MA5810

Logistic Regression and Clustering

Assessment 2

Bievenido Hiyas Jr
13824819

29-Nov-2020

Assessment2

Bienvenido Hiyas Jr

29/11/2020

Question 1

Imagine you are asked to use a logistic regression to classify whether the breast cancer is benign or malignant. The data Breast cancer.csv is a subset data from Breast Can- cer Wisconsin, which is available from the UCI Data Repository http://archive.ics. uci.edu/ml/datasets/breast+cancer+wisconsin+%28diagnostic%29. The variables are summarised as follows: • V1: ID number • V2: Diagnosis (M = malignant, B = benign) • V3: radius (mean of distances from center to points on the perimeter) • V4: texture (standard deviation of gray-scale values) • V5: perimeter • V6: area • V7: smoothness (local variation in radius lengths) • V8: compactness (perimeter2 / area — 1.0) • V9: concavity (severity of concave portions of the contour) • V10: concave points (number of concave portions of the contour) • V11: symmetry • V12: fractal dimension ("coastline approximation" — 1) Assignment tasks:

```
#Importing and data exploration
data = read.table("breast_cancer.csv", header = TRUE, sep = ",")
summary(data)
str(data)

##cleaning up data: transform "0" to NA
data[data==0] <- NA
#delete rows with NA's
data = na.omit(data)
str(data)

#transform V2 from char to factor and the reset from char to numeric
data$V1 = as.numeric(data$V1)
data$V2 = as.factor(data$V2)
data$V3 = as.numeric(data$V3)
data$V4 = as.numeric(data$V4)
data$V5 = as.numeric(data$V5)
data$V6 = as.numeric(data$V6)
data$V7 = as.numeric(data$V7)
data$V8 = as.numeric(data$V8)
data$V9 = as.numeric(data$V9)
data$V10 = as.numeric(data$V10)
```

```
data$V11 = as.numeric(data$V11)
data$V12 = as.numeric(data$V12)
```

(a) Implement the logistic regression to classify the breast cancer type in R. (Note: You need to provide details of all steps relating to the implementation of the logistic regression such as data preparation, training the model, evaluating the perforamnce of the model on both training and test data and discuss the results.)

```
#partition dta set into train and set and randomly split it
set.seed(123)
#split into training (80%) and test
ind <- createDataPartition(data$V2, p = 0.8, list = F)
train <- data[ind, ]
test <- data[-ind, ]
# print number of observations in test vs. train
c(nrow(train), nrow(test))
```

```
## [1] 446 110
```

```
# Proportions of people that defaulted and did not default
table(train$V2) %>% prop.table()
```

```
##
##         B         M
## 0.6188341 0.3811659
```

```
#Train the model to predict the likelihood of Malignant based on all p
redictors using Logistic Regressin in Caret Package
log_Reg <- glm(V2 ~., data = train, family = "binomial")
summary(log_Reg)$coef
```

```
##                  Estimate   Std. Error    z value     Pr(>|z|)
## (Intercept) -1.364534e+01 1.662111e+01 -0.8209644 4.116665e-01
## V1          -2.619782e-09 4.583330e-09 -0.5715891 5.676004e-01
## V3           1.295133e+00 4.761855e+00  0.2719808 7.856368e-01
## V4           4.699485e-01 9.239399e-02  5.0863530 3.650145e-07
## V5          -5.927599e-01 6.386194e-01 -0.9281896 3.533092e-01
## V6           4.359927e-02 2.121923e-02  2.0547058 3.990742e-02
## V7           1.229228e+02 4.331153e+01  2.8381074 4.538191e-03
## V8           4.924540e+00 2.542764e+01  0.1936688 8.464352e-01
## V9           4.805464e+01 1.618486e+01  2.9691097 2.986640e-03
## V10          1.117339e+01 3.642063e+01  0.3067874 7.590052e-01
## V11          1.488913e+01 1.463627e+01  1.0172763 3.090220e-01
## V12         -1.098807e+02 1.081331e+02 -1.0161613 3.095526e-01
```

