Wine Clustering Project

Micah Mayanja

2024-06-09

Contents

Principal component analysis & Cluster analysis.

Unsupervised learning! Using Principal component analysis for dimension reduction and then clustering analysis.

The following descriptions are adapted from the UCI webpage: These data are the results of a chemical analysis of wines grown in the same region in Italy but derived from three different cultivars.

```
data1 <- read.csv("~/R/Wine clustering/wine-clustering.csv")
```

```
head(data1)
```

```
##   Alcohol Malic_Acid  Ash Ash_Alcanity Magnesium Total_Phenols Flavanoids
## 1   14.23       1.71 2.43         15.6       127          2.80       3.06
## 2   13.20       1.78 2.14         11.2       100          2.65       2.76
## 3   13.16       2.36 2.67         18.6       101          2.80       3.24
## 4   14.37       1.95 2.50         16.8       113          3.85       3.49
## 5   13.24       2.59 2.87         21.0       118          2.80       2.69
## 6   14.20       1.76 2.45         15.2       112          3.27       3.39
##   Nonflavanoid_Phenols Proanthocyanins Color_Intensity  Hue OD280 Proline
## 1                 0.28            2.29            5.64 1.04  3.92    1065
## 2                 0.26            1.28            4.38 1.05  3.40    1050
## 3                 0.30            2.81            5.68 1.03  3.17    1185
## 4                 0.24            2.18            7.80 0.86  3.45    1480
## 5                 0.39            1.82            4.32 1.04  2.93     735
## 6                 0.34            1.97            6.75 1.05  2.85    1450
```

```
dim(data1)
```

```
## [1] 178  13
```

```
sum(is.na(data1))
```

```
## [1] 0
```

The data contains 178 observations with 13 variables. It should also be noted that the data has no missing values.

Principal Component analysis

```r
wine_data <- scale(data1)
pr.out <- prcomp(data1, scale=TRUE)
summary(pr.out)
```

```
## Importance of components:
##                          PC1    PC2    PC3     PC4     PC5     PC6     PC7
## Standard deviation     2.169 1.5802 1.2025 0.95863 0.92370 0.80103 0.74231
## Proportion of Variance 0.362 0.1921 0.1112 0.07069 0.06563 0.04936 0.04239
## Cumulative Proportion  0.362 0.5541 0.6653 0.73599 0.80162 0.85098 0.89337
##                            PC8     PC9   PC10    PC11    PC12    PC13
## Standard deviation     0.59034 0.53748 0.5009 0.47517 0.41082 0.32152
## Proportion of Variance 0.02681 0.02222 0.0193 0.01737 0.01298 0.00795
## Cumulative Proportion  0.92018 0.94240 0.9617 0.97907 0.99205 1.00000
```

A biplot showing how the variables are represented in a reduced dimensional space defined by the first two principal components.

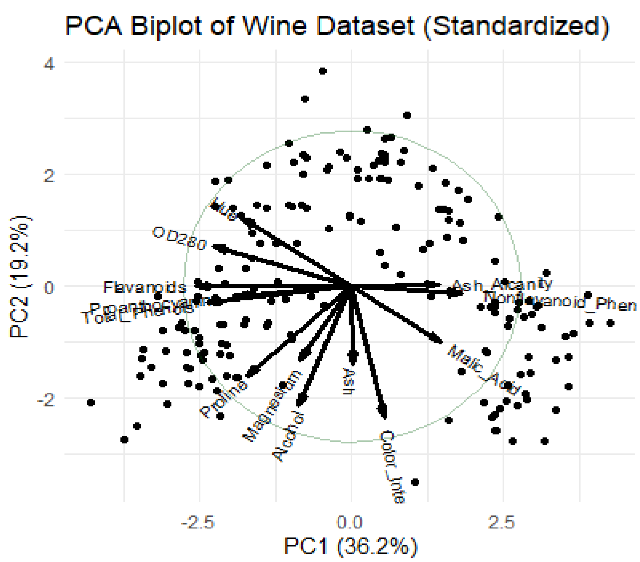
```r
library(ggbiplot)
```

```
## Warning: package 'ggbiplot' was built under R version 4.3.3

## Loading required package: ggplot2

## Warning: package 'ggplot2' was built under R version 4.3.1
```

```r
ggbiplot(pr.out,obs.scale = 1, var.scale = 1,
         circle = TRUE,
         var.axes = TRUE) +
  theme_minimal() +
  ggtitle('PCA Biplot of Wine Dataset (Standardized)')
```

