# Tugas7

Kenneth Manuel (160419041), Jehuda Rivaldo (160419133)

4/18/2021

```
library("readxl")
library("factoextra")
library("dplyr")
library("MASS")
library("nnet")
```

```
data <- read_excel('Salespeople-data.xlsx')
head(data)
```

```
## # A tibble: 6 x 7
##   Salegrow saleproft Newsale createst Mechtest absttest mathtest
##      <dbl>     <dbl>   <dbl>    <dbl>    <dbl>    <dbl>    <dbl>
## 1       93        96    97.8        9       12        9       20
## 2       88.8      91.8  96.8        7       10       10       15
## 3       95       100.   99          8       12        9       26
## 4      101.      104.  107.        13       14       12       29
## 5      102       108.  103         10       15       12       32
## 6       95.8      97.5  99.3       10       14       11       21
```

## (1) Buatkan Fungsi Diskriminan Linear dan Quadratik dalam R untuk data Salespeople. Gunakan cara Clustering dengan target 3 cluster untuk menciptakan respon yang akan diukur ketepatan klasifikasinya dengan menggunakan fungsi diskriminan anda.

Clustering dengan 3 target cluster

```
kmeans <- kmeans(data, centers = 3)
kmeans
```

```
## K-means clustering with 3 clusters of sizes 17, 15, 18
##
## Cluster means:
##   Salegrow saleproft   Newsale  createst Mechtest  absttest mathtest
## 1  90.40588  95.37647  97.91765  9.058824 11.64706  8.941176 17.82353
## 2 106.18000 110.65333 107.44000 13.400000 16.33333 11.733333 41.73333
## 3 101.01111 107.21667 103.57222 11.444444 14.77778 11.111111 31.05556
##
## Clustering vector:
##  [1] 1 1 1 3 3 1 1 2 3 2 3 3 3 3 3 1 3 3 3 1 2 1 2 2 3 1 2 2 1 3 1 2 2 1 3
## [30] 2 2 3 1 3 1 1 2 1 1 3 2
##
## Within cluster sum of squares by cluster:
## [1] 1497.588 1009.337 1023.857
##  (between_SS / total_SS =  77.5 %)
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
## [6] "betweenss"    "size"         "iter"         "ifault"
```

Plot cluster

```
fviz_cluster(kmeans, data = data, ellipse.type = "norm")
```


Attach cluster result label ke data

```
data.cluster <-
  cbind(data, `kmeans` = kmeans[["cluster"]]) # simpan label hasil clustering
head(data.cluster)
```

```
##   Salegrow saleproft Newsale createst Mechtest absttest mathtest kmeans
## 1     93.0      96.0    97.8        9       12        9       20      1
## 2     88.8      91.8    96.8        7       10       10       15      1
## 3     95.0     100.3    99.0        8       12        9       26      1
## 4    101.3     103.8   106.8       13       14       12       29      3
## 5    102.0     107.8   103.0       10       15       12       32      3
## 6     95.8      97.5    99.3       10       14       11       21      1
```

Dengan mengasumsi bahwa masing-masing variabel berdistribusi normal univariate. Split data ke training dan test set

```
# Split data to training and test set
set.seed(123)
training.samples <-
  sample(seq(nrow(data.cluster)), size = floor(0.75 * nrow(data.cluster)), replace = F)
train.data <- data.cluster[training.samples, ]
test.data <- data.cluster[-training.samples, ]
```

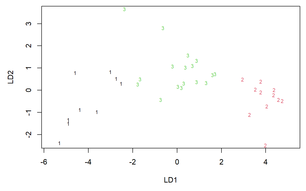
Compute LDA

```
model <- lda(formula = kmeans ~ ., data = train.data)
model
```

```
## Call:
## lda(kmeans ~ ., data = train.data)
##
## Prior probabilities of groups:
##         1         2         3
## 0.2432432 0.2972973 0.4594595
##
## Group means:
##   Salegrow saleproft   Newsale createst Mechtest  absttest mathtest
## 1  91.46667  96.92222  98.58889 10.44444 12.44444  8.777778 18.55556
## 2 106.82727 119.00000 107.04545 13.18182 16.81818 11.909091 42.18182
## 3 100.95294 107.65294 103.58824 11.35294 14.94118 10.941176 31.23529
##
## Coefficients of linear discriminants:
##                 LD1         LD2
## Salegrow   0.1634183  0.6651413
## saleprofit 0.5077856 -0.3382152
## Newsale    0.1479681  0.5033317
## createst  -0.1603818 -0.4649611
## Mechtest  -0.3731997  0.1544920
## absttest   0.1271026 -0.7029662
## mathtest  -0.2256012 -0.1739269
##
## Proportion of trace:
##   LD1   LD2
## 0.947 0.053
```