Intercept shows log odds that the breast cancer is Benign or Malignant when the predictors is 0 when log odds is < 0 then prob is < 0.5, likely Benign, when log odds is > 0 then prob is > 0.5, likely Malignant log odds ratio ->log odds increase or become Malignant per 1 unit

increase in predictors for Radius, it means for 1 unit of Radius, the log of the odds of breast cancer being Malignant decreases by (-)4.504. While for Area, it means for 1 unit of Area, the log of the odds of breast cancer being Malignant increases by 3.941

```
#Making predictions in R
# Evaluating the performance of the model for training data
lodds_train <- predict(log_Reg, type = "link")#log odds
preds_lodds_train <- ifelse(lodds_train > 0, "M", "B") #using log odds
#confusion matrix and accuracy for train data
confusionMatrix(as.factor(preds_lodds_train), train$V2)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   B   M
##          B 268  13
##          M   8 157
##
##                Accuracy : 0.9529
##
```

Using the training data set, our model has 94.39% accuracy. It has successfully predicted 265 out of 276 B or 96.0% Benign cases and 156 out of 170 M or 91.76% Malignant cases successfully.

```
# Evaluating the performance of the model for test data
lodds_test <- predict(log_Reg, newdata = test, type = "link")#log odds
preds_lodds_test <- ifelse(lodds_test > 0, "M", "B") #using log odds
#confusion matrix and accuracy for test data
confusionMatrix(as.factor(preds_lodds_test), test$V2)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  B  M
##          B 62  4
##          M  6 38
##
##                Accuracy : 0.9091
##
##             Prevalence : 0.6182
##         Detection Rate : 0.5636
##    Detection Prevalence : 0.6000
##       Balanced Accuracy : 0.9083
##
##        'Positive' Class : B
##
```

Using the test data set, our model has 92.73% accuracy. It has successfully predicted 65 out of 68 B or 95.58% Benign and 37 out of 42 M or 88.09 % Malignant type successfully

