|  |  |
| --- | --- |
|  | |
| Data Advanced Data Analytics  CA TWO | |
| Module code : B8IT109 | |
| Ciaran Finnegan  Student No : 10524150  24/05/2020 |  |
|  |  |

Table of Contents

1 Question One 3

1.1 Question 1 – from PDF 3

1.2 Output From RStudio Cloud Console 3

2 Question Two 4

2.1 Question 2 – from PDF 4

2.2 Output from RStudio Cloud Console 4

3 Question Three 5

3.1 Question 2 – from PDF 5

3.2 Output from RStudio Cloud Console 5

4 Question Four 6

4.1 Question 2 – from PDF 6

4.2 Output from RStudio Cloud Console 7

# Question One

## Question 1 – from PDF

Use in-built dataset ‘airquality’,

a) explore the general feature of dataset using appropriate R functions.

(**5 Marks**)

b) perform data cleansing if required. (**5 Marks**)

c) consider ‘Temp’ attributes and compute the central and variational measures. (**10 Marks**)

d) apply boxplot technique to detect outlier of ‘wind’ attribute if any.

(**10 Marks**)

**(Total: 30 Marks)**

## Output From RStudio Cloud Console

To follow...

# Question Two

## Question 2 – from PDF

Use dataset available on http://users.stat.ufl.edu/~winner/data/nfl2008\_fga.csv , then:

(a) Train the model using 80% of this dataset and suggest an appropriate GLM to model **homekick** to **togo, ydline** and **kicker** variables.

**(5 Marks)**

(b) Specify the significant variables on **homekick** at the level of 𝛼=0.05, and estimate the parameters of your model.

**(5 Marks)**

(c) Predict the test dataset using the trained model. **(5 Marks)**

(d) Provide the confusion matrix and obtain the probability of correctness of predictions. **(10 Marks)**

(Total: 25 Marks)

## Output from RStudio Cloud Console

To follow..

# Question Three

## Question 2 – from PDF

Using Yahoo Finance API, select a specific stock market price, apply time series analysis, consider ‘***close price*** as your time series variable:

1. Validate the assumptions using graphical visualization. (5 Marks)
2. Fit the optimized model for ‘***close price’*** and provide the coefficient estimates for the fitted model. **(5 Marks)**
3. What is the estimated order for AR and MA? (5 Marks)

(d) Forecast h=10 step ahead prediction of ***wage*** on the plot of the original time series. **(10 Marks)**

**(Total: 25 Marks)**

## Output from RStudio Cloud Console

To follow..

# Question Four

## Question 2 – from PDF

Use dataset available on http://users.stat.ufl.edu/~winner/data/nfl2008\_fga.csv

1. Use LDA to classify the dataset into few classes so that at least 90% of information of dataset is explained through new classification. (**Hint**: model the variable “**qtr”** to variables “**togo**”, “**kicker**”, and “**ydline**”). How many LDs do you choose? Explain the reason.

**(5 Marks)**

*2.* Apply PCA, and identify the important principle components involving at least 90% of dataset variation. Explain your decision strategy? Plot principle components

versus their variance (**Hint**: to sketch the plot use the Scree plot). **(5 Marks)**

3. Split the dataset into two sets of variables so that **X**=( togo,kicker,ydline) and **Y**=( distance, homekick). Apply canonical correlation analysis to find the cross-correlation between **X** and **Y**. What is the correlation between ***ydline*** and ***distance***? **(5 Marks)**

4. Use K-means clustering analysis to identify the most important classes. How many classes do you select? Why?

**(6 Marks)**

**(Total: 20 Marks)**

## Output from RStudio Cloud Console

> ## CA Two Advanced Data Analytics : Module Code B8IT109

> ## Student Name : Ciaran Finnegan

>

> ## Student Number : 10524150

>

> ## May 2020

> >

> ## Question Four

>

> ## Q. 4(Part 1)

>

> ## Use LDA to classify the dataset into few classes so that at least 90% of information

> ## of dataset is explained through new classification.

> ## (Hint: model the variable “qtr” to variables “togo”, “kicker”, and “ydline”).

> ## How many LDs do you choose? Explain the reason.

