KU LEUVEN kulak

Opdracht 1

Academiejaar 2019 – 2020 Statistische modellen en data-analyse

1 Toelichting

Het projectwerk is een onderdeel van het examen Statistische modellen en data-analyse, telt mee voor 3 van de 20 punten en wordt in groepen van twee studenten gemaakt.

Het is de bedoeling om de leerstof in de praktijk te gebruiken. Met behulp van onderstaande onderzoeksvragen en opdrachten worden dan gepaste analyses uitgevoerd en conclusies getrokken. De evaluatie gebeurt op basis van een script `naam1_voornaam1_naam2_voornaam2_Project1.R` met alle gebruikte commando's en een rapport `naam1_voornaam1_naam2_voornaam2_Project1.pdf` van *maximaal* 4 pagina's (zonder grafieken en tabellen mee te rekenen).

Vermeld op het titelblad van het rapport duidelijk jullie namen en studentennummers. Beide bestanden worden ingediend via Toledo ten laatste op vrijdag 24 april.

2 Chemische samenstelling van wijn naargelang het druivenras

In dit project wordt nagegaan of op basis van de chemische samenstelling van wijn is uit te maken met welk druivenras deze wijn is gemaakt. De dataset bevat gegevens over 178 soorten wijnen uit eenzelfde Italiaanse wijnstreek, maar van drie verschillende rassen. Telkens zijn dertien verschillende chemische eigenschappen geregistreerd. De verschillende veranderlijken worden beschreven in tabel 1.

2.1 Clustering

Onderzoek of de wijnen kunnen opgedeeld worden in groepen op basis van de 13 chemische eigenschappen. Zijn de gestandaardiseerde of de oorspronkelijke gegevens aangewezen? Beschrijf (indien van toepassing) de bekomen clusters. Zijn er overeenkomsten met de druivenrassen? Rapporteer enkel de duidelijkste resultaten en stel deze grafisch voor.

2.2 Multivariate normaliteit

Bespreek de (multivariate) verdeling van de data als geheel en in de drie afzonderlijke groepen: is deze normaal/sferisch/symmetrisch of zijn er duidelijke afwijkingen? Zijn er transformaties mogelijk die leiden tot een betere verdeling?

Onderzoek de hypothese van multivariaat normale verdeling van de chemische eigenschappen. Hou rekening met de conclusies in het vervolg van het onderzoek. Voorzie enkele typische en markante illustraties.

2.3 Principaalcomponentenanalyse

Voer principaalcomponentenanalyse uit op de dataset (13 chemische eigenschappen). Welke gegevens zijn hiervoor het interessantst: geschaald? getransformeerd? Hoeveel componenten zijn belangrijk? Tracht indien mogelijk deze componenten te interpreteren. Welke chemische eigenschappen zijn het belangrijkste?

2.4 Classificatie

Ga na in hoeverre het mogelijk is om het druivenras te identificeren aan de hand van chemische eigenschappen van de wijn. Ga te werk als volgt.

1. Selecteer willekeurig 50 wijnen als testset, 50 als validatieset en de rest als trainingsset.
2. Maak modellen met lineaire en kwadratische discriminantmethode en nearest neighbour classificatie voor alle relevante parameterwaarden op basis van de trainingssgegevens.

Tabel 1: Veranderlijken in de dataset druivensoorten

	Variabele	Omschrijving
1	druivenras	Categorische veranderlijke met labels 1, 2 en 3
2	alcohol	Indicator voor het suikergehalte van de druif
3	appelzuur	Belangrijkste zuur in wijn
4	asgehalte	Anorganisch materiaal (mineralen) dat overblijft na verbranding
5	alkaliteit	Alkaliteit van de as
6	magnesium	Hoeveelheid magnesium
7	fenolen	Aromatische verbindingen afkomstig van schil, pitten en steeltjes
8	flavonoiden	Type fenolen, draagt bij tot smaak en mondgevoel (onder meer tannines)
9	nonflavonoiden	Type fenolen, voornamelijk aromatische zuren
10	proanthocyanidine	Blauwe kleurstof in schil en pitten van druiven
11	intensiteit	Kleurintensiteit van de wijn
12	tint	Kleur van de wijn
13	proteines	Proteïne concentratie in de wijn
14	proline	Meest voorkomende aminozuur in wijn

3. Maak een grafiek van de error rate op basis van de validatiegegevens, in functie van het aantal buren. Bepaal voor hoeveel buren de error rate minimaal is.

4. Geef op de grafiek ook de error rate aan van het lineaire en kwadratische discriminantmodel en bepaal voor welk van beide modellen deze minimaal is.

5. Bereken nu voor beide gevonden modellen de error rate op basis van de testgegevens, rapporteer deze en beslis welke model de beste resultaten geeft.

Voorzie een afbeelding waarop alle gegevens volgens de beste methode worden ingedeeld. Gebruik kleuren en symbolen om het werkelijke en het voorspelde druivenras voor te stellen.

Instructies

Bundel al je commando's in één script en zorg dat het script correct werkt op basis van de originele gegevens. Verwijder alle overbodige lijnen en voeg zeer summier wat commentaar toe aan elke stap, in het bijzonder bij berekeningen die het verslag niet halen.

Neem van de uitvoer van het script enkel die statistieken en grafieken in je verslag over die werkelijk relevant zijn voor de opbouw van het verhaal. Noteer alle statistieken met de juiste eenheid en een gepast aantal beduidende cijfers. Zorg er voor dat je grafieken duidelijk leesbaar zijn en voorzien van titel, astitels en eenheden.

Maak van je rapport een degelijk wetenschappelijk verslag, een doorlopende tekst die los te lezen is van de opgave en begrijpelijk is voor een buitenstaander met dezelfde kennis van statistiek als jijzelf. Focus op de interpretatie, maar zorg er voor dat de lezer begrijpt hoe tot het gevonden model en bijhorende conclusies wordt gekomen.

Hou je aan de paginalimiet, bestandsnamen en deadline.