PCA Shiny

Dieses Skript ermöglicht es dem Benutzer den Iris Datensatz in 2D und 3D zu visualisieren und das PCA Ergebnis in einem 2D Plot anzuzeigen.
Zuerst liest das Skript den Iris Datensatz ein. Anschließend wird ein neues Dataframe erstellt das nur alle Spalten mit Zahlen beinhaltet, um anschließend eine PCA durchführen zu können. Die PCA Ergebnisse werden wiederum in einem Datenframe gespeichert.
Die Shiny Library erlaubt es nun aus dem originalen Datensart ein 2D und 3D Plot (mit Hilfe der car Library) zu erstellen und anzuzeigen. Der PCA Tab zeigt die PCA Ergebnisse in einer Tabelle an und erstellt ein 2D Plot davon.

Das Skript wurde getestet unter:
R version 3.6.2 (2019-12-12) -- "Dark and Stormy Night"
Ubuntu 16.04.6 LTS (Xenial Xerus)"
Kernel 4.4.0-173-generic

Installation

Um die passende Version von R auf Ubuntu 16 zu installieren müssen folgende Schritte im Terminal ausgeführt werden:

```
sudo add-apt-repository 'deb https://cloud.r-project.org/bin/linux/ubuntu xenial-cran35/'
sudo apt-key adv --keyserver keyserver.ubuntu.com --recv-keys E084DAB9
sudo apt-get install  libx11-dev mesa-common-dev libglu1-mesa-dev
sudo apt-get install r-base r-base-dev
sudo apt-get install libcurl4-openssl-dev
sudo add-apt-repository -y ppa:cran/imagemagick; sudo apt-get update; sudo apt-get install -y libmagick++-dev
```

Um die benötigten R libraries runterzuladen reicht es das Install.R Skript auszuführen:
```
Rscript Installs.R
```

Folgende Funktionen wurden in dem Skript benutzt:

read.csv - Diese Funktion liest die csv. Datei ein und speichert sie als data.frame
prcomp - Als input bekommt diese Funktion ein data.frame und erstellt dann eine PCA

Funktionen der Shiny Library:
tabPanel - Erstellt eine Seite die per Tab ansteuerbar ist
headerPanel - Erstellt ein Header auf einer Seite
sidebarPanel - Erstellt eine Sidebar Panel auf einer Seite
selectInput - Erstellt ein drop down Auswahlmenü auf einer Seite
reactive - Funktion die jedesmal wenn sich ein bestimmter Output ändert getriggert wird
plot - Erstellt ein 2D Plot aus einem dataframe mit zwei Spalten
scatter3d - Erstellt ein 3D Plot aus einem dataframe mit drei Spalten

Funktionen der rgl library
rglwidgetOutput - Container für ein rglwidget

rglwidget() - Erstellt eine html Version eines Plots

Funktion der car library
scatter3d - Erstellt ein 3D Plot aus einem dataframe mit drei Spalten

Die Daten

Der Iris Datensatz stammt aus der im Jahr 1936 erschienenen Veröffentlichung The use of multiple measurements in taxonomic problems" Annual Eugenics, 7, Part II, 179-188 von Fisher,R.A.

Jede Zeile der Tabelle repräsentiert eine Irisblume dar, einschließlich ihrer Art und Abmessungen ihrer Blätter, Kelchblatt und Blütenblatt, in Zentimetern.

Der Datensatz enthält 3 Klassen mit jeweils 50 Instanzen, wobei sich jede Klasse auf eine Art Irispflanze bezieht. Eine Klasse ist linear von den anderen 2 trennbar. Letztere sind nicht linear voneinander trennbar.

Attributinformationen:

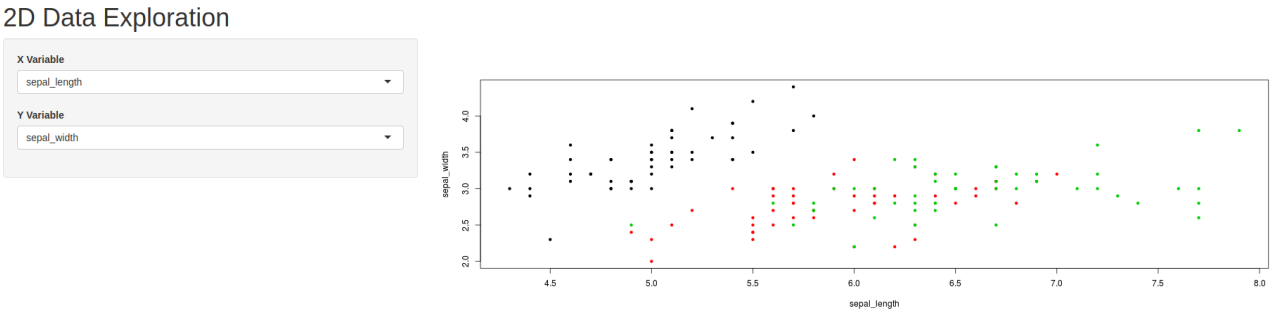
1. Kelchblattlänge in cm
2. Kelchblattbreite in cm
3. Blütenblattlänge in cm
4. Blütenblattbreite in cm
5. Klasse:
-- Iris Setosa
-- Iris Versicolour
-- Iris Virginica

Der Datensatz wurde als csv File aus dem folgenden github Projekt heruntergeladen
https://gist.github.com/curran/a08a1080b88344b0c8a7

Um das Programm auszuführen:
Rscript pcashiny.R

Analyse

In der 2D Darstellung der Daten lässt sich eindeutig erkennen das eine Art Iris linear von den anderen 2 trennbar ist:



Im PCA TAB sind folgende zwei Tabellen abgebildet:

Sample	PC1	PC2	PC3	PC4
sepal_length	0.51	-0.45	0.71	0.21
sepal_width	-0.30	-0.89	-0.32	-0.10
petal_length	0.58	-0.03	-0.20	-0.79
petal_width	0.57	-0.04	-0.59	0.57

Summary	PC1	PC2	PC3	PC4
Standard deviation	1.71	0.96	0.37	0.17
Proportion of Variance	0.73	0.23	0.03	0.01
Cumulative Proportion	0.73	0.96	0.99	1.00

Die obere Tabelle bildet die Rotationsmatrix ab. In dieser Matrix sieht man wie stark jedes Attribut (sepal_length, sepal_widht, …) zu jeder Hauptkomponente (PC!, PC2, …) beinträgt.

In dem 2D Plot der PCA Daten lässt sich eine eindeutige Trennung der zwei nicht linear trennbaren Arten, erkennen: