Principle Component Analysis and Clustering

Khanin Sisaengsuwanchai

2022-08-18

```
# 1.
# Read the data as follows (x is the data and y is the ID of the digit)
library(dslabs)
library(cluster)
library(factoextra)
```

```
## Loading required package: ggplot2
```

```
## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa
```

```
df0 = read_mnist()
x = data.frame(df0$test$images)
y = data.frame(df0$test$labels)
names(y) = "ID"
df = data.frame(y,x)
dim(df)
```

```
## [1] 10000   785
```

```
# b) How many rows of each digit are in df?
table(df$ID)
```

```
##
##    0    1    2    3    4    5    6    7    8    9
##  980 1135 1032 1010  982  892  958 1028  974 1009
```

```
# The first row: digits
# The second row: number of digits

# 0 digit number = 980
# 1 digit number = 1135
# 2 digit number = 1032
# 3 digit number = 1010
# 4 digit number = 982
# 5 digit number = 892
# 6 digit number = 958
# 7 digit number = 1028
# 8 digit number = 974
# 9 digit number = 1009
```

```r
# c) Create new dataframes df34 and x with digits ID values 3 and 4 only
df34 = df[df$ID==3|df$ID==4,]

x1   = df34[-1] # Remove ID
dim(x1)
```

```
## [1] 1992  784
```

```r
# K-mean clustering
set.seed(1)
k34 = kmeans(x1,centers = 2, nstart = 20, iter.max=20)

head(k34$cluster) # The first row is index, and the second row is its cluster
```

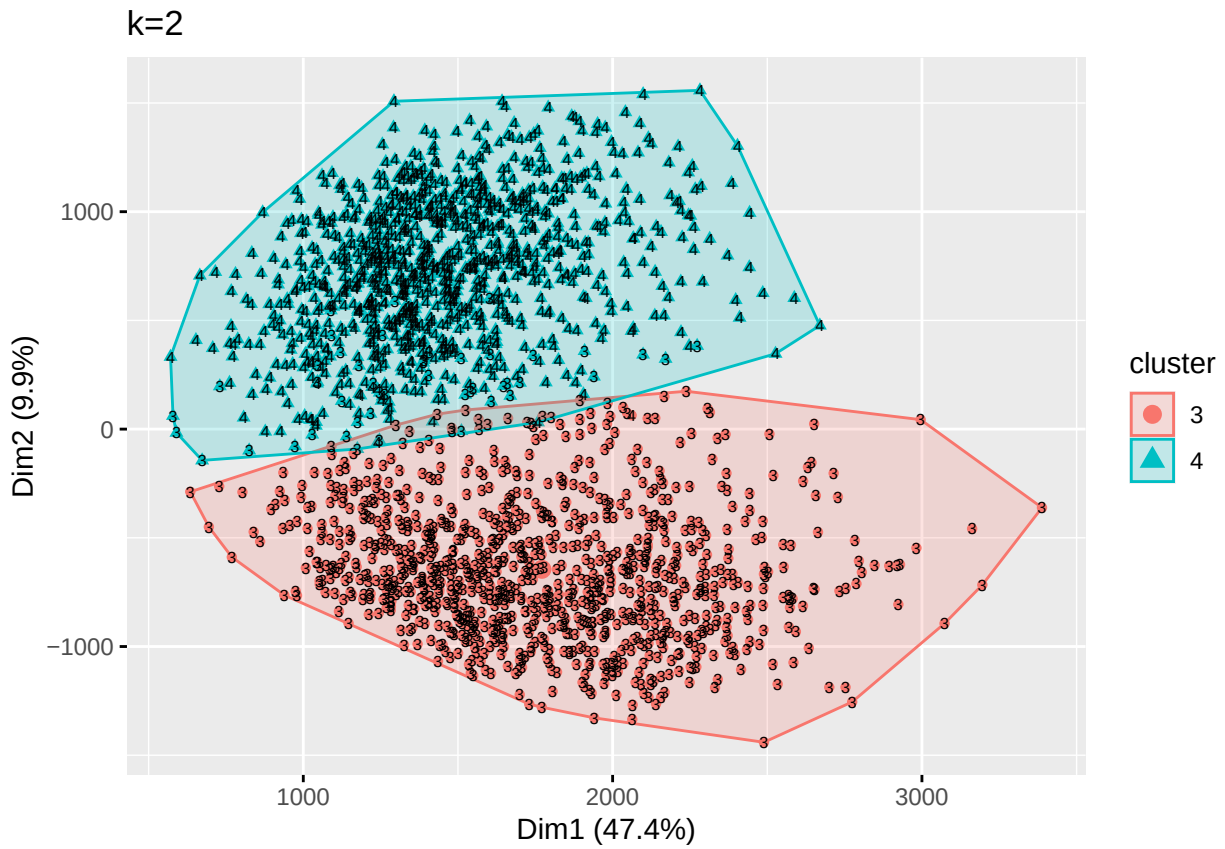
```
##  5  7 19 20 25 28
##  2  2  1  2  2  2
```

```r
df34[1:6,1:6]
```

```
##    ID X1 X2 X3 X4 X5
## 5   4  0  0  0  0  0
## 7   4  0  0  0  0  0
## 19  3  0  0  0  0  0
## 20  4  0  0  0  0  0
## 25  4  0  0  0  0  0
## 28  4  0  0  0  0  0
```

```r
# Since df34 and k34 do not have the same cluster number, I need to align
# their class names.
k34$cluster[k34$cluster==1]=3
k34$cluster[k34$cluster==2]=4

# d) Show the clusters in PC1, PC2 space. Label each point with the actual
# digit number (different color for each different digit).
fviz_cluster(k34, data=x1, stand=FALSE, geom = 'point') + ggtitle("k=2") +
  geom_text(aes(label = df34$ID), size = 2)
```




```
# e) Construct a cross-tab table showing how many digits are correctly
# grouped in each cluster and how many are not.

table(ID=df34$ID,Cluster=k34$cluster)
```

```
##    Cluster
## ID    3   4
##   3 964  46
##   4   1 981
```

```
# Accuracy (hit rate) for the unsupervised learning model
mean(df34$ID == k34$cluster)
```

```
## [1] 0.9764056
```

```
# As expected, the accuracy is very high.
```