(b) Based on the output obtained from the logistic regression in (a), can we identify which factors determine the odds of having breast cancer? Justify your answers.

```
summary(log_Reg)
```

```
##
## Call:
## glm(formula = V2 ~ ., family = "binomial", data = train)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.80436  -0.11493  -0.02087   0.00203   2.85851
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.365e+01  1.662e+01  -0.821  0.41167
## V1          -2.620e-09  4.583e-09  -0.572  0.56760
## V3           1.295e+00  4.762e+00   0.272  0.78564
## V4           4.699e-01  9.239e-02   5.086 3.65e-07 ***
## V5          -5.928e-01  6.386e-01  -0.928  0.35331
## V6           4.360e-02  2.122e-02   2.055  0.03991 *
## V7           1.229e+02  4.331e+01   2.838  0.00454 **
## V8           4.925e+00  2.543e+01   0.194  0.84644
## V9           4.805e+01  1.618e+01   2.969  0.00299 **
## V10          1.117e+01  3.642e+01   0.307  0.75901
## V11          1.489e+01  1.464e+01   1.017  0.30902
## V12         -1.099e+02  1.081e+02  -1.016  0.30955
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Ans: Yes, basing on the coefficients results above, we can check which predictors are statistical significant with P value lower than 0.05. In this case, V4 (texture) has the highest factor since it has a very low P-value (lower than 0.05).Other predictors with lower or close to P-value of 0.05 are also factors that can determine the odds of having breast cancer Benign or Malignant like V6 (Area),V7(Smoothness) and V9(Concavity).

(c) Based on the output obtained from the logistic regression in (a), discuss the impacts of radius and area variables on the odds of having breast cancer. Provide your insight on the findings.

Ans: Radius (V3) predictor only has higher p value which is around 70% significant while Area (V6) predictor has lower p value, more than 95% significant. Therefore, we can include area predictor and drop radius predictor in the model to determine whether a breast cancer is benign or malignant.

Question 2

(a) Compare the complete linkage and single linkage clustering. (Please use your own words for the discussion.)

Ans: Single Linkage uses the distance from the point of interest to the CLOSESTpoint of each cluster. Then merge them.

Complete Linkage uses distance from the point of interest to the FURTHEST point of each cluster and not merge them.

(b) Please do not use R or any programming languages for this question. Please solve the problem manually.

Suppose that you have a dissimilarity matrix of 5 observations as follows

$$D = \begin{bmatrix} 0 & 0.2 & 0.45 & 0.7 & 0.8 \\ 0.2 & 0 & 0.1 & 0.5 & 0.35 \\ 0.45 & 0.1 & 0 & 0.55 & 0.6 \\ 0.7 & 0.5 & 0.55 & 0 & 0.3 \\ 0.8 & 0.35 & 0.6 & 0.3 & 0 \end{bmatrix}$$

The D matrix implies that the dissimilarity between the first and the third obser- vation is 0.45, and the dissimilarity between the second and the fourth observation is 0.5.

• Sketch the dendrogram using the complete linkage clustering approach. Ex- plain and provide the detailed procedure of obtaining the height at which each fusion occurs, and the

observations associated with each leaf in the dendro- gram.

	a	b	c	d	e
a	0				
b	0.2	0			
c	0.45	0.1	0		
d	0.7	0.5	0.55	0	
e	0.8	0.35	0.6	0.3	0

Step1: The lowest distance is 0.1 between b and c. Therefore merge them (bc)

	a	bc	d	e
a				
bc	0.45			
d	0.7	0.55	0	
e	0.8	0.6	0.3	0

Step 2: Lowest distance is 0.3 between cluster d and e. Merge them (de)

To update the distance matrix [a,bc,d,e]

The maximum distance between (b,a) and (c,a) is 0.45

The maximum distance between (b,d) and (c,d) is 0.55

The maximum distance between (b,e) and (c,e) is 0.6

The distance between d and a is 0.7, The distance between a and e is 0.8, The distance between d and e is 0.3

	a	bc	de
a	0		
bc	0.45	0	
de	0.8	0.6	0

Step 3: Lowest distance is 0.45 between a and bc. Merge them (abc)

To update the distance matrix [a,bc,de]

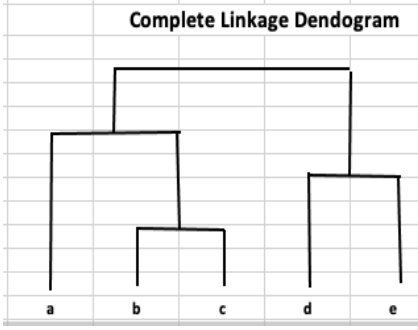
The maximum distance between (a,d) and (a,e) is 0.8

The maximum distance between (bc,d) and (bc,e) is 0.6

The distance between bc and a is 0.45

	abc	de
abc	0	
de	0.8	0

Step 4: Merge the last remaining abc and de into abcde



The procedure for complete linkage is explained on the image above. In each step the pair or clustering is determined by the shortest distance of all the points. In step, the shortest distance is c and b with 0.1. Therefore, we merge or cluster them. Then on step 2, a new distance was made from the new cluster bc. Since we are using complete linkage, the furthest distance between b to other points and c to the other points is used for the new matrix. The process continues until all points are merged or clustered.