Plot

```
plot(model, col = as.integer(train.data$kmeans))
```




Make predictions

```
predictions <- predict(object = model, newdata = test.data)
```

Hasil model accuracy

```
mean(predictions$class == test.data$kmeans)
```

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

## (2) Dengan menggunakan Logistic Regression, lakukan ketepatan klasifikasi pada hasil Clustering di atas.

Split data ke training dan test set

```
train.x <- sample_frac(data.cluster, 0.75)
sample_id <- as.numeric(rownames(train)) # rownames() returns character (therefore use as.numeric)
test <- data.cluster[-sample_id, ]
# Set baseline
#train$kmeans <- relevel(train$kmeans, ref = "3")
```

Use multinom() function to fit model then use summary() to explore beta coefficients

```
multinom.fit <- multinom(kmeans ~ . , data = train) # Training the multinomial model
```

```
## # weights:  27 (16 variable)
## initial  value 41.747267
## iter  10 value 0.012757
## iter  20 value 1.331187
## iter  30 value 0.002098
## iter  40 value 0.000124
## final  value 0.000078
## converged
```

```
summary(multinom.fit) # Checking the model
```

```
## Call:
## multinom(formula = kmeans ~ ., data = train)
##
## Coefficients:
##   (Intercept)   Salegrow saleprofit   Newsale  createst  Mechtest  absttest
## 2    3.415513 -14.538694  10.65329 -4.378194  8.216882 -1.360839 16.421996
## 3   -8.154516   6.148452 -11.68854  1.025055 -1.026111 15.131750 -8.442853
##   mathtest
## 2 15.14617
## 3 15.20410
##
## Std. Errors:
##   (Intercept) Salegrow saleprofit  Newsale   createst  Mechtest absttest
## 2    1.577435 168.3599  189.0829 165.3463   16.05685  25.23893  17.3532
## 3   20.981033 678.3779  477.4473 100.4808 1468.46350 239.59070 308.8196
##    mathtest
## 2  75.91054
## 3 170.20235
##
## Residual Deviance: 0.000155446
## AIC: 32.00016
```

```
#exp(coef(multinom.fit)) ## extracting coefficients from the model and exponentiate
#head(probability.table <- fitted(multinom.fit))
```

Extracting coefficients from the model and exponentiate

```
## extracting coefficients from the model and exponentiate
exp(coef(multinom.fit))
```

```
##    (Intercept)     Salegrow    saleprofit     Newsale     createst     Mechtest
## 2 3.043254e+01 4.852053e-07 4.233180e+04 0.01254801 3702.9370969 2.564455e-01
## 3 2.874343e-04 4.679924e+02 8.389392e-06 2.78724979    0.1966931 3.729370e+06
##      absttest mathtest
## 2 1.355134e+07  3783525
## 3 2.154346e-04  4009190
```

Make predictions

```
# Predicting values for train dataset
train$precticed <- predict(multinom.fit, newdata = train, "class")

# Building classification table
ctable <- table(train$kmeans, train$precticed)
```

Model Accuracy

```
round((sum(diag(ctable)) / sum(ctable)), 2)
```

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

## (3) Bandingkan hasil klasifikasi anda dengan hasil fungsi diskriminan anda. Mana yang anda pilih (yang lebih tepat cara klasifikasinya – cara diskriminan atau Logistic Regression–?)

Kesimpulan: Baik hasil klasifikasi cara diskriminan maupun logistic regression pada kasus Sales People model keduanya mengklasifikasi 100% observasi dengan benar sehingga keduanya cukup tepat.