|  |  |
| --- | --- |
|  | |
| Data Advanced Data Analytics  Advanced Data Aanlyics Exam | |
| Module code : B8IT109 | |
| Ciaran Finnegan  Student No : 10524150  14/06/2020 |  |
|  |  |

Table of Contents

1 ADA – Exam Paper Submission 3

1.1 Course Details 3

1.2 Exam Declaration 3

2 Question One 4

2.1 Question 1 – from PDF 4

2.2 Output From RStudio Cloud Console 5

3 Question Two 6

3.1 Question 2 – from PDF 6

3.2 Output from RStudio Cloud Console 7

4 Question Three 8

4.1 Question 3 – from PDF 8

4.2 Output from RStudio Cloud Console 9

5 Question Four 10

5.1 Question 4 – from PDF 10

5.2 Output from RStudio Cloud Console 11

# ADA – Exam Paper Submission

## Course Details

|  |  |
| --- | --- |
| Module Code | B8IT109 (B8IT109\_1920\_SME3\_BHD08DNW) |
| Module Name | Advanced Data Analytics – Fri/Sat Part-time March 2019 Intake |
| Date | 15th June 2020 |
| Student number | 10524150 |
| Student name | Ciaran Finnegan |

## Exam Declaration

*By uploading this exam from my Moodle account I Ciaran Finnegan am confirming that this document is all my own work.*

*I understand that DBS will carry out checks such as text-matching (via Urkund), bench-marking and viva voce exams in order to verify the authenticity of submissions.*

# Question One

## Question 1 – from PDF

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

(a) Use LDA to classify the dataset into few classes so that at least 85% of information of dataset is explained through new classification. (**Hint**: model the output variable “**carclass\_id”** to input variables “**msrp**”, “**accelrate**”, and “**mpg**”). How many LDs do you choose? Explain the reason. **(10 Marks)**

(b) Apply PCA to input variables, 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)**

(c) Use K-means clustering analysis to input variables and identify the most important classes. How many classes do you select? Why?

(5 Marks)

(d) Split the dataset into two sets of variables so that **X**=( msrp, mpgmpge) and **Y**=( accelrate, mpg). Apply canonical correlation analysis to find the cross-correlation between **X** and **Y**. What is the correlation between ***msrp*** and ***mpg***?

**(5 Marks)**

**(Total: 25 Marks)**

## Output From RStudio Cloud Console

## CA Two Advanced Data Analytics : Module Code B8IT109

> ## Advanced Data Analytics : Module Code B8IT109

> ## Student Name : Ciaran Finnegan 10524150

>

> ## Exam Submission

>

> ## June 15th 2020

>

> ## Question N : LDA, PCA, K-Means, Canonical Correlation

> ## Multivariate analysis and unsupervised learning methods

>

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

> ## Perform initial data load, analysis, and clean up operations before starting

> ## Question 4 solution

>

>

>

> ## Load MASS library to use LDA function

> library(MASS)

> ## Load CCA library to use functions for Canonical Correlation

> library(CCA)

> ## Load 'factoextra' for visualization - Scree plot

> library(factoextra)

>

>

>

>

> ## Inbuilt dataset

> ## dataBsn = Boston

>

> ## dataframe

> ## data('EuStockMarkets')

> ## Dax.df = data.frame(EuStockMarkets)

> ## t = Dax.df$DAX

>

>

>

> ## Read in the NFL dataset

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

> datasetNFL=read.csv(link)

> ## Brief Review of number of rows, head and tail of dataset records

> ## the and structure of dataset

> nrow(datasetNFL)

[1] 1039

> head(datasetNFL)

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

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

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

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

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

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

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

Missed Blocked

1 0 0

2 0 0

3 0 0

4 0 0

5 0 0

6 0 0

> tail(datasetNFL)

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

1034 20081102 TB KC 5 -5 27 TB KC 4 1 36 16 M.Bryant 34 0 0 -273 27 27 2008 1

1035 20081102 GB TEN 5 -6 41 TEN GB 3 2 37 23 R.Bironas 41 1 0 -319 16 16 2008 1

1036 20081211 NO CHI 5 -3 19 CHI NO 2 13 7 18 R.Gould 35 1 0 -161 24 24 2008 1

1037 20081222 GB CHI 5 -4 33 CHI GB 3 10 7 20 R.Gould 38 1 0 -207 17 17 2008 1

1038 20081116 PHI CIN 5 -15 13 CIN PHI 4 5 8 29 S.Graham 47 1 0 -887 13 13 2008 0

1039 20081019 NYJ OAK 5 -13 35 OAK NYJ 4 11 28 39 S.Janikowski 57 1 0 -745 13 13 2008 1

Missed Blocked

1034 0 0

1035 0 0

1036 0 0

1037 0 0

1038 1 0

1039 0 0

> str(datasetNFL)

'data.frame': 1039 obs. of 23 variables:

$ GameDate: int 20081130 20081005 20081228 20081012 20080907 20081116 20081123 20081207 20081130 20090118 ...

$ AwayTeam: chr "IND" "IND" "TEN" "BAL" ...

$ HomeTeam: chr "CLE" "HOU" "IND" "IND" ...

$ qtr : int 1 1 1 1 1 1 1 1 1 1 ...

$ min : int 47 54 45 45 50 50 46 52 46 49 ...

$ sec : int 2 47 20 42 56 43 45 34 12 46 ...

$ kickteam: chr "IND" "IND" "IND" "IND" ...

$ def : chr "CLE" "HOU" "TEN" "BAL" ...

$ down : int 4 4 4 4 4 4 4 4 4 4 ...

$ togo : int 11 3 3 1 21 7 5 7 7 9 ...

$ kicker : int 15 15 15 15 15 15 15 18 18 29 ...

