalt in welke cluster een Simpson valt (gegeven de clusters uit de vorige vraag). Gebruik hiervoor het CART algoritme.

2 Studenten

1. lees het bestand "studenten.csv". Hierin staan de scores van 100 studenten op bepaalde vakken
2. bereken de afstand tussen de eerste twee studenten op de volgende manieren:
    - euclidisch
    - gestandaardiseerd euclidisch
    - manhattan
    - gestandaardiseerd manhattan
3. we zoeken 4 clusters in deze set. Kies volgende centroids:
    - (9, 3, 14, 1, 6, 10, 10, 15)
    - (12, 18, 9, 5, 18, 1, 3, 18)
    - (6, 15, 13, 18, 9, 15, 20, 18)
    - (5, 4, 7, 18, 20, 17, 1, 15)

    Bereken nu de euclidische afstand van studenten 0, 9, 19 en 29 tot deze centroids
4. bij welke centroid zou je deze studenten indelen?
5. bepaal de categorieën (clusters) van iedere student adh van het K-means algoritme. Maak 4 clusters en gebruik 300 als seed. Hoeveel studenten zitten er in iedere categorie?
6. Zoek eens uit hoe je het algoritme kan laten starten met de 4 centroids die hierboven staan. Hoeveel studenten zitten er dan in iedere categorie?
7. Gebruik nu het CART algoritme om een beslissingsboom op te stellen die de cluster voorspelt (volgens de laatst gevonden clusters).
8. in welke categorie valt een nieuwe student met de volgende punten: (10, 15, 12, 11, 13, 14, 9, 10)?

3 Extraterrestrial life...

Na een lange reis, stort het ruimteschip van spaceman Spiff neer op een verre planeet. Alles lijkt rustig, maar Spiff ontdekt dat er levende wezens op deze planeet rondlopen. Als goed onderzoeker, begint hij onmiddellijk data te verzamelen over deze wezens. Met zijn "Mertilizer" kan hij de wezens vangen. Per wezen schrijft hij het volgende op:

- hoeveel poten hebben ze?
- hoeveel ogen hebben ze?
- lengte
- breedte
- hoogte
- heeft het wezen een staart?
- wat is de kleur?
- heeft het wezen vleugels?

Je vindt deze data in het bestand: "spiffs metingen.csv".

1. lees het bestand in. Zet de booleaanse waarden om naar 0 en 1. Verwijder de kolom kleur, want die is nominaal
2. maak een dendrogram van deze data. Gebruik de gestandaardiseerde euclidische afstand en average linkage
3. knip dit dendrogram zodat er 3 clusters zijn. Hoeveel aliens zitten er in iedere cluster?
4. maak een scatterplot waarbij je het aantal ogen in de X-as zet en de lengte in de Y-as. Gebruik de cluster als kleur. Wat zie je?
5. doe dit ook voor 4 clusters. Welke cluster werd er opgesplitst? Hoe kan je dit zichtbaar maken?
6. zoek met kmeans 3 clusters (gebruik ‘random_state=42’ als parameter zodat je altijd hetzelfde resultaat bekomt). Bekijk de centroids. Welke soorten wezens zijn er op deze planeet?
7. Zoek nu een beslissingsboom die bepaalt tot welke categorie een ruimtewezen behoort. Gebruik het CART algoritme. Welke eigenschap is volgens deze boom het belangrijkst in het onderscheid?