> >

>

> ## Load MASS library to use LDA function

> library(MASS)

> library(CCA)

> >

> ## Read in the NFL dataset

> link='http://users.stat.ufl.edu/~winner/data/nfl2008\_fga.csv'

> datasetNFL=read.csv(link)

> head(datasetNFL)

GameDate AwayTeam HomeTeam qtr min sec kickteam def down togo kicker ydline name distance homekick kickdiff timerem offscore defscore season GOOD Missed

1 20081130 IND CLE 1 47 2 IND CLE 4 11 15 12 A.Vinatieri 30 0 -3 2822 0 3 2008 1 0

2 20081005 IND HOU 1 54 47 IND HOU 4 3 15 28 A.Vinatieri 46 0 0 3287 0 0 2008 1 0

3 20081228 TEN IND 1 45 20 IND TEN 4 3 15 10 A.Vinatieri 28 1 7 2720 7 0 2008 1 0

4 20081012 BAL IND 1 45 42 IND BAL 4 1 15 19 A.Vinatieri 37 1 14 2742 14 0 2008 1 0

5 20080907 CHI IND 1 50 56 IND CHI 4 21 15 21 A.Vinatieri 39 1 0 3056 0 0 2008 1 0

6 20081116 HOU IND 1 50 43 IND HOU 4 7 15 22 A.Vinatieri 40 1 -3 3043 0 3 2008 1 0

Blocked

1 0

2 0

3 0

4 0

5 0

6 0

>

> ## Minor Clean up of NFL dataset

> sum(is.na(datasetNFL))

[1] 4

> datasetNFL <- na.omit(datasetNFL)

> sum(is.na(datasetNFL))

[1] 0

>

> # Display the values for 'qtr'

> table(datasetNFL$qtr)

1 2 3 4 5

214 373 182 255 13

>

>

>

>

> ## Use LDA function to classify dataset. The output variable is 'qtr' and the input variables are

> ##'togo', 'kicker', and 'ydline'.

> datasetNFL.lda <- lda(qtr~togo+kicker+ydline, data=datasetNFL)

> datasetNFL.lda

Call:

lda(qtr ~ togo + kicker + ydline, data = datasetNFL)

Prior probabilities of groups:

1 2 3 4 5

0.20636451 0.35969142 0.17550627 0.24590164 0.01253616

Group means:

togo kicker ydline

1 6.481308 19.64486 17.22897

2 6.973190 18.77212 19.30027

3 6.543956 19.96703 19.03297

4 6.792157 20.20000 18.53725

5 5.923077 22.61538 19.53846

Coefficients of linear discriminants:

LD1 LD2 LD3

togo 0.06665269 0.12498308 0.20996464

kicker -0.04134867 -0.06009657 0.05013225

ydline 0.07726467 -0.07173243 -0.02257770

Proportion of trace:

LD1 LD2 LD3

0.615 0.322 0.063

>

> ## Two LDs are required - LD1 and LD2 - to explain at least 85% of formation of the NFL dataset is explained

> ## LD1 explains 61.5%. LD2 explains a further 32%. Hence LD1 and LD2 will explain 93.7 % together.

>

>

>

## Q. 4(Part 2)

>

> ## Apply PCA, and identify the important principle components involving at least 90% of dataset variation.

> ## Explain your decision strategy?

>

> ## We only use the input variables for the PCA question.This analysis is a type of 'unsepervised' learning.

> datasetNFL2 = cbind(datasetNFL$togo, datasetNFL$kicker, datasetNFL$ydline)

> fit <- princomp(datasetNFL2, cor = TRUE)

> summary(fit)

Importance of components:

Comp.1 Comp.2 Comp.3

Standard deviation 1.146726 0.9998433 0.8278479

Proportion of Variance 0.438327 0.3332289 0.2284441

Cumulative Proportion 0.438327 0.7715559 1.0000000

>

> ## Looking at the 'Cumulative Proportion' output line we can see that Comp1 captures 43.8% of dataset variation.

> ## Comp 1 and Comp2 togther capture 77.2% (approx) of dataset variation.

> ## However, ll three components (Comp1, Comp2, Comp3) are important to capture 90% of the dataset variation.

>

>

>

> ## Plot principle components versus their variance

> ## (Hint: to sketch the plot use the Scree plot).