• Sketch the dendrogram using the single linkage clustering approach. Explain and provide the detailed procedure of obtaining the height at which each fusion occurs, and the

observations associated with each leaf in the dendrogram.

	a	b	c	d	e
a	0				
b	0.2	0			
c	0.45	0.1			
d	0.7	0.5	0.55	0	
e	0.8	0.35	0.6	0.3	0

Step1: The lowest distance is 0.1 between b and c. Therefore merge b and c

	a	bc	d	e
a	0			
bc	0.2	0		
d	0.7	0.5	0	
e	0.8	0.35	0.3	0

To update the distance matrix [a,bc,d,e]

The minimum distance between (b,a) and (c,a) is 0.2

The minimum distance between (b,d) and (c,d) is 0.5

The minimum distance between (b,e) and (c,e) is 3.5

The distance between d and a is 0.7, The distance between a and e is 0.8, The distance between d and e is 0.3

Step 2: Lowest distance is 0.2 between cluster bc and a. Merge them (abc)

	abc	d	e
abc	0		
d	0.5	0	
e	0.35	0.3	0

To update the distance matrix [abc,d,e]

The minimum distance between (a,d) and (bc,d) is 0.5

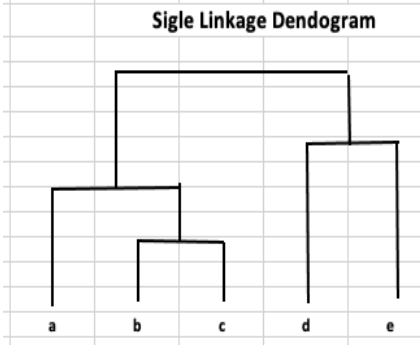
The minimum distance between (a,e) and (bc,e) is 0.35

The distance between d and e is 0.3

Step 3: Lowest distance is 0.3 between e and d. Merge them (de)

	abc	de
abc	0	
de	0.35	0

Step 4: Merge the last remaining abc and de into abcde



The procedure for single linkage is explained on the image above. In each step the pair or clustering is determined by the shortest distance of all the points. In step 1, the shortest distance is c and b with 0.1. Therefore, we merge or cluster them. Then on step 2, a new distance was made from the new cluster bc. Since we are using single linkage, the shortest distance between b to other points and c to the other points is used for the new matrix. The process continues until all points are merged or clustered.

It can be noticed that the height of the which fusion occured were different from complete and single linkage. Although they both started clustering from bc, the next height or clustering was different because they used different methods(shortest and furthest distance). cluster de came first after cluster bc on complete linkage while cluster, abc came first after cluster bc on single linkage.

Question 3

Clustering is a common exploratory technique used in bioinformatics where researchers aim to identify subgroups within diseases using gene expression. Imagine you are asked to analyse the gene expression dataset available in the leukemia dat.Rdata file. This data was originally generated by [Golub et al., Science, 1999]https://science.sciencemag.org/content/sci/286/5439/531.full.pdf and contains the expression level of 1867 selected genes from 72 patients with different types of leukemia.

The data in each column are summarized as follows:

• Column 1: patient id = a unique identifier for each patient (observation)

• Column 2: type = A factor variable with two subtypes of leukemia; acute lym- phoblastic leukemia (ALL, n = 47) and acute myeloblastic leukemia (AML, n = 25).

• Columns 3: - 1869. Gene expression data for 1867 genes, Gene 1, …, Gene 1867.
Assignment Tasks:

The researchers hypothesized that patient samples will cluster by subtype of leukemia based on gene expression. Your task is to use a clustering technique to address this scientific hypothesis and report your results back to the researcher.

(a) Select a clustering technique to apply. Justify your choice.

Ans: K means clustering is chosen for this data because K means can handle big data. It also Doesn't make any assumptions about data distribution.

(b) Implement your chosen clustering technique in R. Describe your implementation (You need to provide details of all steps relating to the implementation of the clus- tering algorithms, such as data preparation including any transformations performed on the data prior to clustering, training the model and evaluating the performance of the model.)

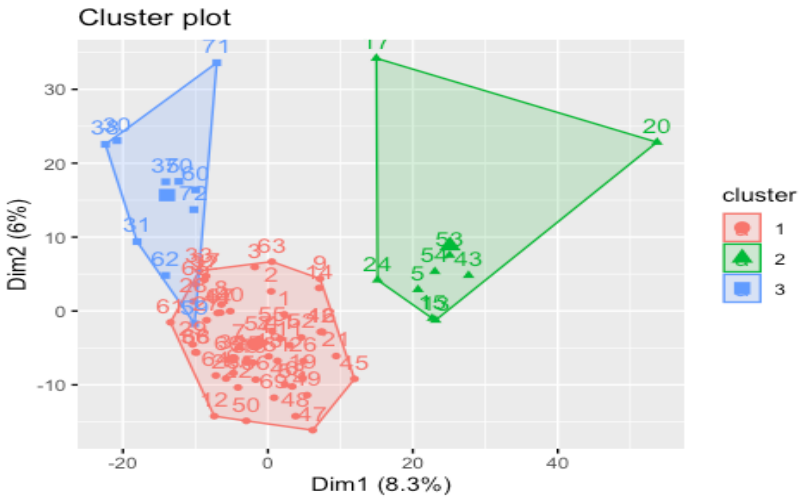
```r
#load the leukemia_dat.Rdata file to R
load("leukemia_dat.Rdata")
#create object data2 <- leukemia_dat
data2 <- leukemia_dat
#Data transformation: Delete patient_id and type variables and retain only Gene variables(Continuous/numericals)
data2 <- data2[,3:1869]
#scale the data to normalize
data2_scaled <- scale(data2)