$ ydline : int 12 28 10 19 21 22 5 8 20 27 ...

$ name : chr "A.Vinatieri" "A.Vinatieri" "A.Vinatieri" "A.Vinatieri" ...

$ distance: int 30 46 28 37 39 40 23 26 38 45 ...

$ homekick: int 0 0 1 1 1 1 0 0 0 0 ...

$ kickdiff: int -3 0 7 14 0 -3 0 0 -3 -7 ...

$ timerem : int 2822 3287 2720 2742 3056 3043 2805 3154 2772 2986 ...

$ offscore: int 0 0 7 14 0 0 0 0 0 0 ...

$ defscore: int 3 0 0 0 0 3 0 0 3 7 ...

$ season : int 2008 2008 2008 2008 2008 2008 2008 2008 2008 2008 ...

$ GOOD : int 1 1 1 1 1 1 1 1 1 1 ...

$ Missed : int 0 0 0 0 0 0 0 0 0 0 ...

$ Blocked : int 0 0 0 0 0 0 0 0 0 0 ...

>

>

> ## Minor Clean up of NFL dataset

> sum(is.na(datasetNFL))

[1] 4

> datasetNFL <- na.omit(datasetNFL)

> sum(is.na(datasetNFL))

[1] 0

> nrow(datasetNFL) # Confirm rows after missing data removed = nn

[1] 1037

>

>

>

>

>

> ## Q. 4(Part 1)

>

> ## Use LDA to classify the dataset into a small number of classes

> ## so that at least 90% of the information of the 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.

>

>

>

> # 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 90% of

> ## formation of the NFL dataset

> ## Reading values under the 'Proportion of trace:' output I can see...

> ## LD1 explains 61.5%.

> ## LD2 explains a further 32.2%.

> ## 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 'unsupervised' learning.

> ## Just focusing on raw data on 'togo', 'kicker', and 'ydline' from

> ## dataset and extracting PCs from the correlation matrix

>

> ## I could have used the

> ## cbind(datasetNFL$togo, datasetNFL$kicker, datasetNFL$ydline) function

> ## but I want to preserve the dataset column names

> datasetNFL2 = datasetNFL[10:12] # datasetNFL$togo, datasetNFL$kicker, datasetNFL$ydline

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

> summary(fit) # Print variance

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, all 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).

>

> #

> ## Use function to extract loadings for factor analysis

> ## - small loadings are usually not printed.

> loadings(fit) # PC loadings

Loadings:

Comp.1 Comp.2 Comp.3

togo 0.707 0.707

kicker 0.999

ydline 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

>

> #

> ## Visualize eigenvalues (scree plot). Show the percentage

> ## of variances explained by each principal component.

> fviz\_eig(fit)

>

> # <Insert graph here>

>

> #

> plot(fit, type = "lines") # Another Scree Plot view. A plot of variation

>

> #

> ## <Insert graph here>

> ## Component 2 is just on or over the Variance value of '1' so I can

> ## determine that Component 1 and 2 are the most important components

> ## to consider.

> ## However, in order to meet the 90% level of variance requested in

> ## this question we still need to consider Component 3.

>

>

> biplot(fit) # Graph that shows two components and role of each variable

> ##(relationship between components and variables)

>

> #

> # <Insert graph here>

> ## For example, as you increase 'ydline' there is a

> ## noticeable increase in Component 1

> ## An increase in 'ydline' shows a very minor increase in Component 2

>

>

>

> #

> ## The Plots confirms that all three components are important to capture 90%.

> ## 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)

> X1 <- data.frame(datasetNFL$togo, datasetNFL$kicker, datasetNFL$ydline)

>

> ## Set up 'Y' variable

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

> Y1 <- data.frame(datasetNFL$distance, datasetNFL$homekick)

>

> ## Run 'cor' function to produce Correlation Matrix

> cor(X, Y)

[,1] [,2]

[1,] 0.315641454 -0.04838438

[2,] -0.001951722 -0.02363159

[3,] 0.998947222 0.04295427

> cor(X1, Y1)

datasetNFL.distance datasetNFL.homekick

datasetNFL.togo 0.315641454 -0.04838438

datasetNFL.kicker -0.001951722 -0.02363159

datasetNFL.ydline 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.

>

>

>

> # 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)

> # Use a default of number of classes = 10 to start the analysis

>

> # <Insert Graph here..>

>

>

> #

> ## In the Cluster graph we can see a definite 'elbow' 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.

>

>

> ## K-Means : Clustering Analysis on NFL Dataset

> k.means.fit <- kmeans(datasetNFL2, 4) # k = 4, the number of classes in type (see above)

> attributes(k.means.fit)

$names

[1] "cluster" "centers" "totss" "withinss" "tot.withinss" "betweenss" "size" "iter" "ifault"

$class

[1] "kmeans"

>

> ## Centroids(arithmetic mean)

> k.means.fit$centers

togo kicker ydline

1 5.268293 28.439024 10.22997

2 7.961373 30.841202 27.05579

3 5.952586 8.336207 10.53879

4 7.859649 10.536842 26.82807

>

> ## Cluster size - shows the breakdown of the number of datapoints in the NFL dataset

> ## into my chosen cluster grouping

> k.means.fit$size

[1] 287 233 232 285

>

> ## All value above sum to nnnn, which is the size of the dataset (rows)

|  |
| --- |
|  |
|  |
| |  | | --- | |  | |

# Question Two

## Question 2 – from PDF

## Output from RStudio Cloud Console

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

# Question Three

## Question 3 – from PDF

## Output from RStudio Cloud Console

## CA Two Advanced Data Analytics : Module Code B8IT109

# Question Four

## Question 4 – from PDF

## Output from RStudio Cloud Console

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