Unsupervised ML

Tammy Rex

R Markdown

This is a practice exercise on unsupervised methods such as dimentionality reduction and clustering. Principal component analysis (PCA) will be used on the iris dataset and k-means clustering.

Background on data set: The Iris dataset is a famous dataset in the field of statistics and machine learning. It was introduced by the British statistician and biologist Ronald Fisher in his 1936 paper "The use of multiple measurements in taxonomic problems" as an example of discriminant analysis. The dataset consists of measurements of various features of iris flowers belonging to three species: Setosa, Versicolor, and Virginica. The features measured include the lengths and widths of the sepals and petals. The dataset contains 150 observations, with 50 samples from each of the three specie

```
# Load the iris dataset
data(iris)

# Display the structure of the dataset
str(iris)
```

```
## 'data.frame':    150 obs. of  5 variables:
##  $ Sepal.Length: num  5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9 ...
##  $ Sepal.Width : num  3.5 3 3.2 3.1 3.6 3.9 3.4 3.4 2.9 3.1 ...
##  $ Petal.Length: num  1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1.5 ...
##  $ Petal.Width : num  0.2 0.2 0.2 0.2 0.2 0.4 0.3 0.2 0.2 0.1 ...
##  $ Species     : Factor w/ 3 levels "setosa","versicolor",..: 1 1 1 1 1 1
1 1 1 1 ...
```

```
# Display the first few rows of the dataset
head(iris)
```

```
##   Sepal.Length Sepal.Width Petal.Length Petal.Width Species
## 1          5.1         3.5          1.4         0.2  setosa
## 2          4.9         3.0          1.4         0.2  setosa
## 3          4.7         3.2          1.3         0.2  setosa
## 4          4.6         3.1          1.5         0.2  setosa
## 5          5.0         3.6          1.4         0.2  setosa
## 6          5.4         3.9          1.7         0.4  setosa
```

```
# Summary statistics of the dataset
summary(iris)
```

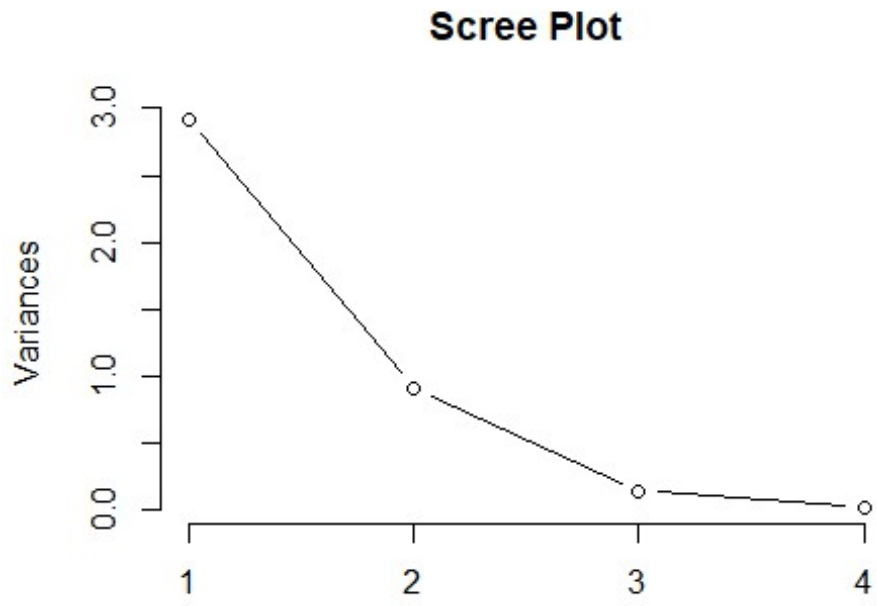
```
##   Sepal.Length    Sepal.Width     Petal.Length    Petal.Width
##  Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100
```

```
##  1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300
##  Median :5.800   Median :3.000   Median :4.350   Median :1.300
##  Mean   :5.843   Mean   :3.057   Mean   :3.758   Mean   :1.199
##  3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.800
##  Max.   :7.900   Max.   :4.400   Max.   :6.900   Max.   :2.500
##        Species
##  setosa    :50
##  versicolor:50
##  virginica :50
##
##
##
```

```r
# Extracting features (predictors)
iris_features <- iris[, 1:4]

# Standardize the features
scaled_features <- scale(iris_features)

# Perform PCApp
pca_result <- prcomp(scaled_features, scale. = TRUE)

# Summary of PCA
summary(pca_result)
```

```
## Importance of components:
##                           PC1    PC2     PC3     PC4
## Standard deviation     1.7084 0.9560 0.38309 0.14393
## Proportion of Variance 0.7296 0.2285 0.03669 0.00518
## Cumulative Proportion  0.7296 0.9581 0.99482 1.00000
```

```r
# Screen plot to visualize the variance explained by each principal component
plot(pca_result, type = "l", main = "Scree Plot")
```