set.seed(6)
#Appply K means algorithm
kmeans_res <- kmeans(data2_scaled, centers = 3, nstart = 25)
str(kmeans_res)
```

```
## List of 9
##  $ cluster     : int [1:72] 1 1 1 1 2 1 1 1 1 1 ...
##  $ centers     : num [1:3, 1:1867] 0.0274 -0.16917 0.00701 -0.07362
-0.46332 ...
##   ..- attr(*, "dimnames")=List of 2
##   .. ..$ : chr [1:3] "1" "2" "3"
##   .. ..$ : chr [1:1867] "Gene_1" "Gene_2" "Gene_3" "Gene_4" ...
##  $ totss       : num 132557
##  $ withinss    : num [1:3] 80404 15575 22244
##  $ tot.withinss: num 118223
##  $ betweenss   : num 14334
##  $ size        : int [1:3] 53 9 10
##  $ iter        : int 2
##  $ ifault      : int 0
##  - attr(*, "class")= chr "kmeans"
```

As shown above, we are interested in the tot.withinss because that is the total within sums of squares for all of the cluster, in this attempt 118223. Cluster is the particular cluster that each observation is in. Centers are the centroid for each of the variables.

(c) Produce two or more plots to visualize your results. Describe your results as you would in your report to the researcher.

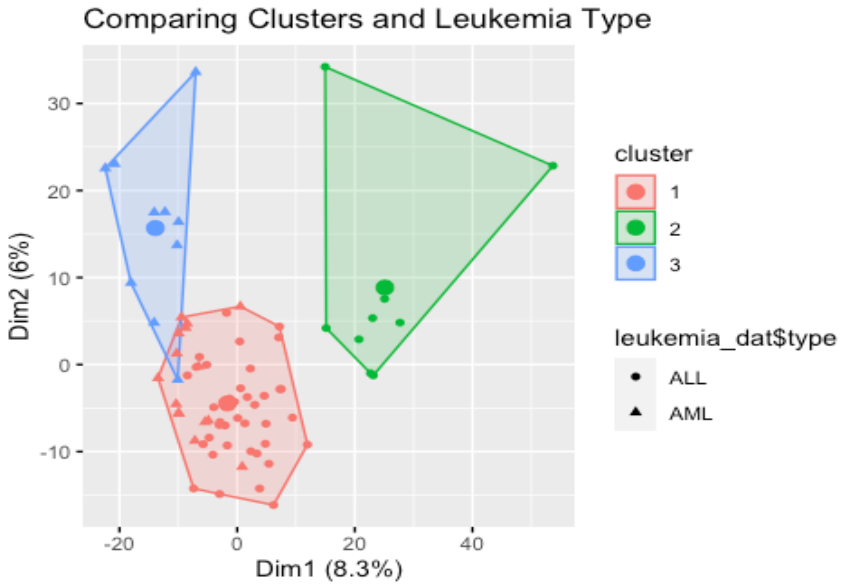
```
#visualization using factoextra library
fviz_cluster(kmeans_res, data = data2_scaled)
```



The plot above shows the clustering of the data(data2) using k =2. It can be noticed that they are clustered properly

```
#visualization of clusters and leukemia type ()
fviz_cluster(kmeans_res, #set up plot
             data = data2_scaled,
             geom = "point", # only shows points and not labels
             shape = 19,# define one shape for all clusters (a circle)
```

```
            alpha = 0)+ # make circles see-though
  geom_point(aes(colour = as.factor(kmeans_res$cluster),
                 shape = leukemia_dat$type))+ #colour by species
  ggtitle("Comparing Clusters and Leukemia Type") #add a title
```



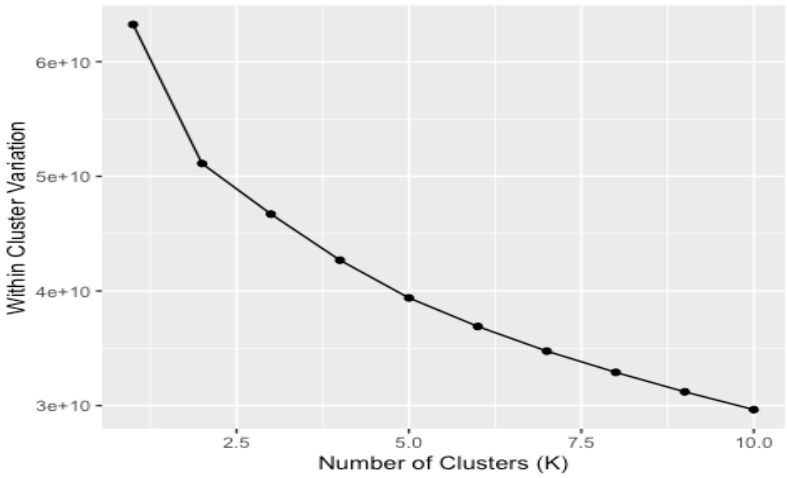
The plot above compares the clustering the the Leukemia Type ALL and AML using K =3. It can be noticed all of the data in cluster 2(Green) are all ALL while all of the data in Cluster 3 (Blue) are all AML. However, the data in cluster 1(red) are mostly ALL with some AML.

