R Notebook

Course: HUDK 4050, Week 8

Author: Guotai Sun

Assignment: ICE6

Objectives: At the end of this ICE, I will be able to:

1.develop intuitions about principal component analysis

2.implement the PCA algorithm for dimension reduction purposes

Here are the variables:

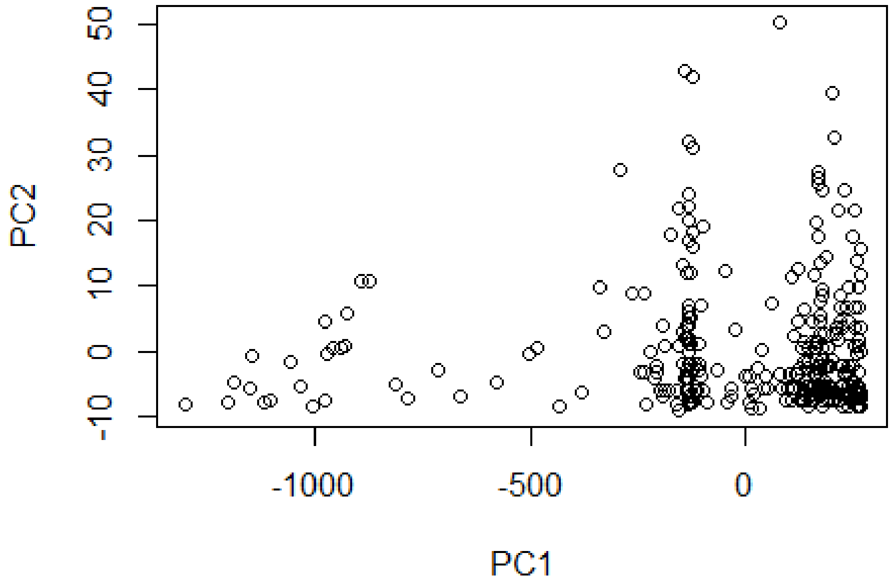
id - student id prior_prob_count - The number of problems a student has done in the system prior to the current session prior_percent_correct - The percentage of problems a student has answered correctly piror to the current session problems_attempted - The number of problems a student has attempted in this current session mean_correct - The percentage of correct problems in this currect session mean_hint - The average number of hints the student requested in the current session mean_attempt - The average attempts for each problem mean_confidence - The reported confidence a student has reported at the end of the session

This is an R Markdown Notebook. When you execute code within the notebook, the results appear beneath the code.

Try executing this chunk by clicking the *Run* button within the chunk or by placing your cursor inside it and pressing *Ctrl+Shift+Enter*.

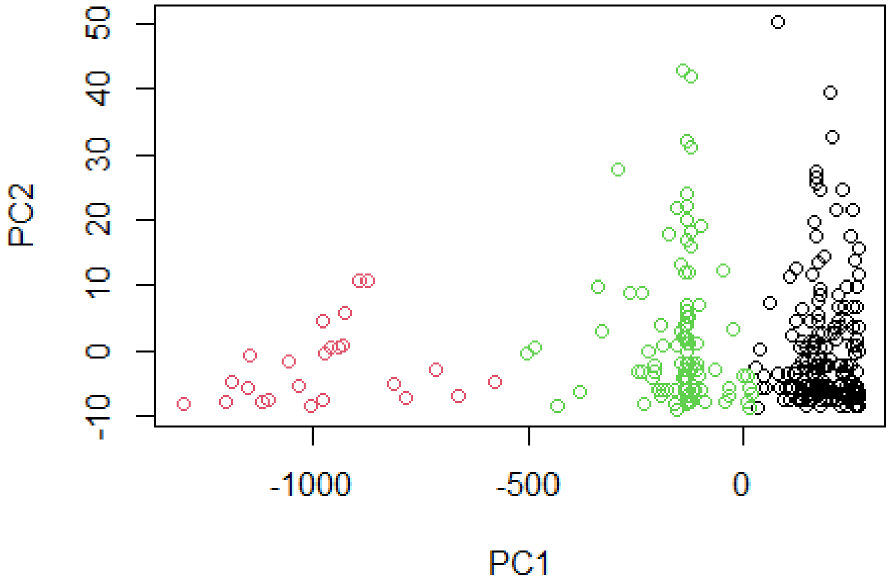
```
library(tidyverse)

## Warning: package 'tidyverse' was built under R version 4.0.5

## -- Attaching packages ---------------------------------------
tidyverse 1.3.1 --

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.4     v dplyr   1.0.7
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.1

## Warning: package 'ggplot2' was built under R version 4.0.5

## Warning: package 'tibble' was built under R version 4.0.5

## Warning: package 'tidyr' was built under R version 4.0.5
## Warning: package 'dplyr' was built under R version 4.0.5

## -- Conflicts ------------------------------------------
tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()

ICEdata <- read_csv("ICE6_Data.csv")

##
## -- Column specification --------------------------------------------
------------
## cols(
##   id = col_double(),
##   prior_prob_count = col_double(),
##   prior_percent_correct = col_double(),
##   problems_attempted = col_double(),
##   mean_correct = col_double(),
##   mean_hint = col_double(),
##   mean_attempt = col_double(),
##   mean_confidence = col_double()
## )

ICEdata

## # A tibble: 342 x 8
##         id prior_prob_count prior_percent_correct problems_attempted
mean_correct
##      <dbl>            <dbl>                 <dbl>              <dbl>
<dbl>
##  1 172777              650                 0.723                  4
1
##  2 175658             1159                 0.801                 22
0.455
##  3 175669             1239                 0.657                 11
0.636
##  4 176151             1246                 0.730                 16
0.75
##  5 176165             1299                 0.568                  6
0.333
##  6 176168             1415                 0.685                 11
0.545
##  7 176461              753                 0.499                 11
0.364
##  8 176486              772                 0.576                 10
0.3
##  9 176488              529                 0.675                 19
0.421
## 10 176494             1226                 0.644                 12
0.25
```

```
## # ... with 332 more rows, and 3 more variables: mean_hint <dbl>,
## #   mean_attempt <dbl>, mean_confidence <dbl>

ICEdata_noid <- ICEdata %>% select(-id)
icepca <- prcomp(ICEdata_noid, scale. = FALSE) # Here let's see a
unscaled example
summary(icepca)

## Importance of components:
##                             PC1     PC2     PC3    PC4    PC5    PC6
PC7
## Standard deviation     319.233 9.82811 0.89846 0.6489 0.2033 0.1269
0.1044
## Proportion of Variance   0.999 0.00095 0.00001 0.0000 0.0000 0.0000
0.0000
## Cumulative Proportion    0.999 0.99999 1.00000 1.0000 1.0000 1.0000
1.0000

#It shows variance of the seven variables,and the two most siginificant
ones are PC1 and PC2, which stand for 99.99% explanation of the model

icepca2c <- icepca$x[,1:2] #choose the first and the second variables
plot(icepca2c)
```