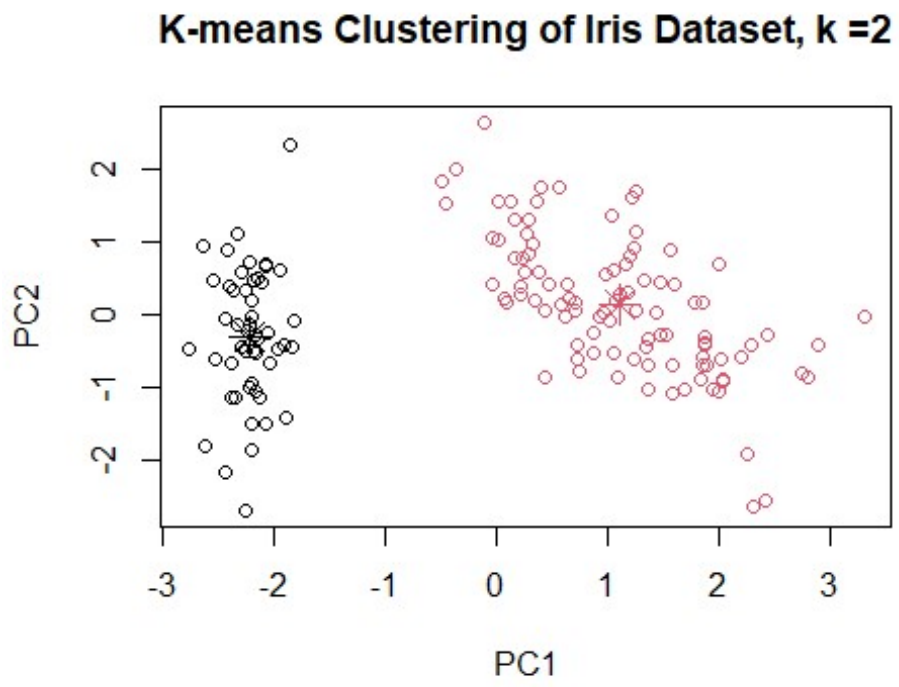


Interpretation of PCA results the plot shows the amount of variance (y-axis) captured by each principal component.

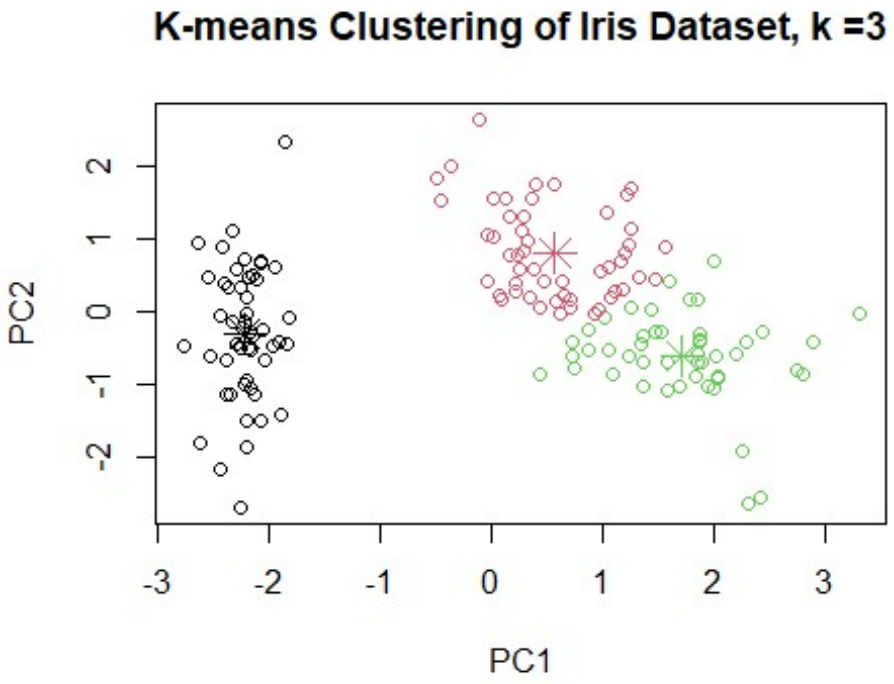
PC1 captures approximately 72.96% of the total variance. PC2 captures approximately 22.85% of the total variance. PC3 captures approximately 3.67% of the total variance. PC4 captures approximately 0.52% of the total variance.

The cumulative proportion indicates that the first two principal components (PC1 and PC2) together capture approximately 95.81% of the total variance, while the first three principal components capture approximately 99.48%. This suggests that the first two principal components are sufficient to capture the majority of the variability in the data.

```
# Starting with 2 clusters
# Extract the scores of the principal components
pc_scores <- as.data.frame(pca_result$x)

# Perform clustering (e.g., k-means clustering)
library(cluster)
set.seed(123)  # for reproducibility
k <- 2 # number of clusters
kmeans_result <- kmeans(pc_scores[, 1:2], centers = k)

# Visualize the clustering results
plot(pc_scores[, 1:2], col = kmeans_result$cluster, main = "K-means
Clustering of Iris Dataset, k =2")
points(kmeans_result$centers, col = 1:k, pch = 8, cex = 2)
```




```r
# Changint to 2 clusters
# Extract the scores of the principal components
pc_scores <- as.data.frame(pca_result$x)

# Perform clustering (e.g., k-means clustering)
library(cluster)
set.seed(123)  # for reproducibility
k <- 3 # number of clusters
kmeans_result <- kmeans(pc_scores[, 1:2], centers = k)

# Visualize the clustering results
plot(pc_scores[, 1:2], col = kmeans_result$cluster, main = "K-means
Clustering of Iris Dataset, k =3")
points(kmeans_result$centers, col = 1:k, pch = 8, cex = 2)
```


K-means Clustering of Iris Dataset, k =3
PC2
PC1
-3
-2
-1
0
1
2
3
-2
-1
0
1
2