```
#perform kmeans & calculate ss
total_sum_squares <- function(k){
  kmeans(data2, centers = k, nstart = 25)$tot.withinss
}

#define a sequence of values for k up to 10 sequences
all_ks <- seq(1,10,1)

#apply to all values of k
choose_k <- sapply(seq_along(all_ks), function(i){
  total_sum_squares(all_ks[i])
})
choose_k
```

```
##  [1] 63245643704 51117195720 46702946588 42685144717 39392938366 36
903172869
##  [7] 34750493684 32901642486 31202105532 29651967287
```

The data above are corresponds the each value of K each performed 25 times (nstart=25). So, cluster variation 63245643704 is the best result from the 25 start when K =1 and 29607967824 is the best result from the 25 start when K = 10

```r
# dataframe for plotting
choose_k_plot <- data.frame(k = all_ks,
                            within_cluster_variation = choose_k)
# plot
ggplot(choose_k_plot, aes(x = k,
                          y = within_cluster_variation))+
  geom_point()+
  geom_line()+
  xlab("Number of Clusters (K)")+
  ylab("Within Cluster Variation")
```



The above graph shows how to select value of K=2 to reduce within cluster of variation of our data

Conclusion: We can therfore conclude that the researchers hypothesis confirms that patients samples cluster by subtype of leukemia based on gene expression as proved with the data and visualisations above.

Appendix:

Rmarkdown code:

---

title: "Assessment2"

author: "Bienvenido Hiyas Jr"

date: "29/11/2020"

output:

  word_document: default

html_document: default

pdf_document: default

Question 1

Imagine you are asked to use a logistic regression to classify whether the breast cancer is benign or malignant. The data Breast cancer.csv is a subset data from Breast Can- cer Wisconsin, which is available from the UCI Data Repository http://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+%28diagnostic%29. The variables are summarised as follows:

• V1: ID number

• V2: Diagnosis (M = malignant, B = benign)

• V3: radius (mean of distances from center to points on the perimeter) • V4: texture (standard deviation of gray-scale values)

• V5: perimeter

• V6: area

• V7: smoothness (local variation in radius lengths)

• V8: compactness (perimeter2 / area — 1.0)

• V9: concavity (severity of concave portions of the contour)

• V10: concave points (number of concave portions of the contour)

• V11: symmetry

• V12: fractal dimension ("coastline approximation" — 1) Assignment tasks:

```{r setup, include=FALSE, results='hide'}

#Import Libraries

library(ISLR)

library(caret, warn.conflicts = F, quietly = T)

library(dplyr)

library(cluster, warn.conflicts = F, quietly = T) #clustering algorithms
```

library(factoextra, warn.conflicts = F, quietly = T) #data visualization

```

```{r , include=TRUE, results='hide'}

#Importing and data exploration

data = read.table("breast_cancer.csv", header = TRUE, sep = ",")

summary(data)

str(data)

##cleaning up data: transform "0" to NA

data[data==0] <- NA

#delete rows with NA's

data = na.omit(data)

str(data)

#transform V2 from char to factor and the reset from char to numeric

data$V1 = as.numeric(data$V1)

data$V2 = as.factor(data$V2)

data$V3 = as.numeric(data$V3)

data$V4 = as.numeric(data$V4)

data$V5 = as.numeric(data$V5)

data$V6 = as.numeric(data$V6)

data$V7 = as.numeric(data$V7)

data$V8 = as.numeric(data$V8)

data$V9 = as.numeric(data$V9)

data$V10 = as.numeric(data$V10)

data$V11 = as.numeric(data$V11)

data$V12 = as.numeric(data$V12)

```

(a) Implement the logistic regression to classify the breast cancer type in R. (Note: You need to provide details of all steps relating to the implementation of the logistic regression such as data preparation, training the model, evaluating the perforamnce of the model on both training and test data and discuss the results.)

```{r , include=TRUE, warning=FALSE}

#partition dta set into train and set and randomly split it

set.seed(123)

#split into training (80%) and test

ind <- createDataPartition(data$V2, p = 0.8, list = F)

train <- data[ind, ]

test <- data[-ind, ]

# print number of observations in test vs. train

c(nrow(train), nrow(test))

# Proportions of people that defaulted and did not default

table(train$V2) %>% prop.table()

#Train the model to predict the likelihood of Malignant based on all predictors using Logistic Regressin in Caret Package

log_Reg <- glm(V2 ~., data = train, family = "binomial")

summary(log_Reg)$coef
```