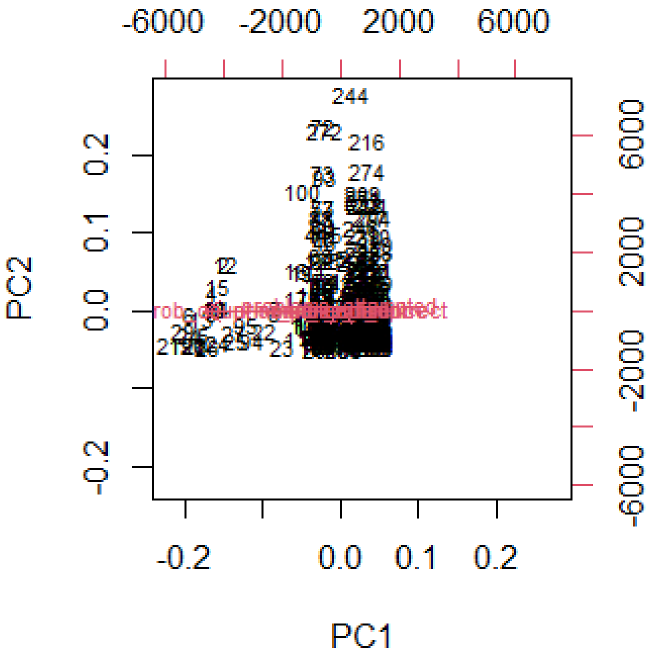
```
#since there are three clusters in the graph of icepca2s, we use 3-mean
clusters to group data
```

```
cl <- kmeans(icepca2c, centers = 3)
plot(icepca2c, col = cl$cluster)
```



```
#It looks like the clustering works pretty nicely. So now, it is your
turn. Think about two things: what could the two dimensions mean? How
would you interpret the KMeans clustering?

biplot(icepca, cex=.7)

## Warning in arrows(0, 0, y[, 1L] * 0.8, y[, 2L] * 0.8, col = col[2L],
length =
## arrow.len): zero-length arrow is of indeterminate angle and so
skipped

## Warning in arrows(0, 0, y[, 1L] * 0.8, y[, 2L] * 0.8, col = col[2L],
length =
## arrow.len): zero-length arrow is of indeterminate angle and so
skipped

## Warning in arrows(0, 0, y[, 1L] * 0.8, y[, 2L] * 0.8, col = col[2L],
length =
## arrow.len): zero-length arrow is of indeterminate angle and so
skipped

## Warning in arrows(0, 0, y[, 1L] * 0.8, y[, 2L] * 0.8, col = col[2L],
length =
## arrow.len): zero-length arrow is of indeterminate angle and so
```

```
skipped

## Warning in arrows(0, 0, y[, 1L] * 0.8, y[, 2L] * 0.8, col = col[2L],
length =
## arrow.len): zero-length arrow is of indeterminate angle and so
skipped
```



Add a new chunk by clicking the *Insert Chunk* button on the toolbar or by pressing *Ctrl+Alt+I*.

When you save the notebook, an HTML file containing the code and output will be saved alongside it (click the *Preview* button or press *Ctrl+Shift+K* to preview the HTML file).

The preview shows you a rendered HTML copy of the contents of the editor. Consequently, unlike *Knit*, *Preview* does not run any R code chunks. Instead, the output of the chunk when it was last run in the editor is displayed.