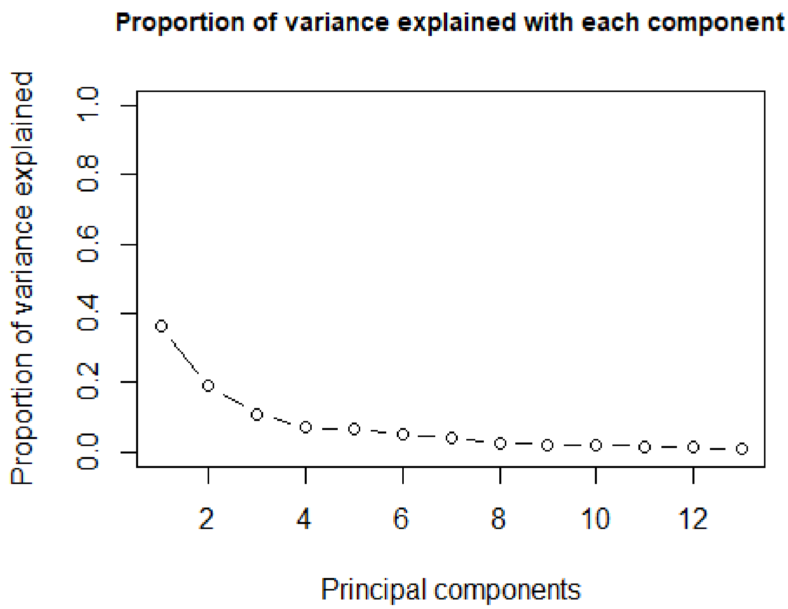


```
#Alternatively
#options(repr.plot.width = 8, repr.plot.height = 6)  #Adjust plot size
#biplot(pr.out,scale = 0, col = c("blue", "red"),cex=0.8)
```

Determine the proportion of variance explained.

```
pr.var <- pr.out$sdev^2 #find variance from standard deviation
pve <- pr.var/sum(pr.var) #compute the proportion of variance explained

#scree plots
plot(pve, xlab="Principal components",ylab="Proportion of variance
explained",
     ylim=c(0,1),main = "Proportion of variance explained with each
component", cex.main = 0.9, type="b")
```



We choose the smallest number of principal components that are required in order to explain a sizable variation of the data (elbow in the scree plot). Therefore, we can reduce the dimensions from 13 variables to 3 principal components.

```
#Extract the first 3 Principal components
Newdata <- as.data.frame(pr.out$x[,1:3])
head(Newdata)

##         PC1        PC2        PC3
## 1 -3.307421 -1.4394023 -0.1652728
## 2 -2.203250  0.3324551 -2.0207571
## 3 -2.509661 -1.0282507  0.9800541
## 4 -3.746497 -2.7486184 -0.1756962
## 5 -1.006070 -0.8673840  2.0209873
## 6 -3.041674 -2.1164309 -0.6276254
```

```
dim(Newdata)

## [1] 178   3
```

Cluster Analysis

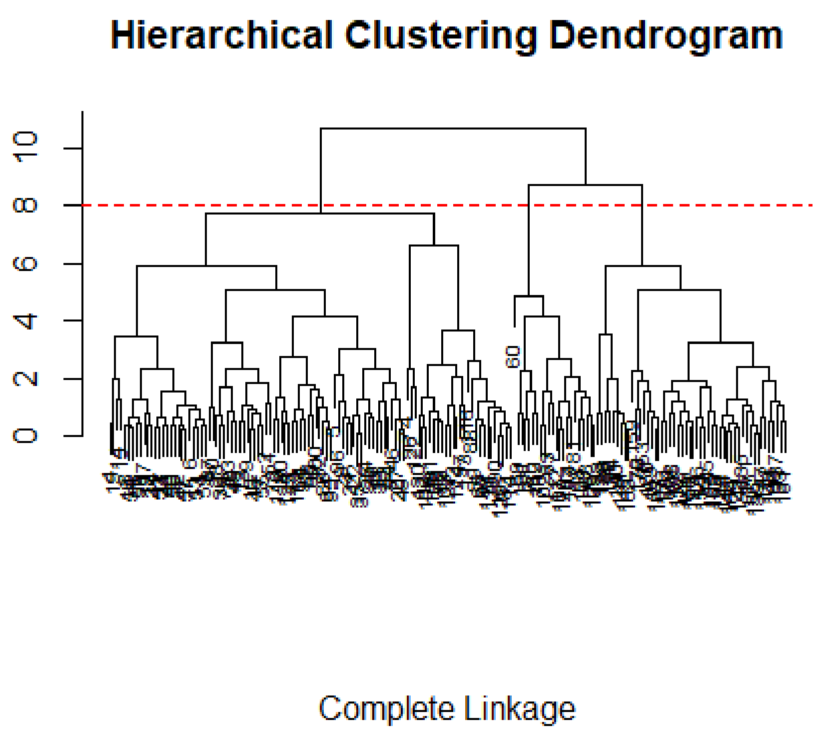
Hierarchical clustering

```
data.dist=dist(Newdata)

hcluster <- hclust(data.dist)

#Plot dendrogram
plot(hcluster,main="Hierarchical Clustering Dendrogram", sub = "Complete
Linkage",xlab ="",ylab ="", cex = 0.6)

#add horizontal line for clustering
abline(h=8,col="red", lty = 2)
```



K-means clustering

```
set.seed(2)
km.out <- kmeans(Newdata,3,nstart =20)
#km.out
km.clusters <-  km.out$cluster

clusters <- as.factor(km.clusters)
```

```r
# Create a PCA biplot with cluster colors
library(ggplot2)
ggplot(Newdata, aes(x = PC1, y = PC2, color = clusters)) +
  geom_point(size = 3) +
  scale_color_manual(values = c("blue", "green", "red")) +  # Customize
cluster colors
  labs(x = "PC1", y = "PC2", color = "Cluster") +
  ggtitle("Clustered Principal components")
```