Intercept shows log odds that the breast cancer is Benign or Malignant when the predictors is 0

when log odds is < 0 then prob is < 0.5, likely Benign,

when log odds is > 0 then prob is > 0.5, likely Malignant

log odds ratio ->log odds increase or become Malignant per 1 unit increase in predictors

for Radius, it means for 1 unit of Radius, the log of the odds of breast cancer being Malignant

decreases by (-)4.504. While for Area, it means for 1 unit of Area,
```

the log of the odds of breast cancer being Malignant increases by 3.941

```{r, include=TRUE, warning = FALSE}
#Making predictions in R
# Evaluating the performance of the model for training data
lodds_train <- predict(log_Reg, type = "link")#log odds
preds_lodds_train <- ifelse(lodds_train > 0, "M", "B") #using log odds
#confusion matrix and accuracy for train data
confusionMatrix(as.factor(preds_lodds_train), train$V2)
```

Using the training data set, our model has 94.39% accuracy. It has successfully predicted 265 out of 276 B or 96.0% Benign cases

and 156 out of 170 M or 91.76% Malignant cases successfully.

```{r, include=TRUE, warning = FALSE}
# Evaluating the performance of the model for test data
lodds_test <- predict(log_Reg, newdata = test, type = "link")#log odds
preds_lodds_test <- ifelse(lodds_test > 0, "M", "B") #using log odds
#confusion matrix and accuracy for test data
confusionMatrix(as.factor(preds_lodds_test), test$V2)
```

Using the test data set, our model has 92.73% accuracy. It has successfully predicted 65 out of 68 B or 95.58% Benign and 37 out of 42 M or 88.09 %

Malignant type successfully

(b) Based on the output obtained from the logistic regression in (a), can we identify which factors determine the odds of having breast cancer? Justify your answers.

```{r, include=TRUE, warning = FALSE}
summary(log_Reg)
```

```

Ans: Yes, basing on the coefficients results, we can check which predictors are statistical significant with P value lower than 0.05. In this case, V4 (texture) has the highest factor since it has a very low P-value (lower than 0.05).Other predictors with lower or close to P-value of 0.05 are also factors that can determine the odds of having breast cancer Benign or Malignant.

##### (c) Based on the output obtained from the logistic regression in (a), discuss the impacts of radius and area variables on the odds of having breast cancer. Provide your insight on the findings.

Ans: radius (V3) predictor only has higher p value which is around 70% significant while area (V6) predictor has lower p value, more than 95% significant. Therefore, we can include area predictor and drop radius predictor in the model to determine whether a breast cancer is benign or malignant.

### Question 2

##### (a) Compare the complete linkage and single linkage clustering. (Please use your own words for the discussion.)

Ans: Single Linkage uses the distance from the point of interest to the CLOSESt point of each cluster. Then merge them.

Complete Linkage uses distance from the point of interest to the FURTHEST point of each cluster and not merge them.

##### (b) Please do not use R or any programming languages for this question. Please solve the problem manually.

Suppose that you have a dissimilarity matrix of 5 observations as follows


```

The D matrix implies that the dissimilarity between the first and the third obser- vation is 0.45, and the dissimilarity between the second and the fourth observation is 0.5.

• Sketch the dendrogram using the complete linkage clustering approach. Ex- plain and provide the detailed procedure of obtaining the height at which each fusion occurs, and the observations associated with each leaf in the dendro- gram.

The procedure for complete linkage is explained on the image above. In each step the pair or clustering is determined by the shortest distance of all the points. In step, the shortest distance is c and b with 0.1. Therefore, we merge or cluster them. Then on step 2, a new distance was made from the new cluster bc. Since we are using complete linkage, the furthest distance between b to other points and c to the other points is used for the new matrix. The process continues until all points are merged or clustered.

• Sketch the dendrogram using the single linkage clustering approach. Explain and provide the detailed procedure of obtaining the height at which each fusion occurs, and the observations associated with each leaf in the dendrogram.

The procedure for single linkage is explained on the image above. In each step the pair or clustering is determined by the shortest distance of all the points. In step 1, the shortest distance is c and b with 0.1. Therefore, we merge or cluster them. Then on step 2, a new distance was made from the new cluster bc. Since we are using single linkage, the shortest distance between b to other points and c to the other points is used for the new matrix. The process continues until all points are merged or clustered.

It can be noticed that the height of the which fusion occured were different from complete and single linkage. Although they both started clustering from bc, the next height or clustering was different because they used different methods(shortest and furthest distance). cluster de came first after cluster bc on complete linkage while cluster, abc came first after cluster bc on single linkage.

Question 3

Clustering is a common exploratory technique used in bioinformatics where researchers aim to identify subgroups within diseases using gene expression. Imagine you are asked to analyse the gene expression dataset available in the leukemia dat.Rdata file. This data was originally generated by [Golub et al., Science, 1999]https://science.sciencemag.

org/content/sci/286/5439/531.full.pdf and contains the expression level of 1867 selected genes from 72 patients with different types of leukemia.

The data in each column are summarized as follows:

• Column 1: patient id = a unique identifier for each patient (observation)

• Column 2: type = A factor variable with two subtypes of leukemia; acute lym- phoblastic leukemia (ALL, n = 47) and acute myeloblastic leukemia (AML, n = 25).

• Columns 3: - 1869. Gene expression data for 1867 genes, Gene 1, ..., Gene 1867.
Assignment Tasks:

The researchers hypothesized that patient samples will cluster by subtype of leukemia based on gene expression. Your task is to use a clustering technique to address this scientific hypothesis and report your results back to the researcher.

(a) Select a clustering technique to apply. Justify your choice.

Ans: K means clustering is chosen for this data because K means can handle big data. It also Doesn't make any assumptions about data distribution.

(b) Implement your chosen clustering technique in R. Describe your implementation (You need to provide details of all steps relating to the implementation of the clus- tering algorithms, such as data preparation including any transformations performed on the data prior to clustering, training the model and evaluating the performance of the model.)

