2018R2 Data Mining (STAT5104) Assignment 1 Q1

Yiu Chung WONG 1155017920

```
knitr::opts_chunk$set(echo = TRUE)
library(dplyr)
set.seed(17920)
```

mdist function

```
mdist<-function(x) {
   t<-as.matrix(x)         # transform x to a matrix
   m<-apply(t,2,mean)      # compute column mean
   s<-var(t)               # compute sample covariance matrix
   mahalanobis(t,m,s)      # using built-in mahalanobis function
}
```

Outlier detection

a)

```
d <- read.csv("Telephone.csv", header = TRUE)
d0 <- filter(d, Change == 0) #select observations with Change == 0
d1 <- filter(d, Change == 1) #select observations with Change == 1
```

b)

```
x0 <- select(d0, -c(2,3,18))  #remove column 2, 3, 18
x1 <- select(d1, -c(2,3,18))  #remove column 2, 3, 18
md0 <- mdist(x0)              #calculate distance
md1 <- mdist(x1)              #calculate distance
```

c)

```
cutoff <- qchisq(0.99,df = ncol(d) - 3) #calculate Chi-Square cut-off
dc0 <- filter(d0, md0 < cutoff)         #select observations where distance is below cutoff
dc1 <- filter(d1, md1 < cutoff)         #select observations where distance is below cutoff
```

There are 37 outliners in `d0`; there are 8 outliners in `d1`

d)

```
dc <- rbind(dc0, dc1)                    #row bind observations
write.table(dc, "tele.csv", sep = ",")   #write to csv files
```