> loadings(fit)

Loadings:

Comp.1 Comp.2 Comp.3

[1,] 0.707 0.707

[2,] 0.999

[3,] 0.706 -0.707

Comp.1 Comp.2 Comp.3

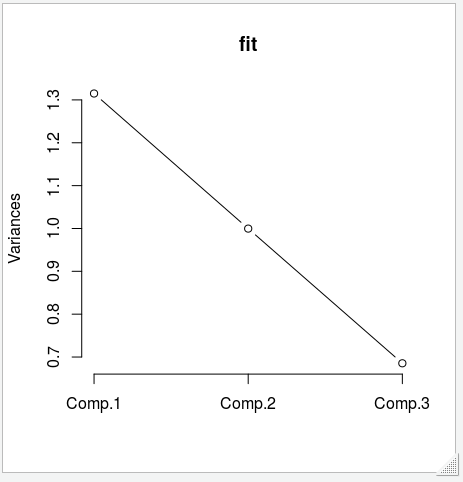
SS loadings 1.000 1.000 1.000

Proportion Var 0.333 0.333 0.333

Cumulative Var 0.333 0.667 1.000

> plot(fit, type = "lines")

>



> # The Plot confirms that all three components are important. There is no 'bend' in the line indicating that higher

> # components contribute less to the capture of dataset variation

>

>

>

>

## Q. 4(Part 3)

>

> ## Split the dataset into two sets of variables so that X=(togo, kicker, ydline) and Y=(distance, homekick).

> ## Apply canonical correlation analysis to find the cross-correlation between X and Y.

>

>

> ## Set up 'X' variable

> X <- cbind(datasetNFL$togo, datasetNFL$kicker, datasetNFL$ydline)

>

> ## Set up 'Y' variable

> Y <- cbind(datasetNFL$distance, datasetNFL$homekick)

> cor(X, Y)

[,1] [,2]

[1,] 0.315641454 -0.04838438

[2,] -0.001951722 -0.02363159

[3,] 0.998947222 0.04295427

>

>

> ## What is the correlation between 'ydline' and 'distance'?

>

> ## Read three down the X value and one across the Y value

> ## The correlation between 'ydline' and 'distance' is equal to '0.998947222'

> ## This value shows a high level of correlation between the 'ydline' and 'distance' values

>

>

>

>

>

>

>

>

## Q. 4(Part 4)

>

> ## Use K-means clustering analysis to identify the most important classes.

> ## How many classes do you select? Why?

>

> ## Again consider the input variables. We use the 'datasetNFL2' dataset because I want to

> ## just consider the 'togo', 'kicker', and 'ydline' input variables.

>

> ## K-Mean

> ##k.means.fit <- kmeans(datasetNFL2, 4)

> ##attributes(k.means.fit)

>

> # Centroids(arithmetic mean)

> ##k.means.fit$centers

>

> # Cluster size

> ##k.means.fit$size

>

>

> # Generate the plot K-Means clustering

> ## Write function for plot generation

> wssplot <- function(datasetNFL2, nc=10, seed=2343){

+

+ wss <- (nrow(datasetNFL2)-1) \* sum(apply(datasetNFL2, 2, var))

+

+ for (i in 2 : nc){

+

+ set.seed(seed)

+ wss[i] <- sum(kmeans(datasetNFL2, centers = i)$withinss)

+

+ }

+

+ plot(1:nc, wss, type = "b", xlab = "Numbers of Clusters", ylab = "Within Groups Sum of Squares")

+

+

+ }

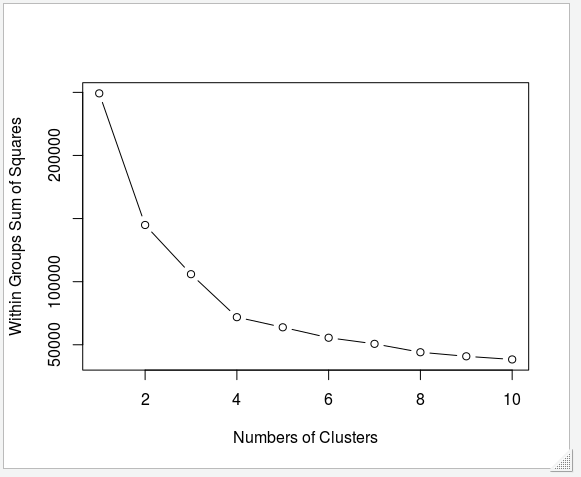
>

>

# Invoke plot function

> wssplot(datasetNFL2, nc = 10)

>



> ## In the Cluster graph we can see a definite 'elbox' at Number of Clusters = 4.

> ## After Cluster 4 the changes in variation are noticeably less

> ## Therefore the main cluster are clusters 1 through to cluster 4.

> ## We would select four classes as an answer to this question.