```{r, include=TRUE, warning = FALSE}

#load the leukemia_dat.Rdata file to R

load("leukemia_dat.Rdata")

#create object data2 <- leukemia_dat
```

data2 <- leukemia_dat

#Data transformation: Delete patient_id and type variables and retain only Gene variables(Continuous/numericals)

data2 <- data2[,3:1869]

#scale the data to normalize

data2_scaled <- scale(data2)

set.seed(6)

#Appply K means algorithm

kmeans_res <- kmeans(data2_scaled, centers = 3, nstart = 25)

str(kmeans_res)

```

As shown above, we are interested in the tot.withinss because that is the total within sums of squares for all of the cluster, in this attempt 118223. Cluster is the particular cluster that each observation is in. Centers are the centroid for each of the variables.

##### (c) Produce two or more plots to visualize your results. Describe your results as you would in your report to the researcher.

```{r, include=TRUE, warning = FALSE}

#visualization using factoextra library

fviz_cluster(kmeans_res, data = data2_scaled)

```

The plot above shows the clustering of the data(data2) using k =2. It can be noticed that they are clustered properly

```{r, include=TRUE, warning = FALSE}

#visualization of clusters and leukemia type ()

fviz_cluster(kmeans_res, #set up plot

        data = data2_scaled,

        geom = "point", # only shows points and not labels

```
        shape = 19,# define one shape for all clusters (a circle)
        alpha = 0)+ # make circles see-though
  geom_point(aes(colour = as.factor(kmeans_res$cluster),
           shape = leukemia_dat$type))+ #colour by species
  ggtitle("Comparing Clusters and Leukemia Type") #add a title
```

The plot above compares the clustering the the Leukemia Type ALL and AML using K =3. It can be noticed all of the data in cluster 2(Green) are all ALL while all of the data in Cluster 3 (Blue) are all AML. However, the data in cluster 1(red) are mostly ALL with some AML.

```{r, include=TRUE, warning = FALSE}
#perform kmeans & calculate ss
total_sum_squares <- function(k){
  kmeans(data2, centers = k, nstart = 25)$tot.withinss
}
#define a sequence of values for k up to 10 sequences
all_ks <- seq(1,10,1)
#apply to all values of k
choose_k <- sapply(seq_along(all_ks), function(i){
  total_sum_squares(all_ks[i])
})
choose_k
```

The data above are corresponds the each value of K each performed 25 times (nstart=25). So, cluster variation 63245643704 is the best result from the 25 start when K =1 and 29607967824 is the best result from the 25 start when K = 10

```{r, include=TRUE, warning = FALSE}
```

```
# dataframe for plotting
choose_k_plot <- data.frame(k = all_ks,
                            within_cluster_variation = choose_k)
# plot
ggplot(choose_k_plot, aes(x = k,
                          y = within_cluster_variation))+
  geom_point()+
  geom_line()+
  xlab("Number of Clusters (K)")+
  ylab("Within Cluster Variation")
```

The above graph shows how to select value of K=2 to reduce within cluster of variation of our data

Conclusion: We can therfore conclude that the researchers hypothesis confirms that patients samples cluster by subtype of leukemia based on gene expression as proved with the data and visualisations above.