KU LEUVEN kulak

Feedback opdracht 1

Academiejaar 2018 – 2019 Statistische modellen en data-analyse

Algemeen

- Geef concrete info over de data: hoeveel en welke veranderlijken/observaties (minstens beschrijvend of zelfs opsommend in een tabel), bron van de gegevens.
- Tracht geen theorie samen te vatten in het verslag van een statistische analyse, ga er van uit dat de lezer de methode kent (in welk geval de uitleg overbodig is) en zorg dat de conclusie/interpretatie helder is (ook voor wie de methode en de data niet kent). Zelf uitleggen hoe een methode werkt, is bovendien meestal zeer onnauwkeurig en dus fout.
- Zorg steeds voor objectieve informatie (statistieken en grafieken) naast je eigen interpretatie zodat lezer zich een beeld kan vormen van wat je schrijft. Als de beste clustering Afrika scheidt van de rest van de wereld: geef dan een figuur die dat toont.
- Maak een strenge selectie van grafieken: toon geen twee keer (laat staan tien keer) hetzelfde, dat is verwarrend. Vergelijkbare resultaten moeten een meerwaarde bieden en dus moet het verschil uitgelegd worden. Is er geen duidelijk verschil, toon dan slechts één grafiek.
- Verzorg elke figuur in verslag tot in de puntjes. Assen, titel, legende en bijschrift van een grafiek bevatten telkens andere informatie, bijvoorbeeld:
  - Assen: PC1 en PC2
  - Titel: Principaalcomponenten van doodsoorzaken
  - Legende: overzicht van de gebruikte symbolen/kleuren/lijnstijlen (continent/ontwikkeling)
  - Bijschrift: geeft duiding bij de grafiek: niet wat er op staat (objectief, dat is al vervat in het vorige) maar wat de lezer er zou moeten op zien. Bijvoorbeeld: In Afrikaanse landen hebben infectieziekten de overhand (lage PC1) terwijl Europa eerder lijdt onder beschavingsziekten (lage PC2). De andere continenten hebben minder uitgesproken waarden voor de eerste principaalcomponent en zijn op basis hiervan verder niet te onderscheiden. Wat de tweede principaalcomponent betreft...
- Als je verwijst naar punten op de grafiek, zorg dan dat deze duidelijk gelabeld zijn (bijvoorbeeld een pijl of cijfer er bij of een opgevuld in plaats van hol symbool)
- Gebruik steeds vector graphics ('export as pdf' of 'copy as metafile'). Neem geen screenshots van tabellen uit R maar copieer het resultaat in een eenvoudig tabelletje in je tekstverwerker. Zoniet zijn grafieken/tekst korrelig, nauwelijks leesbaar, onprofessioneel.
- Verwijs niet naar grafieken die niet in het verslag staan.

Clustering

- Bestaande groepen herkennen is niet het eerste doel van clustering: tracht eerst (objectief) na te gaan of en hoeveel clusters er zijn (bijvoorbeeld met Silhouette width). Vergelijk pas in tweede instantie met bestaande groepen. Het is absoluut denkbaar dat er een duidelijke structuur gevonden wordt die niets te maken heeft met de gegeven groepen, dat bepaalde groepen door clustering als één worden gezien, of dat er binnen bestaande groepen meerdere clusters worden onderscheiden. Beperk je dus zeker niet tot de aantallen groepen in regio/ontwikkeling.
- Leg uit waarom je precies (niet-)gestandaardiseerde gegevens verkiest. De argumenten bij die keuze geven info aan de lezer. Minder clustering vinden is niet per se slecht, maar wil misschien gewoon zeggen dat er geen groepen in de data zitten ingebakken.

- Geef geen opsomming van alle modellen die je hebt gemaakt, haal er de duidelijkste/meest representatieve uit en vat verschil met andere kort samen (indien relevant).

Principaal

- Tracht principaalcomponenten te beschrijven in een zo vlot mogelijke tekst aan de hand van de hoogste loadings in de rotatiematrix en een aantal landen die extreem hoog of laag scoren voor die component. Enkel veranderlijken opsommen is niet zo duidelijk voor de lezer, tracht inzicht te geven (zie bijschrift hierboven).
- Let op de betekenis van een principaalcomponent: het gaat (meestal) over het contrast tussen twee (groepen) veranderlijken. Dus niet "de meeste mensen sterven aan hart- en vaatziekten"maar ër sterven relatief meer mensen aan hart- en vaatziekten dan aan kanker in vergelijking met andere landen"
- Bij principaalcomponentenanalyse is het wel een duidelijke meerwaarde om zoveel mogelijk info in zo weinig mogelijk veranderlijken te proppen. Dat kan een leidraad zijn bij het kiezen voor al dan niet gestandaardiseerde gegevens.

Normaliteit

- De gegevens zijn 'niet-normaal' kan nog vanalles betekenen, van eerder sferisch tot compleet scheef... in het eerste geval zullen veel methodes nog robuust genoeg zijn, in het laatste dringen zich misschien transformaties op.
- Argumenteer zeker wat de gevolgen van je bevindingen zijn op het vervolg van het onderzoek.

Classificatie

- APER is niet interessant als maat, aangezien ze inherent vertekend is. Als je ze vermeldt, hoort ze een meerwaarde te hebben.
- Een classificatie die niet perfect is, is daarom nog niet 'zeer slecht'.

Besluit

- Focus op wat je uit de data leert in verband met de onderzoeksvraag, de gebruikte methoden zijn hier niet meer zo relevant. Bijvoorbeeld: 'Afrika is redelijk goed van de rest van de wereld te onderscheiden' is een beter besluit dan 'clustering is niet mogelijk'. Schrijf eerder 'De mate waarin infectieziekten, kankers en hart- en vaatziekten doorwegen bij de doodsoorzaken van een land hangen in grote mate samen met de ontwikkelingsstatus' dan 'De eerste PC's verklaren 90% van de variantie'. Schrijf 'Het is in beperkte mate mogelijk om een land te klasseren in een bepaalde regio of volgens de ontwikkelingsstatus' en niet 'Classificatie gaat niet goed'.